Comparing performances of logistic regression, decision trees , and neural networks for classifying heart disease patients

Anchana Khemphila
Software Systems Engineering Laboratory
Department of Mathematics and Computer Science
Faculty of Science, King Mongkut's Institute
of Technology Ladkrabang
Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand.
s0067103@kmitl.ac.th

Veera Boonjing
Software Systems Engineering Laboratory
Department of Mathematics and Computer Science
Faculty of Science, King Mongkut's Institute
of Technology Ladkrabang
Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand.
kbveera@kmitl.ac.th

***Abstract*—In this study, performances of classification techniques were compared in order to predict the presence of the patients getting a heart disease.A retrospective analysis was performed in 303 subjects.We compared the performance of logistic regression(LR),decision trees(DTs) , and Artificial neural networks (ANNs).The variables were medical profiles are age,Sex,Chest Pain Type,Blood Pressure,Cholesterol,Fasting Blood Sugar, Resting ECG,Maximum Heart Rate,Induced Angina,Ole Peak,Slope,Number Colored Vessels,Thal and Concept Class.We have created the model using logistic regression classifiers , artificial neural networks and decision trees that they are often used for classification problems.Performances of classification techniques were compared using lift chart and error rates.In the result, artificial neural networks have the greatest area between the model curve and the baseline curve.The error rates are 0.22,0.198,0.21,respectively for logistic regression , artificial neural networks and decision trees.The neural networks exhibited sensitivity of 81.1% , specificity of 78.7% and accuracy of 80.2%,while the decision tree provided the prediction performance with a sensitivity,specificity and accuracy of 81.7%,76.0% and 79.3%.And the logistic regression provided the prediction performance with a sensitivity,specificity and accuracy of 81.2%,73.1% and 77.7% Artificial neural networks have the least of error rate and has the highest accuracy ,therefore Artificial neural networks is the best technique to classify in this data set.**

Keywords: **data mining, data mining techniques, heart disease , Logistic regression classifiers , Artificial neural networks , Classification trees , Decision trees.**

I. Introduction

The heart is the organ that pumps blood, with its life-giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer.And if the heart stops working altogether, death occurs within minutes. Life itself is completely dependent on the efficient operation of the heart. Heart disease is not contagious you can't catch it like you can the flu or a cold. Instead, there are certain things that increase a person's chances of getting cardiovascular disease. Doctors call these things risk factors.Some of these risk factors a person can't do anything about, like being older and having other people in the family who have had the same problems. But people do have control over some risk factors smoking, having high blood pressure, being overweight, and not exercising can increase the risk of getting cardiovascular disease. Many people do not realize they have cardiovascular disease until they have chest pain, a heart attack, or stroke. These kinds of problems often need immediate attention and the person may need to go to the emergency department of a hospital.

Data mining has been heavily used in the medical field, to include patient diagnosis records to help identify best practices [11]. The difficulties posed by prediction problems have resulted in a variety of problem-solving techniques. For example, data mining methods comprise artificial neural networks and decision trees, and statistical techniques include linear regression and stepwise polynomial regression. It is difficult, however, to compare the efficacy of the techniques and determine the best one because their performance is data-dependent.A few studies have compared data mining and statistical approaches to solving prediction problems [8].Many people compared linear regression, stepwise polynomial regression and neural networks in the context of predicting student GPAs.The comparison studies have mainly considered a specific data set or the distribution of the dependent variable.

In the next section, we review the three data mining techniques i.e., Logistic regression classifiers(LR) ,Artificial neural networks(ANNs) ,and Decision tree(DTs).We compare the classification accuracy among them on section 3 .Section 4 is result of experiment. Finally, section 5 contains concluding.

II. Data mining techniques

Few works have been published on the comparison of classification techniques in different areas.Moisen and Frescino (2002) compared linear models, generalized additive models, classification and regression tree (CART), Multi-

978-1-4244-7818-7/10/$26.00 ©2010 IEEE

variate Additive Regression Splines (MARS), and artificial neural networks for mapping forest characteristics in the Interior Western United States using forest inventory field data and ancillary satellite-based information [15], Delen, Walker, and Kadam (2004) compared LR, decision tree (C5) and artificial neural networks for predicting the survivability of diagnosed cases for breast cancer [7]. Stark and Pfeiffer (1999) compared LR, classification tree algorithms (ID3, C4.5, CHAID, CART) and artificial neural networks to solve classification problems in complex data sets in veterinary epidemiology [17]. Colombet et al. (2000) evaluated the implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database [4].King, Feng, and Sutherland (1995) compared symbolic learning (CART, C4.5, NewID, AC2, ITrule, Cal5, and CN2), statistics (Naive Bayes, k-nearest neighbor, kernel density, linear discriminant, quadratic discriminant, LR, projection pursuit, and Bayesian networks), and neural networks (back-propagation and RBF) algorithms on twelve datasets with respect to large real-world problems [12].Engin Avci and Ibrahim Turkoglu (2009) study an intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases [1].Marcel A.J. van Gerven , Rasa Jurgelenaite ,Babs G. Taal, Tom Heskes and Peter J.F. Lucas (2006)predict carcinoid heart disease with the noisy-threshold classifier [18].Resul Dasa, Ibrahim Turkoglub, Abdulkadir Sengurb(2009) diagnosis of valvular heart disease through neural networks ensembles [5].Imran Kurt , Mevlut Ture , A. Turhan Kurum (2008) compare performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease [10]. In this research,we would like to mention about classification.We used the classification techniques as the logistic regression classifiers(LR) ,artificial neural networks(ANNs) ,and decision trees(DTs) to deduce the real defalut probability and offered the solutions to the following two questions:
1.Is there any difference of classification accuracy among the three data mining techniques?
2.Could the estimated probability of default produced from data mining methods represent the real probability of default?
LR,ANNs and DTs are often used for classification problems.LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of set of independent variables which are continuous, categorical, or both. Furthermore, it assumes that measures of dependent variables are independently and randomly sampled, all potentially relevant independent variables are in the model and all independent variables in the model are relevant [9].Neural networks have been used to model medical and functional outcomes of dangerous disease. They have become a popular tool for classification, as they are very flexible, not assuming any parametric form for distinguishing between categories [13]. CART is inherently non-parametric that no assumptions are made regarding the underlying distribution of values of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure [3]. We would like to explain tree classification techniques as follows.

A. Logistic regression(LR)

Logistic regression can be considered a special case of linear regression models. However, the binary response variable violates normality assumptions of general regression models. A logistic regression model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. The weaknesses are that LR cannot properly deal with the problems of non-linear and interactive effects of explanatory variables.LR is a regression method for predicting a dichotomous dependent variable. In producing the LR equation, the maximum-likelihood ratio was used to determine the statistical significance of the variables [9], [16].LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of set of predictor variables.It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous.LR model for p independent variables can be written as

$$H(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_{\mathrm{p}} x_{\mathrm{p}})}} \quad (1)$$

where $P(Y=1)$ is probability of presence of CAD. and β_0,β_1,β_2,...,β_{p} are regression coefficients.There is a linear model hidden within the logistic regression model.Te natural logarithm of the ratio of $P(Y=1)$ to $(1-P(Y=1))$ gives a linear model in X_{i}:

$$\begin{aligned} g(x) &= ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) \\ &= \beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_{\mathrm{p}} x_{\mathrm{p}} \end{aligned} \quad (2)$$

The $g(x)$, has many of the desirable properties of a linear regression model.The independent variables can be a combination of continuous and categorical variables [9] [16].

LR model can include the main effects and interaction terms. An important step in the process of modeling a set of data is determining whether there is evidence of interaction and confounder term in the data. The term confounder is used to describe a covariate that is associated with both the dependent variable of interest and a primary independent variable. When both associations are present then the relationship between independent variable and the

dependent variable is said to be confounded. LR model to check for the confounder status of a covariate is to compare the estimated coefficient for the independent variable from models containing and not containing the covariate.Any clinically important change in the estimated coefficient for the independent variable suggests that the covariate is a confounder and should be included in the model, regardless of the statistical significance of its estimated coefficient. One way to test for confounder and interactions in LR is to start with a main effects model,and use a forward-selection method to find interaction terms which significantly reduce the likelihood ratio test statistic [9].

B. Artificial neural networks(ANNs)

Artificial neural networks was inspired by attempts to simulate biological neural systems. The human brain consists primarily of nerve cells called neurons,linked together with other neurons via stand of fiber called axons. Axons are used to transmit nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites,which are extensions from the cell body of the neurons.The contact point between a dendrite and an axon is called a synapse.

Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm.It is supervised networks so they require a desired response to be trained.It learns how to transform input data in to a desired response,so they are widely used for pattern classification.With one or two hidden layers,they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems.The most popular static network in the multilayer.The multilayer is trained with error correction learning,which is appropriate here because the desired multilayer response is the arteriographic result and as such known.Error correction learning works in the following way from the system response at neuron j at iteration t, $y_j(t)$,and the desired response $d_j(t)$for given input pattern an instantaneous error $e_j(t)$is defined by

$$e_j(t) = d_j(t) - y_j(t) \tag{3}$$

Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e.

$$w_{jk}(t+1) = w_{jk}(t) + \eta\delta_j(t)x_k(t) \tag{4}$$

The $\eta(t)$ is the learning-rate patameter.The $w_{jk}(t)$ is the weight connecting the output of neuron k to the input neuron j at iteration t.The local error$\delta_j(t)$ can be computed as a weighted sum of errors at the internal neurons.

C. Classification and Regression Trees(CART)

In a CART structure, each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes. The top-most node in a tree is the root node. CART are applied when the response variable is qualitative or quantitative discrete. CART perform a classification of the observations on the basis of all explanatory variables and supervised by the presence of the response variable. The segmentation process is typically carried out using only one explanatory variable at a time. CART are based on minimizing impurity, which refers to a measure of variability of the response values of the observations. CART can result in simple classification rules and can handle the nonlinear and interactive effects of explanatory variables. But their sequential nature and algorithmic complexity can make them depends on the observed data, and even a small change might alter the structure of the tree. It is difficult to take a tree structure designed for one context and generalize it for other contexts.

CART is a recursive partitioning method to be used both for regression and classification.CART is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set.The best predictor is chosen using a variety of impurity or diversity measures(Gini, twoing, ordered twoing and least-squared deviation).The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable [3].In this study, we used measute of Gini impurity that used for categorical target variables.

Gini Impurity Measure :

The Gini index at node $t, g(t)$, is defined as

$$g(t) = \sum_{j\neq i} p(j|t)p(i|t) \tag{5}$$

where i and j are categories of the target variable.The equation for the Gini index can also be written as

$$g(t) = 1 - \sum_{j} p^2(j|t)$$

Thus, when the cases in a node are evenly distributed across the categories, the Gini index takes its maximum value of $1-(1-k)$, where k is the number of categories for the target variable. When all cases in the node belong to the same category, the Gini index equals 0. If costs of misclassification are specified, the Gini index is computed as

$$g(t) = \sum_{j\neq i} C(i|j)p(j|t)p(i|t) \tag{6}$$

where $C(i|j)$ is the probability of misclassifying a category j case as category i. The Gini criterion function of split s

at node t is defined as

$$\Phi(s,t) \quad = \quad g(t) - p_L g(t_L) - p_R g(t_R) \qquad (7)$$

where p_L is the proportion of cases in t sent to the left child node,and p_R is the proportion send to the right child node.The split s is chosen to maximize the value of $\Phi(s,t)$.This value is reported as the improvement in the tree [3].

III. Classification accuracy among data mining techniques.

A. Description of the data.

Our study took the heart disease patients data from Dr.Robert Detrano at the VA Medical Center in Long Beach California. The dataset consists of 303 patients.One hundred thirty-eight with the heart disease.The original dataset contains 13 numeric attributes and a fourteenth attribute in dicating whether the patient has a heart condition.This dataset is interesting because it represents real patient data and has been used extensively for testing various data mining techniques.We can use this data together with one of more data mining techniques to help us develop profiles for differentiating individuals with heart disease from those without known heart conditions.This study reviewed the literature and used the following 14 variables as explanatory variables and Table.1:

−Age=X1 : Age in years.
−Sex=X2 : Patient gender.
−Chest Pain Type=X3 : NoTang=Nonanginal pain.
−Blood Pressure=X4 : Resting blood pressure upon hospital admission.
−Cholesterol=X5 : Serum cholesterol.
−Fasting Blood Sugar<120=X6 : Is fasting blood sugar less than 120?
−Resting ECG=X7 : Hyp=Left ventricular hypertrophy.
−Maximum Heart Rate=X8 : Maximum heart rate achieved.
−Induced Angina=X9 : Does the patient experience angina as a result of exercise?
−Ole Peak=X10 : ST depression induced by exercise relative to rest.
−Slope=X11 : Slope of the peak exercise ST segment.
−Number Colored Vessels=X12 : Number of major vessels colored by fluoroscopy.
−Thal=X13 : Normal, fixed defect, reversible detect.
−Concept Class=X14 : angiographic disease status.

Before building models,the data set were randomly split into two subsets,60 percentages(n=182) of the the data for training set and 40 percentages(n=121) of the data for validation set.

The Lift Chart measures the effectiveness of models by calculating the ratio between the result obtained with a model and the result obtained without a model.In the lift chart,it is represented by the random curve.It shows you the lift factor to how many times it is better to use a model in contrast to not using a model.Lift Chart is well know in the data mining community specialized in marketing and sales applications [2].The greater the area between the model curve and the baseline curve,the better model.

Table I
CARDIOLOGY PATIENT DATA.

Attribute	Values	Numberic
Age	Numeric	Numeric
Sex	Male,Female	1,0
Chest Pain Type	Angina,Abnormal,NoTang,Asymp	1-4
Blood Pressure	Numeric	Numeric
Cholesterol	Numeric	Numeric
Fasting Blood Sugar	<120 True,False	1,0
Resting ECG	Normal,Abnormal,Hyp	0,1,2
Maximum Heart Rate	Numeric	Numeric
Induced Angina	True,False	1,0
Old Peak	Numeric	Numeric
Slope	Up,Flat,Down	1,2,3
Number Colored Vessels	0,1,2,3	0,1,2,3
Thal	Number,Fix,rev	3,6,7
Concept Class	Healthy,Sick	1,0

Table II
COMPARISON OF THE PERFORMANCE OF MODELS FOR TRAINING SET

Method	ACC%	SEN%	SPE%	PPR%	NPR%
Decision trees	91.2	90.1	92.8	93.6	89.6
Neural networks	92.3	91.7	92.9	93.7	90.8
Logistic regression	91.7	90.8	92.8	93.7	89.6

IV. Result.

The lift chart of the three data mining techniques are shown as figure 1 and figure 2.In the training data and the validation data, based on area between the model curve and

Table III
COMPARISON OF THE PERFORMANCE OF MODELS FOR VALIDATION SET

Method	ACC%	SEN%	SPE%	PPR%	NPR%
Decision trees	79.3	81.7	76.0	82.8	74.5
Neural networks	80.2	81.1	78.7	78.9	72.5
Logistic regression	77.7	81.2	73.1	80.0	74.5

Table IV
CLASSIFICATION ACCURACY.

Method	Error rate of Training	Error rate of Validation
Decision trees	0.082	0.21
Neural networks	0.077	0.198
Logistic regression	0.082	0.22

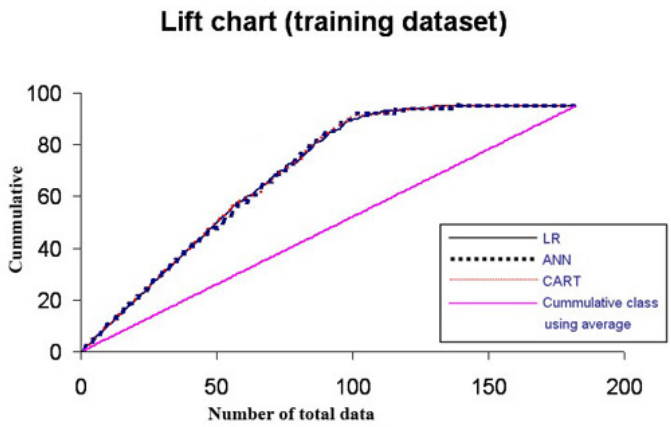

Figure 1. Lift chart of training dataset.

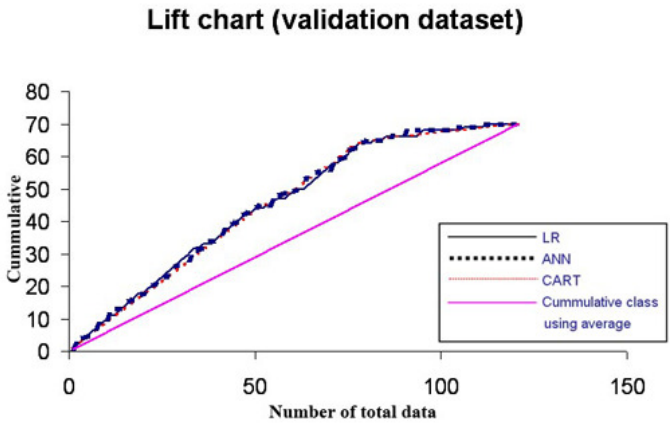

Figure 2. Lift chart of validation dataset.

the base line curve.Neural networks ,Logistic regression and Decision tree have the similarly area.A comparison of the error rate for training set and validation set of classification techniques are shown in Table.4.,in the training data.Neural networks and have the lowest error rates(=0.077),next is decision tree have the error rates(=0.081).The biggest error rates is logistic regression(=0.082) .In the validation data,neural networks have the lowest error rates(=0.198).next is decision tree have the error rates(=0.21).The biggest error rates is logistic regression(=0.22).

A comparison of the accuracy(ACC) and sensitivity(SEN),specificity(SPE),positive predictive rate(PPR)and negative predictive rare(NPR) for training set and validation set are shown in table2,3.Thus,prediction accuracy of 80.2% was obtained by neural networks,while 79.3% was observed from decision tree and 77.7% was obtained by logistic regression.Neural networks ,logistic regression and decision tree presented sensitivity of 81.1%,81.2% and 81.7%,respectively.Furthermore,the specificity of prediction made by the neural networks was 78.7%,which was higher than that obtained from logistic regression 73.1% and decision tree 76.0%.

V. Conclusion.

This paper examines the three classification techniques in data mining and compares the performance of classification among them.In the classification accuracy among the three data mining techniques, the results show that there are the differences in error rates.However there are relatively differences in area.Neural networks perform classification more accurately than the other methods.For neural networks applications in cardiology,Dassen et al.(1998) reported that a major advantage of using neural networks to model the relationship between the possible signs and symptoms and the diagnosis is the fact that this relationship does not have to be a linear one.However,despite the wide interest in the application of neural networks there are a number of limitations that make the introduction of these tools daily practice difficult.First because of the black box nature of neural networks,it is difficult to explain.Second problem is how to validate a trained neural network. They suggested to cardiologists for avoiding from these problems [6].Similarly to this study,we suggests that the trained neural network should be evaluated using an independent test set; the need for a test set to evaluate criteria is also present for classical systems, but the ability to learn all cases by heart makes it even more essential for neural networks; the most important advantage of neural networks are that it draws consistent conclusions and it can be built and evaluated using a large number of cases.

We suggest that age,sex,chest pain type,blood pressure,cholesterol,fasting blood sugar,resting ECG,maximum heart rate,induced angina,old peak,slope,number colored vessels and thal may be used as reliable indicators to predict presence of heart disease.All model had not very high accuracy because our study had several limitations that were the lack of input variable for risk factors,such as exercise behaviour,lipoprotein,hyperuricemia and homocysteinemia.In our study,we suggest that data should be better explored and processed by high performance modeling methods.

In the future ,we will focus on heart disease or any other clinical conditions with classification using neural network and feature selection.

References

[1] Avci, E. and Turkoglu, I. 2009. An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Syst. Appl. 36, 2 (Mar. 2009), 2873-2878.

[2] Berry, M.J.A. and Linoff, G. (1999): Data Mining Techniques: For Marketing, Sales, and Customer Support. Morgan Kaufmann Publishers.

[3] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Montery: Wadsworth and Brooks/ Cole.

[4] Colombet, I., Ruelland, A., Chatellier, G., Gueyffier, F., Degoulet, P., & Jaulent, M. C. (2000). Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. Proc AMIA Symp, 156160.

[5] Das, R., Turkoglu, I., and Sengur, A. 2009. Diagnosis of valvular heart disease through neural networks ensembles. Comput. Methods Prog. Biomed. 93, 2 (Feb. 2009), 185-191.

[6] Dassen, W. R. M., Egmont-Petersen, M., & Mulleneers, R.G.A. (1998).Arti.cial neural networks in cardiology; a review. In P. E. Vardas(Ed.),Cardiac arrhythmias, pacing and electrophysiology (pp. 205211).Great Britain: Kluwer Academic Publishers.

[7] Delen, D., Walker, G., & Kadam, A. (2004). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2), 113127.

[8] Gorr, W. L., Nagin, D., & Szczypula, J.(1994). Comparative study of artificial neural network and statistical medels for predicting student grade point averages. International Journal of Forecasting, 10, 17-34.

[9] Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. New York: John Wiley & Sons.

[10] Imran Kurt, Mevlut Ture,A.Turhan Kurum.2008.Comparing performances of logistic regression, classi.cation and regression tree, and neural networks for predicting coronary artery disease.Expert Systems with Applications 34(2008),366-374.

[11] J.Morris,"Beyond Clinical Documentation: Using the EMR as a Quality Tool," Health Management Technology, volime 25, issue 11, November 2004, pp. 20,22-24.

[12] King, R. D., Feng, C., & Sutherland, A. (1995). Statlog-comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence, 9(3), 289333.

[13] Lee, H. K. H. (2001). Model Selection for neural network classification. Journal of Classification, 18, 227243.

[14] M.H. Dunham, Data Mining Introductory and Advanced Topics. Prentice Hall,2002.

[15] Moisen, G. G., & Frescino, T. S. (2002). Comparing five modelling techniques for predicting forests characteristics. Ecological Modelling, 157, 209225.

[16] Ozdamar, K. (2004). Paket programlarla istatistiksel veri analizi1. Eskisehir: Kaan Kitabevi.

[17] Stark, K. D. C., & Pfeiffer, D. U. (1999). The application of nonparametric techniques to solve classification problems in complex data sets in veterinary epidemiologyan example. Intelligent Data Analysis, 3, 2335.

[18] van Gerven, M. A., Jurgelenaite, R., Taal, B. G., Heskes, T., and Lucas, P. J. 2007. Predicting carcinoid heart disease with the noisy-threshold classifier. Artif. Intell. Med. 40, 1 (May. 2007), 45-55.