Sains Malaysiana 43(4)(2014): 567–573

A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers

(Model Hibrid untuk Prognosis Kanser Mulut berdasarkan kepada Penanda Klinikopatologi dan Genomik)

SIOW-WEE CHANG*, SAMEEM ABDUL KAREEM, AMIR FEISAL MERICAN ALJUNID MERICAN & ROSNAH BINTI ZAIN

ABSTRACT

There are very few prognostic studies that combine both clinicopathologic and genomic data. Most of the studies use only clinicopathologic factors without taking into consideration the tumour biology and molecular information, while some studies use genomic markers or microarray information only without the clinicopathologic parameters. Thus, these studies may not be able to prognoses a patient effectively. Previous studies have shown that prognosis results are more accurate when using both clinicopathologic and genomic data. The objectives of this research were to apply hybrid artificial intelligent techniques in the prognosis of oral cancer based on the correlation of clinicopathologic and genomic markers and to prove that the prognosis is better with both markers. The proposed hybrid model consisting of two stages, where stage one with ReliefF-GA feature selection method to find an optimal feature of subset and stage two with ANFIS classification to classify either the patients alive or dead after certain years of diagnosis. The proposed prognostic model was experimented on two groups of oral cancer dataset collected locally here in Malaysia, Group 1 with clinicopathologic markers only and Group 2 with both clinicopathologic and genomic markers. The results proved that the proposed model with optimum features selected is more accurate with the use of both clinicopathologic and genomic markers and outperformed the other methods of artificial neural network, support vector machine and logistic regression. This prognostic model is feasible to aid the clinicians in the decision support stage and to identify the high risk markers to better predict the survival rate for each oral cancer patient.

Keywords: ANFIS; clinicopathologic; genomic; oral cancer prognosis; ReliefF-GA

ABSTRAK

Terdapat kurang kajian yang memaparkan penyelidikan prognostik yang menggabungkan kedua-dua klinikopatologi dan genomik. Kebanyakan kajian hanya menggunakan faktor klinikopatologi tanpa mengambil kira biologi tumor dan maklumat molekul, manakala beberapa kajian penyelidik yang lain menggunakan penanda genomik atau maklumat mikroarai sahaja tanpa menggunakan parameter klinikopatologi. Maka, kajian ini tidak dapat membuat prognosis pesakit dengan berkesan. Kajian terdahulu telah menunjukkan bahawa keputusan prognosis adalah lebih tepat dengan menggunakan kedua-dua klinikopatologi dan genomik. Tujuan utama kajian ini adalah untuk mengaplikasikan hibrid teknik kepintaran buatan dalam prognosis kanser mulut berdasarkan kepada korelasi penanda klinikopatologi dan genomik dan untuk membuktikan bahawa prognosis adalah lebih baik dengan kedua-dua penanda. Model hibrid yang dicadangkan terdiri daripada dua peringkat, dengan peringkat pertama terdiri daripada ReliefF-GA sebagai kaedah pemilihan untuk mencari ciri optimum subset dan peringkat dua dengan pengelasan ANFIS untuk mengelaskan sama ada pesakit hidup atau mati selepas beberapa tahun didiagnosis. Model ramalan prognostik yang dicadangkan telah diaplikasikan ke atas dua golongan dataset kanser mulut yang dikumpulkan di Malaysia, iaitu Kumpulan 1 dengan penanda klinikopatologi sahaja dan Kumpulan 2 dengan gabungan kedua-dua penanda klinikopatologi dan genomik. Keputusan yang didapati telah membuktikan bahawa model yang dicadangkan dengan ciri optimum yang dipilih adalah lebih tepat dengan kehadiran kedua-dua penanda klinikopatologi dan genomik dan mengatasi kaedah lain seperti rangkaian saraf buatan, mesin sokongan vektor dan regresi logistik. Model prognostik ini boleh dilaksanakan untuk memberi bantuan kepada pakar klinikal di peringkat membuat sokongan keputusan untuk mengenal pasti penanda risiko yang tinggi supaya dapat meramalkan kadar jangka hayat setiap pesakit kanser dengan lebih tepat.

Kata kunci: ANFIS; genomik; klinikopatologi; prognosis kanser mulut; ReliefF-GA

INTRODUCTION

Artificial intelligent (AI) techniques are suitable to use for diagnosis or prognosis of cancer research as they are good for handling noisy and incomplete data and significant results can be attained with small sample size. In the studies of Dom et al. (2008), Kawazu et al. (2003), Li et al. (2007) and Seker et al. (2003), AI techniques have been proven to generate more accurate predictions than the statistical methods.

From the literature, many studies used clinicopathologic factors only for the cancer prognosis, while some studies utilized the genomic markers or microarray information only without the clinicopathologic parameters. Thus, these studies may not be able to predict the diagnosis or prognosis of patient effectively. In order to make a more accurate prognosis, clinician needs to include both clinicopathologic and genomic markers. It has been proven by Catto et al. (2006), Futschik and Sullivan (2003), Gevaert et al. (2006), Oliveira et al. (2008), Seker et al. (2003) and Sun et al. (2007) that prognosis results are more accurate when using both clinicopathologic and genomic data.

In this research, a hybrid AI techniques model, ReliefF-GA-ANFIS (ReliefF-Genetic Algorithm) is proposed and developed to determine the oral cancer prognosis based on clinicopathologic and genomic markers. The objectives of this research were to apply hybrid AI techniques in the prognosis of oral cancer based on clinicopathologic and genomic markers and to prove that the prognosis is better with both markers. The proposed model is experimented on the existing oral cancer dataset collected locally in Malaysia and these dataset are divided into two different groups, Group 1 with clinicopathologic markers and Group 2 with clinicopathologic and genomic markers. The proposed model consists of two stages, stage one with ReliefF-GA feature selection to find an optimum feature of subset and stage two with Adaptive Neuro-Fuzzy Inference System (ANFIS) classification to classify either the patients are alive or dead after three years of diagnosis. The proposed model is validated with two AI models, which are artificial neural network and support vector machine and a statistical model of logistic regression.

Materials and Methods

Oral Cancer Prognosis Dataset

Two types of data are used for developing the oral cancer prognosis model; these are clinicopathologic data and genomic data. Both types of data are collected from the Malaysian Oral Cancer Database and Tissue Bank System (MOCDTBS) coordinated by the Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya. Thirty-one oral cancer cases have been selected based on the completeness of the clinicopathologic data.

Clinicopathologic Marker The selected cases are based on the oral cancer cases seen in Faculty of Dentistry, University of Malaya and Hospital Tunku Ampuan Rahimah, Klang, a Malaysian government hospital from the year 2003 to 2007. All the cases selected are diagnosed as oral squamous cell carcinomas (OSCC). Based on the review from the literature and discussions with oral cancer experts from OCRCC, 15 key clinicopathologic variables have been identified as important prognostic factors of oral cancer. Table 1 lists out the selected 15 clinicopathologic variables.

TABLE 1. The selected 15 clinicopathologic variables

	Name	Description
1	Age	Age at diagnosis
2	Eth	Ethnicity
3	Gen	Gender
4	Smoke	Smoking habit
5	Drink	Alcohol drinking habit
6	Chew	Quid chewing habit
7	Site	Primary site of tumor
8	Subtype	Subtype and differentiation for SCC*
9	Inv	Depth of invasion front
10	Node	Neck nodes
11	PT	Pathological tumor staging
12	PN	Pathological lymph nodes
13	Stage	Overall stage
14	Size	Size of tumor
15	Treat	Type of treatment

*SCC = Squamous cell carcinoma

Genomic Marker In this research, two genomic markers are selected, both are tumour suppressor genes, namely *p53* and *p63*. The selection of genomic markers is based on the literature studies and discussions with the oral pathologists from Department of Oral Pathology and Oral Medicine, Faculty of Dentistry, University of Malaya. *p53* is the most frequently associated marker in the head and neck cancers (Mehrotra & Yadav 2006; Oliveira et al. 2008). *p53* is being named as 'the guardian of the genome', having important role in maintaining genomic stability, cell cycle progression, cellular differentiation, DNA repair and apoptosis. Meanwhile, *p63* is a homolog of *p53* and located in chromosome *3q21-29* and it is found to be associated with prognostic outcome in oral cancer (Muzio et al. 2005; Thurfjell et al. 2005).

In this research, the oral cancer dataset is divided into two different groups, Group 1 with 15 clinicopathologic variables only and Group 2 with 15 clinicopathologic variables and 2 genomic variables. The oral cancer 3-year prognosis dataset is used in this experiment.

ReliefF-GA Feature Selection

Feature selection is used to select the inputs which are most significant in the modeling process, in order to produce more accurate outputs. The purpose of feature selection is to reduce the number of inputs in the modeling process, but retain the accuracy of the outputs if compared to the full-input model. In this study, a hybrid feature selection of ReliefF-GA approach is proposed. This approach consists of two steps: First, it is a filter approach of Relief-F (Kononenko 1994) and second, it is a wrapper approach of genetic algorithm (GA). In the first step, each feature input is ranked and weighted using *k*-nearest neighbours classification, in which $k=1$. The top 10 features with large positive weights are selected and feed into the second stage of GA approach.

In the second step, a GA algorithm for the oral cancer prognosis dataset is proposed (Siow-Wee et al. 2011). The solutions of the GA will form the clinicopathologic or genomic variables that will subsequently be used in the oral cancer prognosis and the output will indicate how well the solutions can predict the oral cancer survival. Finally, a best solution is selected. The pseudo-code of the proposed GA is listed as in Figure 1 and is repeated for *n*-input model with *n* is the optimal number of input with lowest error rates. The increase in the number of *n* will increase the chances of over fitting problems.

In the feature subset selection problem, a solution is specific feature subset that can be encoded as a string of *n* binary digits (bits). Each feature is represented by binary digits of *1* or *0*. If a bit is equal to *1*, the feature is selected; consequently, if a bit is equal to *0*, the feature is not selected. For example, in the oral cancer prognosis dataset, if the solution is *011 001 000 010 000 00* strings of 17 binary digits, it indicates that features 2, 3, 6, and 11 are selected as the feature subset.

The initial population is generated randomly to select a subset of variables (solutions). If the variables are all different, the subset is included in the initial population. If not, it generates again until an initial population with desired size has been created.

The fitness function is used to classify between two groups, which are alive and dead. The error rate of the classification will be calculated using a 10-fold cross-validation. The fitness function is the final error rate obtained. The subset of variables with the lowest error rate will be selected. The roulette wheel selection is used in this research together with the scattered crossover and uniform mutation (Siow-Wee et al. 2011). A stopping criterion of 100 generations or a time limit of 600 s was used.

ANFIS CLASSIFIER

Next, the ANFIS classifier is implemented on the dataset with optimum feature subset generated from the ReliefF-GA feature selection method. In the input membership, the number of membership function is define by m_i, with $i = 2, 3, 4$. The rules generated are based on the number of input and the number of input membership functions, and it is represented as $(m_2^{n1} \times m_3^{n2} \times m_4^{n3})$ rules, in which n^1, n^2, and n^3 represent the number of input with m_i membership functions, respectively and $n^1 + n^2 + n^3 = n$. The type of membership function used is Gaussian and the number and name of membership functions for each input variable are shown in Table 2.

The rules generated are the output membership functions which will be computed as the summation of contribution from each rule towards the overall output. The output is the survival condition, either alive or dead after 3 years of diagnosis. The output is set as 1 for dead and -1 for alive; the psedo-code is:

```
if output >= 0
    then set output = 1, classify as dead
else output < 0,
    then set output = -1, classify as alive
```

Each ANFIS was run for 10 epochs. A 5-fold cross-validation is implemented on the dataset in which the 31 samples of oral cancer prognosis data are divided into 5 subsets of equal size and trained for 5 times, each time leaving out a sample for validation data.

PERFORMANCE MEASURES

In a medical prognosis problem, a person with positive condition (alive) who is predicted as alive is termed as true positive (TP), whereas a person with positive condition (alive) who is predicted as negative is termed as false negative (FN). On the other hand, a person with negative condition (dead) who is predicted as positive is termed as false positive (FP), while a person with negative condition (dead) who is predicted as negative is termed as true negative (TN). Table 3 lists the confusion matrix for oral cancer prognosis.

```
While selecting initial population with n-input
Generate initial population randomly without repetition variables
End while
Evaluate the fitness function of each individual using classification error rate
estimated using 10-fold cross-validation
While stopping criteria not exceeded
    Select parents from the population
    Perform crossover operation
    Perform mutation operation
    Evaluate the fitness function using classification error rate estimated using
    10-fold cross-validation
    Replace the fittest individual
End while
Return the best solution for n-input model
```

FIGURE 1. Pseudo-code for the proposed GA

TABLE 2. Membership functions for each input variable

Name	No. of membership functions	Name of Membership function
Age	4	40-50, >50-60, >60-70,>70
Eth	3	Malay, Chinese, Indian
Gen	2	Male, Female
Smoke	2	Yes, No
Drink	2	Yes, No
Chew	2	Yes, No
Site	4	Buccal mucosa, tongue, floor, others
Subtype	3	Well differentiated, moderate differentiated, poorly differentiated
Inv	2	Non-cohesive, cohesive
Node	2	Negative, positive
PT	4	T1, T2, T3, T4
PN	4	N0, N1, N2A, N2B
Stage	4	I, II, III, IV
Size	4	0-2cm, >2-4cm, >4-6cm, >6cm
Treat	3	Surgery only, Surgery+Radiotherapy, Surgery+Chemotherapy
p53	2	Negative, positive
p63	2	Negative, positive

TABLE 3. Confusion matrix for oral cancer prognosis

		Actual conditions	
		Alive (Positive)	Dead (Negative)
Predicted outcomes	Alive (Positive)	True positive (TP)	False positive (FP)
	Dead (Negative)	False negative (FN)	True negative (TN)

The measures used in this research are accuracy, sensitivity, specificity and receiver operating characteristic (ROC) curve. The performance of the model is defined as the area under the ROC curve (AUC). Accuracy is the proportion of true results in the samples, the higher the accuracy, the better the model is. Sensitivity is the true positive conditions divided by all the living patients. The specificity is the true negative conditions divided by all the dead patients. The ROC curve is a plot of *sensitivity* versus (*1 - specificity*) for different test results. The area calculated under the ROC curve is termed as area under curve (AUC).

Results and Discussion

The optimum feature subset for each n (n = 3 to 7) is obtained using the proposed ReliefF-GA feature selection method. The optimum features for each n-input are selected and the features are listed as in Table 4 for both Groups 1 and 2. As shown in Table 4, there are some similarities between the features selected for both groups, for example, in the 3-input model, there is a common feature for both groups which is *Inv* and in the 4-input model, there are two common features which are *Dri* and *Inv*. Obviously, the features selected for all the n-input models in Group 2 are the combinations of both clinicopathologic and genomic variables.

In the second stage of the proposed model, the selected features for each n-input model was fed into the ANFIS classifier for oral cancer prognosis, to classify either the patients are alive or dead after three years of diagnosis. A 5-fold cross-validation was implemented on all the n-input models. The results are shown in Table 5. For Group 1, it is noted that the classification accuracy is the highest for the 5-input model with the accuracy of 67.62% and AUC of 0.59. As regards to Group 2, the highest classification accuracy is achieved by the 3-input model and the 4-input model with the accuracy of 93.81% and AUC of 0.90.

VALIDATION TESTING

For the validation purpose, the proposed model was compared with two other common artificial intelligent models, which are artificial neural network (ANN) and support vector machine (SVM) and a statistical model of logistic regression (LR). The results are shown in Table 6, and the graphs are depicted in Figure 2.

For Group 1, as shown in Figure 2(a), ANFIS achieved the highest classification accuracy for 5-input model (accuracy=67.62%, AUC=0.59) if compared to ANN, SVM and LR. However, the difference is not significant. As for Group 2 as shown in Figure 2(b), ANFIS outperformed the others for both of 3 and 4-input models with an accuracy of 93.81% and AUC of 0.90. This is followed by ANN

TABLE 4. Features selected for each *n*-input model

	Group 1	Group 2
3-input	Gen, Inv, Node	Dri, Inv, *p63*
4-input	Gen, Dri, Inv, Node	Dri, Inv, Tre, *p63*
5-input	Gen, Dri, Inv, Node, PT	Age, Gen, Smo, Dri, *p63*
6-input	Eth, Gen, Dri, Inv, Node, PT	Age, Gen, Smo, Dri, Inv, *p63*
7-input	Age, Eth, Gen, Smo, Dri, Node, Tre	Age, Eth, Inv, Sta, Tre, p53, *p63*

TABLE 5. ANFIS classification results for *n*-input models

Model	3-input		4-input		5-input		6-input		7-input	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Group 1	67.14	0.55	60.48	0.59	67.62	0.59	51.90	0.47	64.76	0.57
Group 2	93.81	0.90	93.81	0.90	65.71	0.63	64.76	0.62	68.10	0.67

AUC - Area under ROC curve

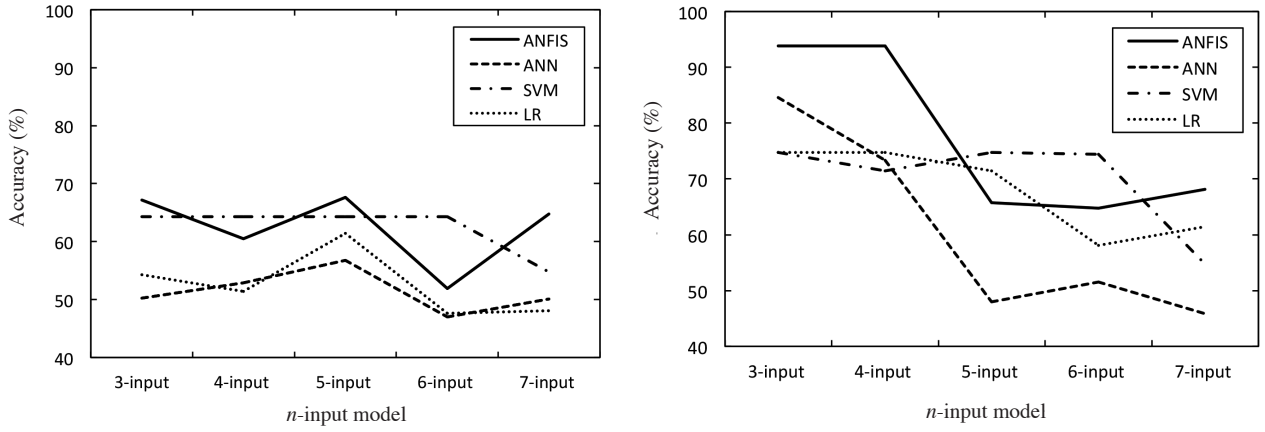

FIGURE 2. Classification results for ANFIS, ANN, SVM and LR for (a) Group 1 and (b) Group 2

classification for the 3-input model with an accuracy of 84.62% and AUC of 0.83. The high accuracy results achieved by ANFIS showed that ANFIS is a suitable classifier for the small sample size if compare to ANN, SVM and LR. The results in Group 2 also showed that higher accuracy is achieved with 3 and 4-input model. The accuracy dropped as the number of inputs increased.

Lastly, the 3-input model for Group 2 with selected features of *Dri, Inv* and *p63* (Table 4) are tested on the oral cancer 1 and 2-year prognosis dataset and the results are very promising with the accuracy for 1-year prognosis is 93.33% and 2-year prognosis is 84.29% as compared with the 3-year prognosis of 93.81%, the results are shown in Table 7.

Since there are two models with the same accuracy, hence, the simpler one is chosen, which is the 3-input model from Group 2 and the optimum subset of features are *Drink, Invasion* and *p63*. These findings confirmed some of the previous studies, which have proved that these features are important prognosis factor for oral cancer survival. Alcohol consumption has always been considered a risk factor and one of the reasons for poor prognosis of oral cancer (Asakage et al. 1998; Jefferies & Foulkes 2001; Leite & Koifman 1998; Reichart 2001; Zain et al. 2001). In Walker et al. (2003), they have shown that the depth of invasion is one of the most important predictors of lymph node metastasis in tongue cancer and in the different research done by Asakage et al. (1998); Giacomara et al. (1999); Morton et al. (1994); Walker et al. (2003) and William et al. (1994), they found a significant link between the depth of invasion and the oral cancer survival. As regards to *p63*, Muzio et al. (2005) proved that *p63* over expression associates with poor prognosis in oral cancer.

The proposed hybrid AI method, which is ReliefF-GA-ANFIS outperformed the other methods of ANN, SVM and LR in the classification accuracy and AUC. The results shown are in accordance with the objective of this research in which the classification performance is much better

TABLE 6. Classification results for ANFIS, ANN, SVM and LR

Model	ANFIS		ANN		SVM		LR	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Group 1								
3-input	67.14	0.55	50.24	0.55	64.29	0.50	54.29	0.54
4-input	60.48	0.59	52.86	0.59	64.29	0.50	51,43	0.52
5-input	67.62	0.59	56.76	0.58	64.29	0.50	61.43	0.62
6-input	51.90	0.47	47.00	0.51	64.29	0.50	47.62	0.55
7-input	64.74	0.57	50.05	0.54	54.76	0.46	48.10	0.51
Group 2								
3-input	93.81	0.90	84.62	0.83	74.76	0.70	74.76	0.70
4-input	93.81	0.90	73.38	0.75	71.43	0.68	74.76	0.70
5-input	65.71	0.63	48.00	0.52	74.76	0.70	71.43	0.68
6-input	64.76	0.62	51.57	0.53	74.43	0.66	58.10	0.55
7-input	68.10	0.67	45.86	0.47	54.76	0.53	61.43	0.60

ANFIS - Adaptive Neuro-Fuzzy Inference System ANN - Artificial neural network
SVM - Support vector machine LR - Logistic regression
AUC - Area under ROC curve

TABLE 7. Classification results for 1 to 3-year oral cancer prognosis

Oral cancer prognosis	Accuracy (%)	AUC
1-year	93.33	0.90
2-year	84.29	0.77
3-year	93.81	0.90

with the existence of genomic markers in Group 2. From the results in Table 6, the best model is ReliefF-GA with ANFIS classification. This proves that the ANFIS is the most optimum classification tool for oral cancer prognosis. The optimum subset of features for oral cancer prognosis has been identified and the features are *Drink, Invasion* and *p63*.

CONCLUSION

In this research, a ReliefF-GA-ANFIS model is proposed for the oral cancer prognosis based on the clinicopathologic and genomic markers. The proposed model consists of two stages, first, ReliefF-GA is the feature selection method and second, the ANFIS model served as the classifier. The classification accuracy obtained by the proposed model had the highest accuracy of 93.81% and AUC of 0.90 if compared to other methods which were tested on. The optimum feature subset for the oral cancer dataset has been determined and the selected features are *Drink*, *Invasion* and *p63*. The results have shown that the oral cancer prognosis is more accurate with the use of combination of both markers. However, more tests and experiments needed to be done in order to further verify the results obtained in this research. Although the sample size is small, it is hoped that this research will serve as a stepping stone for larger multicentre studies in the future.

ACKNOWLEDGEMENTS

This study is supported by the University of Malaya Research Grant (UMRG) with the project number RG026-09ICT. The authors would like to thank Dr Mannil Thomas Abraham from the Tunku Ampuan Rahimah Hospital, Ministry of Health, Malaysia, the staff of Oral & Maxillofacial Surgery Department, the Oral Pathology Diagnostic Laboratory, the OCRCC, Faculty of Dentistry and the ENT Department, Faculty of Medicine, University of Malaya for the preparation of the dataset and related documents for this project.

REFERENCES

Asakage, T., Yokose, T., Mukai, K., Tsugane, S., Tsubono, Y., Asai, M. & Ebihara, S. 1998. Tumor thickness predicts cervical metastasis in patients with stage I/II carcinoma of the tongue. *Cancer* 82: 1443-1448.

Catto, J.W.F., Abbod, M.F., Linkens, D.A. & Hamdy, F.C. 2006. Neuro-Fuzzy modeling: An accurate and interpretable method for predicting bladder cancer progression. *The Journal of Urology* 175: 474-479.

Dom, R.M., Abdul-Kareem, S., Abidin, B., Jallaludin, R.L.R., Cheong, S.C. & Zain, R.B. 2008. Oral cancer prediction model for Malaysian sample. *Austral-Asian Journal of Cancer* 7(4): 209-214.

Futschik, M.E., Sullivan, M., Reeve, A. & Kasabov, N. 2003. Prediction of clinical behaviour and treatment for cancers. *Applied Bioinformatics* 2(3 Suppl): S53-S58.

Gevaert, O., Smet, F.D., Timmerman, D., Moreau, D. & Moor, B.D. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22(14): e184-e190.

Giacomarra, V., Tirelli, G., Papanikolla, L. & Bussani, R. 1999. Predictive factors of nodal metastases in oral cavity and oropharynx carcinomas. *Laryngoscope* 109: 795-799.

Jefferies, S. & Foulkes, W.D. 2001. Genetic mechanisms in squamous cell carcinoma of the head and neck. *Oral Oncology* 37: 115-126.

Kawazu, T., Kazuyuki, A., Yoshiura, K., Nakayama, E. & Kanda, S. 2003. Application of neural networks to the prediction of lymph node metastasis in oral cancer. *Oral Radiology* 2003(19): 137-142.

Kononenko, I. 1994. Estimating Attributes: Analysis and Extension of RELIEF. Paper read at *ECML-94 Proceedings of the European conference on machine learning on Machine Learning*.

Leite, I.C.G. & Koifman, S. 1998. Survival analysis in a sample of oral cancer patients at a reference hospital in Rio de Janeiro, Brazil. *Oral Oncology* 34(5): 347-352.

Li, H., Li, D., Zhang, C. & Nie, S. 2007. An application of machine learning in the criterion updating of diagnosis cancer. *International Conference in Neural Networks and Brain* 2005(1): 187-190.

Mehrotra, R. & Yadav, S. 2006. Oral squamous cell carcinoma: Etiology, pathogenesis and prognostic value of genomic alterations. *Indian Journal of Cancer* 43(2): 60-66.

Morton, R.P., Ferguson, C.M., Lambie, N.K. & Whitlock, R.M. 1994. Tumor thickness in early tongue cancer. *Archives of Otolaryngology-Head & Neck Surgery* 120: 717-720.

Muzio, L.L., Santarelli, A., Caltabiano, R., Rubini, C., Pieramici, T. & Trevisiol, L. 2005. *p63* overexpression associates with poor prognosis in head and neck squamous cell carcinoma. *Human Pathology* 36: 187-194.

Oliveira, L.R., Ribeiro-Silve, A., Costa, J.P.O., Simoes, A.L., Di Matteo, M.A.S. & Zucoloto, S. 2008. Prognostic factors and survival analysis in a sample of oral squamous cell carcinoma patients. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology* 106(5): 685-695.

Reichart, P.A. 2001. Identification of risk groups for oral precancer and cancer and preventive measures. *Clinical Oral Investigations* 5: 207-213.

Seker, H., Odetayo, M.O., Petrovic, D. & Naguib, R.N.G. 2003. A fuzzy logic based-method for prognostic decision making in breast and prostate cancers. *IEEE Transactions on Information Technology in Biomedicine* 7(2): 114-122.

Siow-Wee, C., Kareem, S.A., Kallarakkal, T.G., Merican, A.F., Abraham, M.T. & Zain, R.B. 2011. Feature selection methods for optimizing clinicopathologic input variables in oral cancer prognosis. *Asia Pacific Journal of Cancer Prevention* 12(10): 2659-2664.

Sun, Y., Goodison, S., Li, J., Liu, L. & Farmerie, W. 2007. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 23(1): 30-37.

Thurfjell, N., Coates, P.J., Boldrup, L., Lindgren, B., Bäcklund, B., Uusitalo, T., Mahani, D., Dabelsteen, E., Dahlqvist, Å., Sjöström, B., Roos, G., Vojtesek, B., Nenutil, R. & Nylander, K. 2005. Function and importance of *p63* in normal oral mucosa and squamous cell carcinoma of the head and neck. *Current Research in Head and Neck Cancer. Molecular Pathways, Novel Therapeutic Targets, and Prognostic Factors,* edited by Bier, H. Adv. Otorhinolaryngol. Basel, Karger 62: 49-57.

Walker, D.M., Boey, G. & McDonald, L.A. 2003. The pathology of oral cancer. *Pathology* 35(5): 376-383.

Williams, J.K., Carlson, G.W., Cohen, C., Derose, P.B., Hunter, S. & Jurkiewicz, M.J. 1994. Tumor angiogenesis as a prognostic factor in oral cavity tumors. *The American Journal of Surgery* 168: 373-380.

Zain, R.B. & Ghazali, N. 2001. A review of epidemiological studies of oral cancer and precancer in Malaysia. *Annals of Dentistry University of Malaya* 8: 50-56.

Siow-Wee Chang* & Sameem Abdul Kareem
Department of Artificial Intelligence
Faculty of Computer Science and Information Technology
University of Malaya
50603 Kuala Lumpur
Malaysia

Siow-Wee Chang* & Amir Feisal Merican Aljunid Merican
Bioinformatics Division, Institute of Biological Science
Faculty of Science
University of Malaya
50603 Kuala Lumpur
Malaysia

Rosnah binti Zain
Department of Oral Pathology and
Oral Medicine and Periodontology
Faculty of Dentistry
University of Malaya
50603 Kuala Lumpur
Malaysia

Rosnah binti Zain
Oral Cancer Research and Coordinating Centre (OCRCC)
Faculty of Dentistry
University of Malaya
50603 Kuala Lumpur
Malaysia

*Corresponding author; email: siowwee@um.edu.my

Received: 11 September 2012
Accepted: 22 July 2013