CLASSIFICATION AND REGRESSION TREE IN PREDICTION OF SURVIVAL OF AIDS PATIENTS

Sameem Abdul Kareem, S. Raviraja*, Namir A Awadh*, Adeeba Kamaruzaman**, Annapurni Kajindran***

*Department of Artificial Intelligence,
Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia.
**Division of Infectious Diseases,
University of Malaya Medical Centre, Kuala Lumpur, Malaysia.
Email: sameem@um.edu.my; sraviraja@um.edu.my;

ABSTRACT

Over the years, the advancement in computing technology, the reliability of computers, coupled with the development of easy-to-use but nevertheless sophisticated software has led to significant changes in the way that data are collected and analyzed. Computations has shifted from off-site main frames, dependent on highly trained operators and located in special rooms accessible only to certain authorised staff, to the more accessible desktop and laptop computers. This accessibility has resulted in an increasing number of researches in data mining in which hidden predictive information are extracted from large databases, using techniques from database research, artificial intelligence and statistics, to a wide variety of domains such as finance, manufacturing and medicine. In this research we describe our experiments on the application of Classification And Regression Tree (CART) to predict the survival of AIDS. CART builds classification and regression trees for predicting continuous dependent variables and categorical or predictor variables, and by predicting the most likely value of the dependent variable. In this paper, a total of 998 patients who had been diagnosed with AIDS were grouped according to prognosis by CART. We found that CART were able to predict the survival of AIDS with an accuracy of 60-93% based on selected dependent variables, validated using Receiver Operating Characteristics (ROC). This could be useful in determining potential treatment methods and monitoring the progress of treatment for AIDS patients.

Keywords: Medical prognosis; medical decision making; ANN; AIDS survival; Medical informatics; CART;

1.0 INTRODUCTION

Prognosis - the prediction of the course and the outcome of disease processes - plays an important role in patient management tasks like diagnosis and treatment planning. Prognostic models form, therefore, an integral part of a number of systems supporting these tasks. Furthermore, prognostic models constitute instruments to evaluate the quality of healthcare and the consequences of healthcare policies by comparing predictions according to care norms with actual results. Approaches to developing prognostic models vary from using traditional probabilistic techniques, originating from the field of Statistics, to more qualitative and model-based techniques, originating from the field of Artificial Intelligence (AI). In this paper, various approaches to constructing prognostic models, with emphasis on methods from the field of AI, are described and compared.

The rapid development of medical informatics is due to advances in computing and communications technology, an increasing awareness that the knowledge base of medicine is essentially unmanageable by traditional paper-based methods, and to a growing conviction that the process of informed decision making is as important to modern biomedicine as is the collection of facts on which clinical decisions or research plans are made.

Medical prognosis is a prediction of the future course and outcome of a disease and an indication of the likelihood of recovery from that disease. Prognostic information equally important to healthcare policy makers and administrators, drug manufacturers, clinicians and patients is used to determine healthcare policies, to monitor the progress of regional and national AIDS control programmes, as a tool to assess the efficacy and progress of treatment protocols and programmes, and as an aid in choosing treatment types and methodologies. At the individual level prognostic information can help patients make informed decisions with regards to the quality of life and finance [1].

The purpose of this study is to investigate the use of CART as a tool for data mining, predictive modeling and data processing in the prognosis of AIDS. The goal of any modeling exercise is to extract as much information as possible from available data and provide an accurate representation of both the knowledge and uncertainty about the epidemic.

2.0 BACKGROUND AND RELATED WORK

In this paper we describe the researches carried out in using AI technique, namely, CART, in order to predict the survival of AIDS in the Malaysian scenario.

The analysis of survival data has long been the domain of statistics as can be observed through the number of medical statistics textbooks and journals dedicated to the field [1, 2, 3, 4]. Statistical methods such as the life-table, the Kaplan-Meier method and regression models such as the Cox Proportional Hazard are usually used to model and predict survival data with the ability to handle censored data [5, 6]. However, these methods are usually used to explain the data and to model the progression of the disease rather than to make survival predictions for populations or individual patients.

A major part of the work done in the area of survival analysis is that carried by Ohno-Machado as part of her PhD thesis. Ohno-Machado made comparisons of standard and sequential neural network models in the survival of coronary heart disease, the comparisons of neural network model with that of Cox Proportional Hazards and the use of modular neural networks to predict HIV survival [6, 7, 8, 9, 10].

In writing this paper, the literature review that was carried out on computer-based prognostic systems, namely, on the use of techniques of database research, AI and statistics, in prognostic systems has resulted in at least fifty different articles/papers published by international researchers. In the Malaysian scenario, the number of papers published does not even come close to this figure. A literature search in the National Library of Medicine database (PUBMED), for “Malaysia and AIDS” yielded 169 citations. A PUBMED search on “Malaysia and AIDS and prognosis” resulted in a mere 4 articles, while a search for “Malaysia and AIDS and prognosis and computer”, “Malaysia and AIDS and prognosis and artificial intelligence” and “Malaysia and AIDS and prognosis and neural network” yielded “0” (zero) articles [11, 12, 13]. Thus, there is a great potential for research in this area in the local scenario. In this paper we describe our research on using CART in the prediction of survival of AIDS in the Malaysian scenario.

HIV/AIDS has emerged as one of the leading challenges for global public health. The number of reported HIV cases in Malaysia has increased from a mere three (3) cases in 1986 to 6427 cases in 2004. During the period of 1986 to 2004 the total number of HIV infections in Malaysia is a staggering 64, 439 (59,962 Male, 4477 Female) of which 9442 had developed into AIDS (8596 Male, 846 Female) with a total of 7195 cases reported to be AIDS related deaths (6661, 534 Female). The ages of these HIV/AIDS related deaths vary from as young as less than 2 years of age to that of above 50 years of age [14].

Since HIV/AIDS poses a challenge to the general development of human resource in Malaysia, any research in HIV/AIDS should thus be duly encouraged. The prediction of survival is not the only research that can be carried out in the area as researchers could also investigate the association of certain risk factors to the disease, matching patients to treatment protocols and the application of AI techniques, database research and statistics to the follow up and management of AIDS patients [14].

One of the most challenging tasks in carrying out research in the clinical scenario is the availability of clinical data sets. This is especially so, here in Malaysia, where medical data banks do not currently exist. Medical data are usually only accessible to hospital authorities; even then, inter-hospital data accesses may not be possible. Most researchers will have to depend on the goodwill of clinicians, who may do their own medical record keeping, in order to access medical data. Furthermore, even if a researcher is able to access these data they may not be in an electronic form. For chronic illnesses in particular, clinical data sets may not contain an appreciable number of cases, some data collection may be incomplete or some attributes may not be well represented. Those that are complete may not be error-free. All these add to the difficulties of conducting research using real data sets as opposed to artificial data sets.

However, recently local computer scientists are beginning to make attempts in carrying out researches in medical informatics through collaborations with their medical counterparts.

Recently, various mathematical and statistical approaches have been proposed for the prediction of survival in HIV/AIDS. M. Bonarek used Kaplan-Meier method to determine prognostic factors associated with in-hospital survival in HIV-infected patients admitted to MICUs (medical intensive care units) [15]. A logistic regression model has been used in prognosis of HIV patient's survival [16], Weibull, loglogistic, lognormal distributions has been used in formulating the prognostic model for HIV survivals [17]. Cox models are applied and STATA 7 for statistical analyses is used as well in a study to determine the survival in HIV [18].

The purpose of this study is to examine the use of CART as a tool for data mining, predictive modeling and data processing in the prognosis of AIDS. The goal of any modeling exercise is to extract as much information as possible from available data and provide an accurate representation of both the knowledge and uncertainty about the epidemic.

3.0 DATA ANALYSIS

Data was from 998 patients diagnosed with HIV/AIDS and treated at the University Hospital, Kuala Lumpur from 1987 to 2007. Out of 998 patients, 832 patients were male while 166 patients were female as shown in Fig. 1A.

Fig. 1B shows that the highest prevalence of AIDS is amongst the Chinese, 583; followed by the Malays, 227; the Indians, 145; and others 43; which form the main ethnic composition of multiracial Malaysia.

The highest incidence of AIDS occurs amongst single patients, 609; followed by married patients, 323; divorced or separated patients, 53 and widowed patients, 13 as shown in Fig.1C.

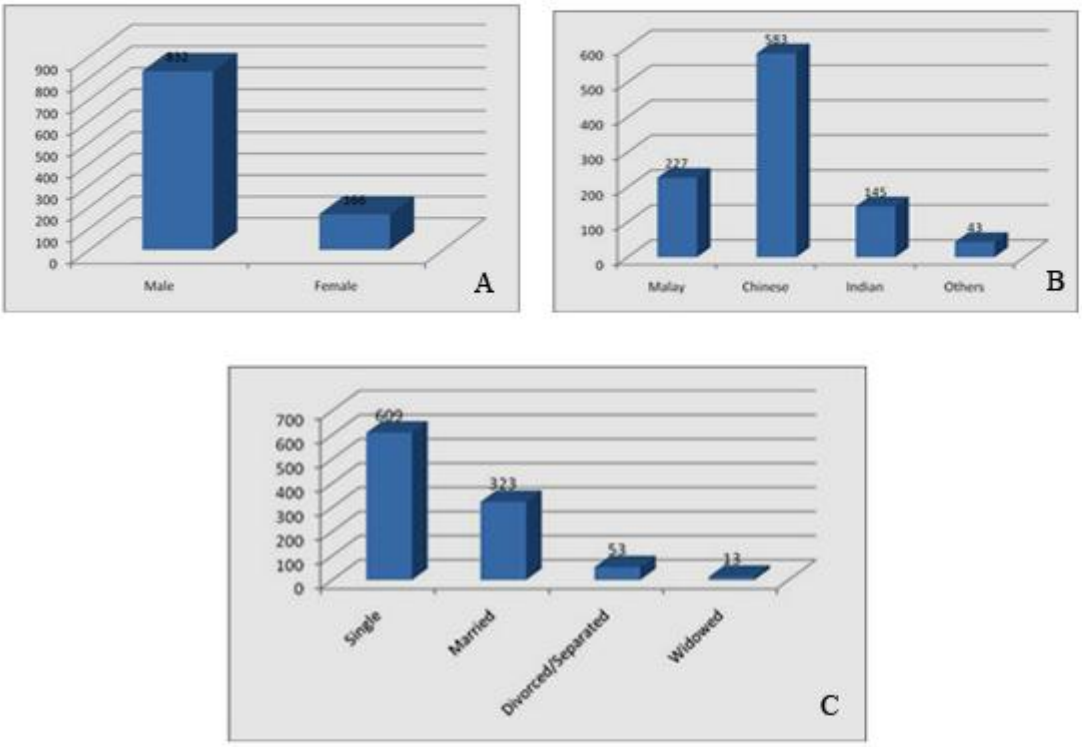

Fig 1: AIDS Data by Gender (A), AIDS Data by Ethnicity (B), AIDS Data by Marital Status (C)

4.0 METHODOLOGY

Data for patients were divided in sets according to survival in different periods after diagnosis of AIDS (1 year, 2 years, etc) as shown in Table 1. Using predefined intervals for the output makes the training of the network easier and the prediction more efficient. This is because the network is able to converge easier to outputs in the form of 1's and 0's based on a predefined threshold as opposed to a real output of, say, 9 years, especially if the input variables are also in binary form.

Table 1: Survival Intervals Status according to the standard, “1”-Dead and “0”- Alive

Surv_Yr	Surv1	Surv3	Surv5	Surv7	Surv9	Surv10
0	1	1	1	1	1	1
1	0	1	1	1	1	1
2	0	1	1	1	1	1
3	0	0	1	1	1	1
4	0	0	1	1	1	1
5	0	0	0	1	1	1
6	0	0	0	1	1	1
7	0	0	0	0	1	1
8	0	0	0	0	1	1
9	0	0	0	0	0	1
10	0	0	0	0	0	0

CART stands for Classification And Regression Trees, a decision-tree procedure representing a classification system or predictive model introduced in 1984 by statisticians, Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone [5]. The results of the decision tree are displayed as a tree diagram using a simple set of if-then rules. In a regression tree the target, or *Y* variable, is dependent on several explanatory variables (*X*s). A regression tree results from recursive partitioning of the set of *Y*-values, according to *X*-values. The end point reached determines the prediction made by the model, which in this research, is the prediction of the survival of AIDS.

CART is used in predictive modeling on a medical dataset; which consists of records for 1500 HIV patients, including demographic data such as, age code as shown in Table 6, gender coded as shown in Table 7, ethnicity coded as shown in Table 8, treatment method coded as shown in Table 9, weight and factors deemed to affect the survival of AIDS, namely, risk exposure coded as shown in Table 10, CD4, CD8 and viral load coded as shown in Table 11 and 12 respectively in Appendix.

Missing values are represented by the mean average value (eg. in Age, missing values are replaced with the Mean Age) or replaced with the most commonly occurring value (eg. in Ethnicity, missing values are replaced with the most commonly occurring Race, i.e. Chinese).

Regression trees are used for predicting the membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. The dependent (target) variable, in this case, the survival of AIDS, is numerical and takes the values of ‘1’ for ‘Dead’ and ‘0’ for ‘Alive’. In order to predict the value of the target variable (Survival) using the regression tree, the model uses the values of the predictor variables to move through the tree until it reaches a terminal (leaf) node, and then ultimately predicts the category shown for that node as in Fig 2.

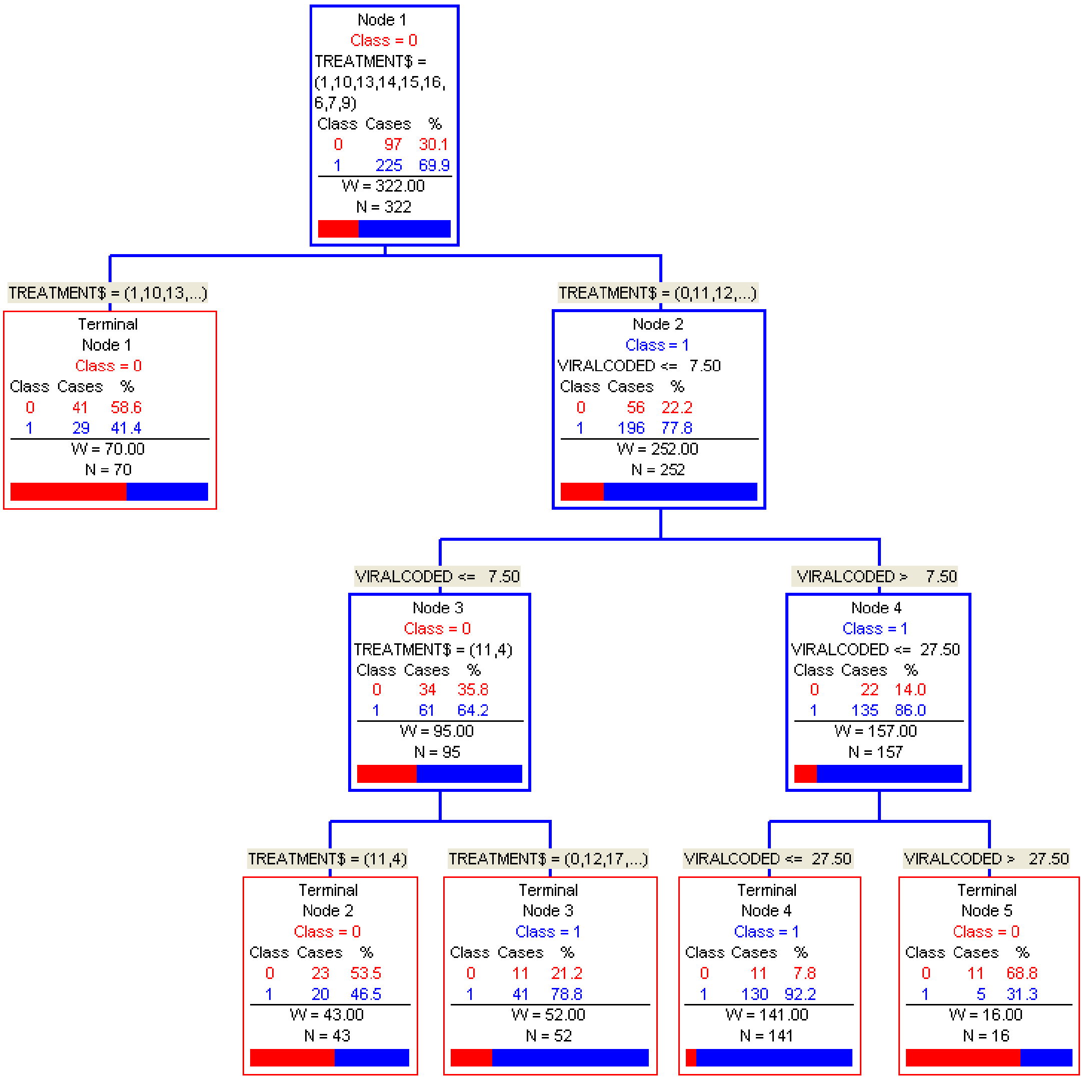

Fig 2: Part of a Regression Tree for the Prediction of Year 7 Survival

At each node the following is evaluated:

i. Is X ≤ d? If X is a ***continuous variable*** and d is a constant within the range of allowable values for X. For example, is Viral-Coded <= 7.5? (Or is coded_age ≤ 3.5?) Or
ii. Is Z = b? If Z is a ***categorical variable*** and b is one of the integer values assumed by Z. For example, is Treatment_method = (1,6,7,9,10, 13, 14, 15, 16) ? Or is Sex = 1?

The number of possible split points on each variable is limited to the number of distinct values each variable assumes in the sample. For example, if a sample size equals N, and if X is a continuous variable and assumes N distinct points in the sample, then the maximum number of split points on X is equal to N. If Z is a categorical variable with m distinct points in a sample, then the number of possible split points on Z equals $(2^{m-1} - 1)$ [1] [19]. Unless otherwise specified, CART software considers that, each split will be based on only a single variable.

5.0 RESULT AND DISCUSSION

Artificial Intelligence and database research in medicine is a relatively new research area, which combines sophisticated representational and computing techniques with the insights of expert physicians to produce tools for improving healthcare. Most existing programs in this area have been judged to be comparable to expert physicians in their competence and had produced outstanding results. This Artificial Intelligence in Medicine (AIM) systems have shown to be statistically indistinguishable from experts in the field, others were judged as giving expert advice by true experts. Although the trials have ranged in rigor from well-controlled experiments to almost anecdotal testimonials, an objective examination of their performance clearly demonstrates that they have captured an important aspect of what it means to be an expert in a particular field of medicine and provided a good demonstration of their capabilities on some significant medical cases.

5.1 Experimental Results

Several test cases are considered to test the proposed technique and is possible for technical and non-technical users to use CART easily, the CART analysis boost perception about the variables and the interactions among those variables, with the window-based interface application, the user can easily load the data file, model the data, create a decision tree, and test results. Also with CART, the user can view the variable importance which can be very important to choose the predictors. Another advantage is the simplicity of results. In most of the cases, the interpretation of results summarized in a tree is very simple to analyse and to decide.

The results are presented as binary decision tree form, which contains nodes connected by branches. Those nodes branched into two new nodes (child nodes) are called parent nodes, otherwise they are terminal nodes. Choosing the right target variable and predictor variables affects the results, for example: For Survival Year 3 and predictor variables: sex, age, ethnicity, marital status, weight, CD4 and CD8. The Tree Sequence is as shown in Table 2 and Table 3 shows the data samples for total of 322 records.

Table 2: Tree Sequence

Tree Number	Terminal Nodes	Cross-Validated Relative Cost	Re-substitution Relative Cost	Complexity
1*	12	0.87131 ± 0.15252	0.01899	0.00000
2	8	0.93776 ± 0.15311	0.07595	0.00714
3	6	0.96624 ± 0.15334	0.12658	0.01267
4	5	0.98840 ± 0.15350	0.16139	0.01742
5	4	1.00738 ± 0.15363	0.19937	0.01900
6	3	0.89135 ± 0.19388	0.26582	0.03324
7	2	0.91983 ± 0.19400	0.39451	0.06436
8	1	1.00000 ± 0.00008	1.00000	0.30275

*Optimal

Table 3: Data Sample, Total Records 322

Class	Learn	%	Total
0	6	1.86	6
1	316	98.14	316
Total:	322	100.00	322

Fig 3 shows a curve which is a relative cost profile and outlines the relationship between classification errors and tree size. The scale is always between 0 and 1, so it is called a relative error curve. 0 means no error or a perfect fit, and 1 represents the performance of random guessing.

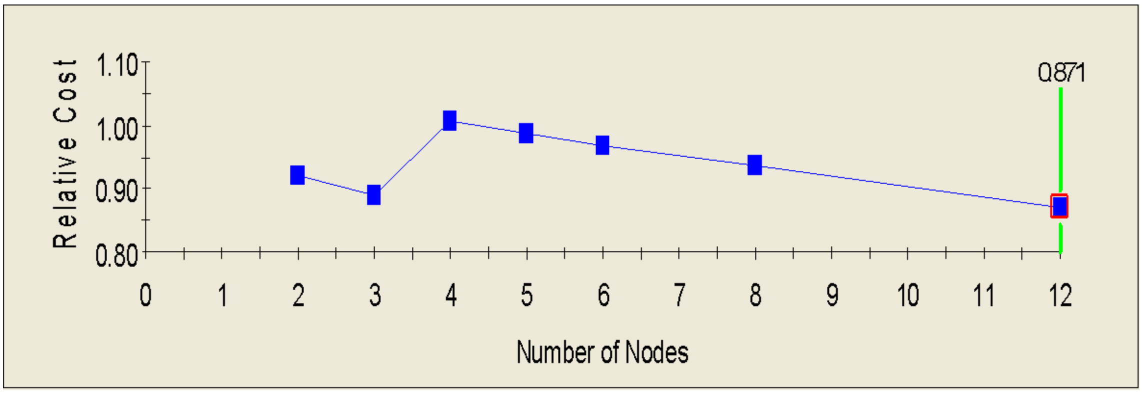

Fig 3: Error Curve

For Target Variable: Year 11 survival and Predictor Variables: age, sex, ethnicity, marital status, weight, CD4 and CD8 hit a relative error of 0.404 as demonstrated in Fig 4.

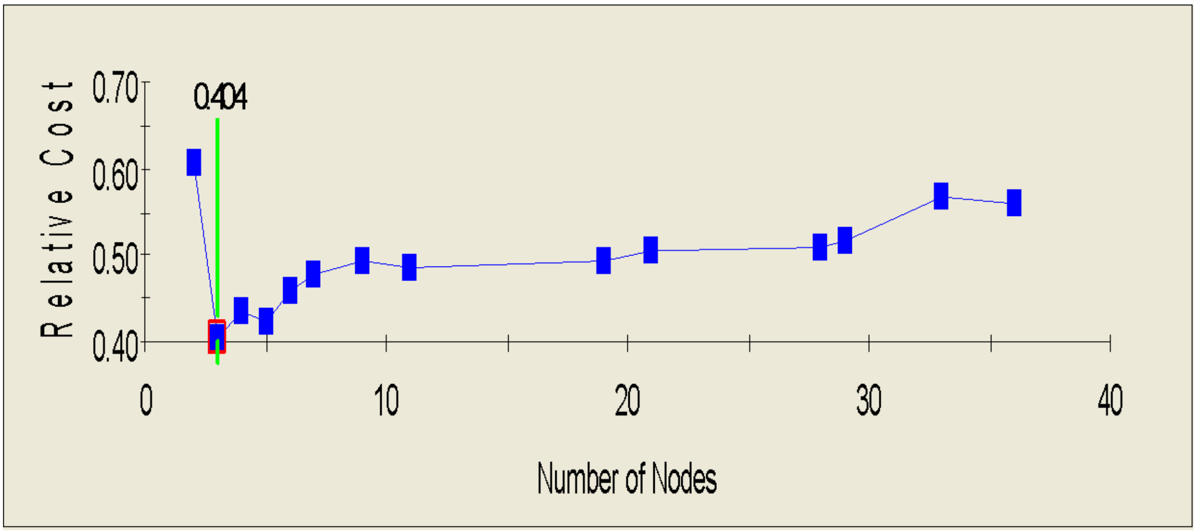

Fig 4: Error Curve for Year 11

A tree with a relative error of 0 or nearly 0 is usually too good to be true. In almost all cases this results from including an inappropriate predictor in the model. This proposed model with surv11 variable shows excellent performance with a test value of the ROC of 0.8166 and train value of ROC is 0.8209. ROC can range between 0 and 1 with higher values indicating better performance as shown in Table 4.

Table 4: ROC with different variables

Survival_Year	ROC	ROC	Overall % Correct
	Training	Testing	Testing
1	N/A	N/A	N/A
3	0.9934	0.5665	93.17
5	0.7924	0.5881	63.35
7	0.7789	0.7194	70.81
9	0.7744	0.7697	75.47
11	0.8209	0.8166	77.02
13	0.7261	0.6693	61.80
15	0.8226	0.733	71.43
17	0.8353	0.7494	74.22
19	0.8136	0.7589	72.67

5.2 Receiver Operating Characteristics and Percent Accuracy

The predictive performance of the CART and hence its generalization capability was measured in terms of the area under the receiver-operating characteristic. In medical prediction, the receiver operating characteristics (ROC) is commonly used to determine the accuracy of predicted values as it can be used across different classification tools.

The ROC is plot sensitivity against specificity for different test results values. A person who is alive and who had a “positive” test result is termed a true positive, whereas a person who is alive but a “negative” test is termed a false negative. On the other hand, a person who died but had a “positive” result is termed a false positive, while person who died and had a “negative” test is termed a true negative [20, 21]. This is summarized in Table 5.

Table 5: The Definition of True Positives/Negatives

	Alive (+)	Dead (-)
Result Positive	A = true positive	B = false positive
Result Negative	C = false negative	D = True Negative

Sensitivity, equation (1) is the true positive test results divided by all the living patients. This is the probability that a patient will be classified as alive when she is alive.

$$\text{Sensitivity} = \quad (a / (a + c)) \ldots\ldots. (1)$$

The specificity, equation (2) of a test is the true-negative test results divided by all the dead patients. This is the probability that a patient will be classified as dead when she is dead. “1-specificity” is the probability that a patient will be classified as alive when she is dead.

$$\text{Specificity} = (d / (b + d)) \ldots\ldots. (2)$$

To generate the ROC curve it is first necessary to determine the sensitivity and specificity for each test result. The X-axis ranges from 0 to 1, or 0% to 100% and is the false positive rate, that is 1-specificity. The Y-axis ranges from 0 to 1, or 0% to 100% and is the true positive rate, that is the sensitivity. The curve starts at (0,0) and increases towards (1,1). The endpoints of the curve will run to these points and an area of the resulting trapezoids can therefore be calculated as shown in Fig. 5. The larger the area under the curve the better is the prediction.

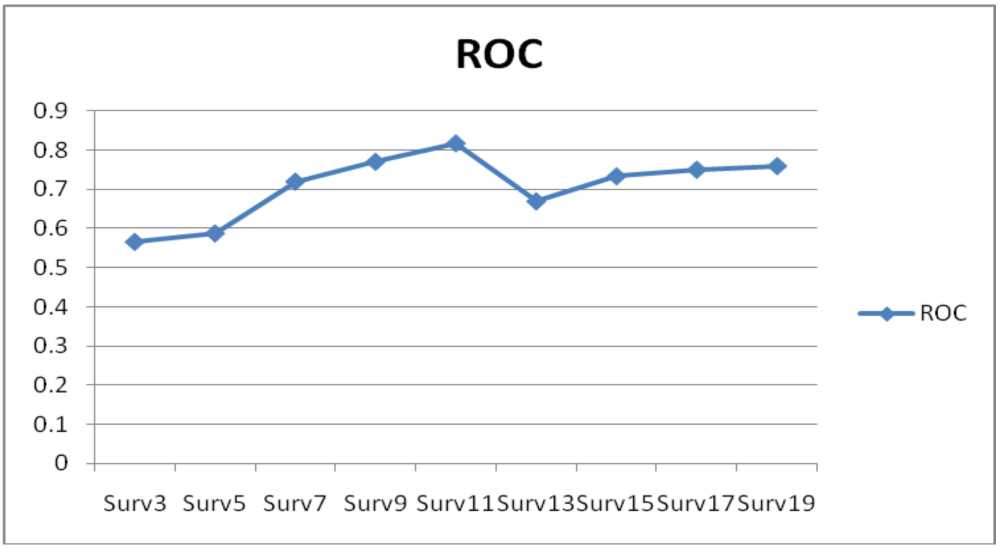

Fig 5: CART Prediction in Terms of ROC

The accuracy of the predictions is measured by the number of correctly predicted cases divided by all the cases in the study (Percent Accuracy) Fig 6.

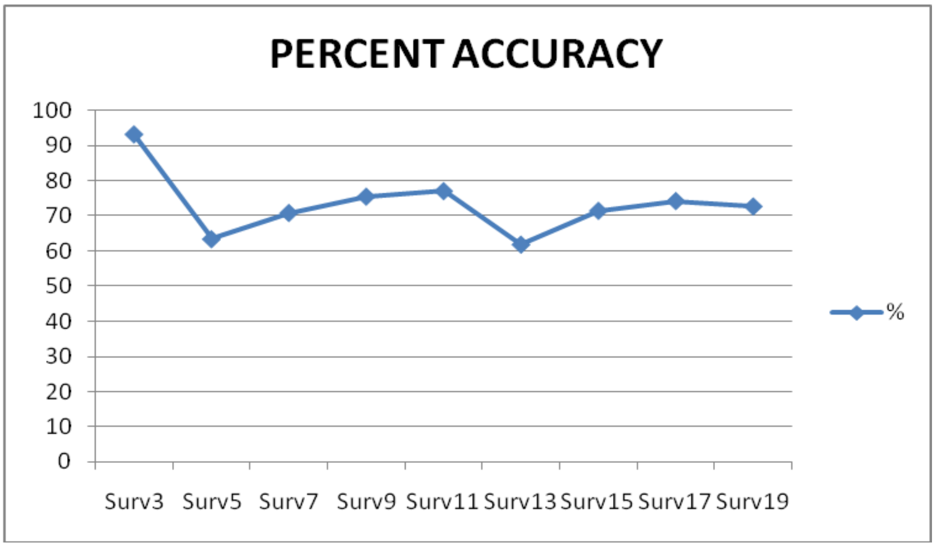

Fig 6: CART Prediction in Terms of Percent accuracy

We found that CART was able to predict the survival of AIDS with an accuracy of 60-93% based on selected dependent variables, validated using ROC. This proposed model for Year 11 survival shows excellent performance with a test value of 0.8166 in terms of ROC and a percent accuracy of 77% and a train value of 0.8209.

6.0 CONCLUSION

CART analysis is a robust, analytical technique; it can provide a means of understanding huge clinical data. However, it can produce useful results using only a few important variables and interpret the results. On the other hand, the use of CART has been growing and may increase in the future, mostly because of the extensive number of important problems for which it is the best available solution.

It has been seen that the CART was able to predict the survival of AIDS with an accuracy of 60%-93% based on selected dependent variables, validated using ROC. The results of this model are significant, this approach experimented and the results obtained in this research could be useful in determining potential treatment methods and monitoring the progress of treatment for AIDS patients.

In medical prognosis, many projects have been shown to be significantly superior to statistical methods. Research efforts in this area ranges from the prediction of the survival of coronary heart disease to HIV survival prediction. A large part of the work in AIM is in the prediction of cancer. The researches on survival analysis using neural network are mostly in the domain of breast cancer, ovarian cancer, bladder cancer and prostate cancer.

As a means of comparisons between neural network models and statistical models we have carried out some analysis using Kaplan-Meier models and the Cox proportional hazards method. The Kaplan-Meier analysis on the NPC data set has enabled us to make survival predictions for groups of patients and to plot their survival curves [21, 22, 23]. Although the Kaplan Meier method is a good nonparametric model of survival data, it is limited in its prognostic prediction. Nonparametric models are usually used to describe the survival of a group of patients and to explain the data rather than to make individual predictions of survival.

Statistical models are usually used for large groups of people based on estimates taken from a sample and are therefore meaningless for an individual. Although there are attempts in statistical tools such SPSS to provide predictions for individual cases, these predictions far from accurate and are affected by the number of similar cases in the sample. As an example if there is only a small number, say, three cases, with a survival of eleven years, the prediction is different if there are, say, fifty such cases.

It has been established that as universal access to treatment for HIV-infected persons has improved even in resource-poor settings, other co-morbidities are becoming important determinants of survival in patients with AIDS. Further we would like to consider including more input variables such as HIV disease stage, co-infection with hepatitis B and C and systemic conditions like diabetes and hypertension to improve the accuracy of prediction in our further research which we have not captured in our model, at present.

One of the greatest difficulties in carrying out research in the clinical scenario is in obtaining clinical data. Thus, collaboration between clinicians and computer scientists is imperative for such research efforts to be a success. On the local scenario there have been published works on the use of neural networks in NPC, and ANN prognostic model for AIDS by Abdul-Kareem S et al, The number of papers published in this field by Malaysian researches does not even come close to that published by their Western counterparts. Thus, there is still room for research in this area in the Malaysian scenario.

Acknowledgements

We have been fortunate in our collaborative efforts with clinicians in the University of Malaya Medical Centre (UMMC) who not only shared their vast experience in AIDS and by playing the role of scientific advisor but also provided the means to access UMMC's data.

REFERENCES

[1] Breiman, L., J. H. Fried man, R. A. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Monterey, Calif., U.S.A.: Wadsworth, Inc.

[2] Collet, D., *Modeling Survival Data in Medical Research, ed*. C. Chatfield, J.V. Zidek. London: Chapman & Hall,1994.

[3] Elandt-Johnson, R.C., Johnson, N.L., *Survival Models and Data Analysis*. John Wiley & Sons, pp.50-83,1976.

[4] Lee, E.T., *Statistical Methods for Survival Analysis*, Lifetime Learning Publications, pp 9-18, California, 1980.

[5] CART FAQ http://clearinghouse.wayne.edu/oldsite/downloads/CARTFAQ.pdf (Accessed, May 2009)

[6] Swinscow, T.D.V., 1999, "*Survival analysis*", Statistics at Square One, *British Medical Journal* (electronic version. Available:http://www.bmj.com/collections/statsbk/12.html

[7] Ohno-Machado, L., "Neural network techniques: utilization in medical prognosis", Electronic Copy of Review. [Emailed by Ohno-Machado],1999.

[8] Ohno-Machado, L., Musen, M.A., "Sequential versus standard neural networks for temporal pattern recognition: an example using the domain of coronary heart disease", Stanford University. Technical Report. School of Medicine. Medical Computer Science. Knowledge Systems Laboratory. 1996.

[9] Ohno-Machado, L., 1994, "Identification of low frequency patterns in backpropagation neural network", Technical Report, Section in Medical Informatics, Stanford University.

[10] Ohno-Machado, L., Walker, M.G., Musen, M.A., 1994, "Hierarchical neural networks for survival analysis", Technical Report, Section in Medical Informatics, Stanford University.

[11] Tan DB, Yong YK, Tan HY, Kamarulzaman A, Tan LH, Lim A, James I, French M, Price P. Immunological profiles of immune restoration disease presenting as mycobacterial lymphadenitis and cryptococcal meningitis. HIV Med. 2008 May;9(5):307-16.

[12] Fadzelly AB, Asmah R, Fauziah O. Effects of Strobilanthes crispus tea aqueous extracts on glucose and lipid profile in normal and streptozotocin-induced hyperglycemic rats. Plant Foods Hum Nutr. 2006 Mar;61(1):7-12.

[13] Mohammad Z, Naing NN. Characteristics of HIV-infected tuberculosis patients in Kota Bharu Hospital, Kelantan from 1998 to 2001. Southeast Asian J Trop Med Public Health. 2004 Mar;35(1):140-3.

[14] United Nations General Assembly Special Session on HIV/AIDS, December 2005, "Monitoring the Declaration of Commitment on HIV/AIDS" Country Report MALAYSIA.

[15] M. Bonarek, "Prognostic score of short-term survival in HIV-infected patients admitted to medical intensive care units (MICUs)". International Conference AID; abstract no: MoPeB2183, July 2000.

[16] Hanson DL; Horsburgh CR Jr; Fann SA; Havlik JA; Thompson SE 3d; "Survival prognosis of HIV-infected patients", Division of HIV/AIDS, Centers for Disease Control, Atlanta,; Georgia 30333. AIDSLINE MED/93267399. Jun 1993.

[17] Matthias Egger, Margaret May, Geneviève Chêne, Andrew N Phillips, Bruno Ledergerber, François Dabis, Dominique Costagliola, Antonella D'Arminio Monforte, Frank de Wolf, Peter Reiss, Jens D Lundgren, Amy C Justice, Schlomo Staszewski, Catherine Leport, Robert S Hogg, Caroline A Sabin, M John Gill, Bernd Salzberger, Jonathan A C Sterne, and the ART Cohort Collaboration," Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies". THE LANCET • Vol 360 • July 13, 2002.

[18] K Porter, A G Babiker, J H Darbyshire, P Pezzotti, K Bhaskaran, A S Walker, "Determinants of survival following HIV-1 seroconversion after the introduction of HAART". THE LANCET • Vol 362 ,October 2003.

[19] Ohno-Machado, L.,“Medical applications of neural networks: connectionist models of survival”. PhD Thesis, Section in Medical Informatics, Stanford University. Palo Alto, California, 1996.

[20] Norazmi MN, Suarn S. Disparity in the percentage of CD4+ T lymphocytes and prognosis of HIV-infected intravenous drug users in Malaysia. Immunol Lett. 1994 Dec;43(3):177-82.

[21] Sameem Abdul Kareem, “Application Of Artificial Neural Network For The Prognosis Of Nasopharyngeal Carcinoma”, Ph.D. Thesis, University of Malaya, 2002.

[22] Seong Ho Park, Jin Mo Goo, Chan-Hee Jo,Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. Korean J Radiol 5(1), March 2004, pp 11-18.

[23] Sameem Abdul Kareem, Sapiyan Baba, Yong Zulina Zubairi, Mohd Ibrahim A Wahid, “Prognostic Systems For NPC: A Comparison Of The Multi Layer Perceptron Model And The Recurrent Model”. Proceedings of The 9th International Conference on Neural Information Processing 18th –22nd November, Singapore, 2002.

Appendix :

Table 6: Age Data Codes

Age / Data	Code
0-10	0
11-20	1
21-30	2
31-40	3
41-50	4
51-60	5
61-70	6
71-80	7
81-90	8
Missing age*	*

* Replace with average age

Table 7: Gender Data Codes

Data	Code
Male	0
Female	1

Table 8: Ethnicity Data Codes

Data	Code
Malay	0
Chinese	1
Indian	2
Others	3
Missing ethnicity*	1

* Replace most Common i.e., Chinese

Table 9: Treatment

Data	Code
BMS232632/Placebo	0
Combivir	1
Didanoside(ddI) 100mg-	2
Didanosine (videx)	3
Efavirenz (DMP266)(Sustiva)(EFV)	4
Efavirenz/Placebo	5
Indinavir(Crixivan)(IDV)	6
Kaletra (lopinavir/rit)	7
Lamivudine (3TC)(Epivir)	8
Nevirapine GENERIC	9
Nevirapine(Viramune) (NVP)	10
Ritonavir (full dose)	11
Ritonavir (low dose)	12
Stavudine (d4T) 15mg	13
Stavudine GENERIC	14
Stavudine(d4T) 30mg	15
Stavudine(d4T) 40mg	16
Stavudine(Zerit)	17
TMC-114	18
Zidovudine (ZDV)(AZT) 250mg	19
Zidovudine(Retrovir)	20
Zidovudine?/(ZDV)(AZT) 100mg	21
Missing Treatment*	4

* Missing Treatment replaced with 4, Efavirenz (DMP266)(Sustiva)(EFV)

Table 10: Exposed Risk Data Codes

Data	Code
Heterosexual	0
Homosexual	1
Blood	2
IDU	3
Bisexual	4
Other	5
Mother	6
Missing exposure*	0

* Exposure replaced with 0-heterosexual

Table 11: CD4, CD8

Data	Code
115	1
219	2
356	3
453	4

Table 12: Viral Code

Data	Code
490	0
15000	1
22900	2
35600	3

BIOGRAPHY

Asso Prof. Datin. Dr. Sameem Abdul Kareem: received the B.Sc. in Mathematics (Hons) from University of Malaya, in 1986, the M.Sc. in Computing from the University of Wales, Cardiff, UK, in 1992, and the Ph.D. in computer science from University of Malaya, Malaysia, in 2002. She is currently an Associate Professor of the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University Of Malaya. Her research interests include medical informatics, information retrieval, data mining and intelligent techniques. She has published over 80 journal and conference papers.

Dr. S. Raviraja: received his B.Sc Computer Science and Masters in Computer Applications from University of Mysore, India and in 2004 PhD in Computer Science from University of Honolulu, USA. Started his career with Motorola (India) as Software Engineer, later as Software Analyst and then as Project lead in reputed software companies in India. He was working as a research scholar and later as Assistant Professor in Unity University in Ethiopia then continued academic and research career with one of the University of Medical Sciences & Technology in Sudan. He joined as a Post Doctoral Fellow and currently his position is a Visiting Sr. Lecturer in Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya since 2008, also Editorial Member of in few Journals of Computer Science. He is a member of Institute of Engineers India, Computer society of India and several other professional associations. His research interest includes Medical Image Analysis, DIP, AI & Robotics and in Software Engineering Methodologies.

Namir A Awadh: Namir A Awadh, MSc student, Department of Artificial Intelligence, Faculty of Computer Science and Information Technology

Dr. Adeeba Kamaruzaman: is a professor and director of Infectious Diseases Unit, Faculty of Medicine, University of Malaya Medical Centre. Also currently she is the president of Malaysian AIDS Council.

Annapurni Kajindran: is a statistician and responsible for the AIDS data management in office of Infectious Diseases Unit, Faculty of Medicine, University of Malaya Medical Centre.