ICIT 2015 The 7th International Conference on Information Technology
doi:10.15849/icit.2015.0058

Analysis of Oral Cancer Prediction using Features Selection with Machine Learning

Fatihah Mohd, Noor Maizura Mohamad Noor
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu (UMT)
21030 K.Terengganu, Terengganu, Malaysia
mpfatihah@yahoo.com, maizura@umt.edu.my

Zainab Abu Bakar
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM)
40450 Shah Alam, Selangor, Malaysia
zainab@tmsk.uitm.edu.my

Zainul Ahmad Rajion
School of Dental Sciences
Universiti Sains Malaysia (USM),
16150 Kubang Kerian, Kelantan, Malaysia
zainul@kck.edu.my

***Abstract*—Accuracy is one of the main elements in the disease diagnose. Thus, it is important to select most relevant attributes to generate the optimal accuracy. The objective of this study is to predict more accurately the presence of oral cancer primary stage with reduced number of attributes. Originally, 25 attributes have been identified in order to predict the oral cancer staging. In this study, the integrated diagnostic model with hybrid features selection methods is used to determine the attributes that contribute the most to the diagnosis of oral cancer, which, indirectly, reduces the number of features that are collected from a variety of patient records. Twenty-five attributes have been reduced to 14 features using hybrid feature selection. Subsequently, four classifiers: Updatable Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine are used to predict the diagnosis of patients with oral cancer. Also, the observations indicate that the Support Vector Machine outperforms other machine learning algorithms after incorporating feature subset selection with SMOTE at preprocessing phases.**

Keywords—diagnose; feature selection; oral cancer; SMOTE;

I. INTRODUCTION

Early clinical diagnosis is seen as an important element in reducing the mortality rate of deadly disease. The process of clinical diagnosis begins with information gathering or eliciting data from a patient's history. It includes data collection from the patient's primary report of symptoms, past medical history, family history, and social history. In this process, sometimes decision making can be done, where the clinician can start the procedure of formulating a list of possible diagnoses [1]. Then, by doing a physical examination, the physician detects abnormalities by looking at, feeling, and listening to all parts of the body. However, the patient's record is a collection of features and data that leads to problems in the diagnosis.

Another issue is most of the diseases share the same clinical features and scaling. Commonly, a biopsy is taken for the diagnosis. However, the diseases often share many histopathological features as well. Besides that, one disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages [2,3].

The difficulty to diagnose clinical diseases has attracted many experts to study the solutions from the perspectives of both medical and computer science. A variety of machine learning methods in data mining and artificial intelligence such as feature selection (FS) and classifications are usually applied in the diagnosis of diseases [4,5]. Both FS processes and classification techniques are capable of produce the most

doi:10.15849/icit.2015.0058

relevant features to build an efficient classifier. In addition, they can also eliminate noise and reduce features to achieve a classification with higher accuracy. Examples of common classification methods include Naïve Bayes (NB) [6,7,8,9], Support Vector Machine (SVM) [10,11,12,13], Genetic Algorithm (GA) [14,15,16], k-Nearest Neighbor (KNN) [17] and Multilayer Perceptron (MLP) [18]. These efficient methods are able to aid doctors in making decision of diagnosis based on the features obtained from the classification. This study aims to produce an efficient predicting diagnosis with deducted number of features that contribute more to the use of oral cancer using feature selection with classification. In this paper, an integrated diagnostic model for selection of the optimum features is proposed. The model is based on integrated a preprocessing phase and hybrid FS which is used to select of features used in the diagnosis process. We also suggested our new hybrid feature selection methods to diagnose the diseases using popular classification techniques such as NB, MLP, SVM and KNN.

Clinical data sets are usually coming with no balance. Class imbalance occurs when one of the classes that are less represented. In the training data, this incident will affect the performance of the algorithm for selecting cases. This often occurs when data collection is not enough [19]. Most classification algorithms aim to minimize the error rate and the percentage of incorrect prediction of class labels [20,21]. To overcome this problem, we propose a preprocessing of imbalanced data set before the features selection stage. In this study, we integrate the Synthetic Minority Oversampling technique (SMOTE) algorithm in our diagnostic model to resolve the problem of imbalance data set.

This paper is organized into four sections. In Section II, the materials and methods included in this study are elaborated. The simulation results of experimental works are presented in Section III. Conclusions are drawn in Section IV.

II. MATERIALS AND METHODS

This section describes oral cancer data set, oversampling method (SMOTE) and features selection algorithms used in this study. The development of the integrated diagnostic model is also presented in this section together with a hybrid feature selection for diagnosis primary stage of oral cancer.

A. Oral Cancer (OC) Data Set

The OC data set in this study consist of 25 variables or features and 82 instances or records [22]. The 25 features are divided into four: (i) demographic features, (ii) clinical signs and symptoms, (iii) histopathological features and (iv) primary stage features (see Table I). The feature of the primary stage is target as class label of disease's diagnostics.

TABLE I. ORAL CANCER DATA SET WITH 25 FEATURES

Demographical Features	**Clinical Features**
F1: Age	F7: Difficulty in Chewing / Swallowing
F2: Gender	F8: Painless Ulceration > 14 Days
F3: Ethnicity	F9: Neck Lump
F4: Smoking	F10: Loss of Appetite
F5: Quid Chewing	F11: Loss of Weight
F6: Alcohol	F12: Hoarseness of Voice
	F13: Bleeding
	F14: Burning Sensation in the Mouth
	F15: Painful
	F16: Swelling
	F17: Numbness
	F18: Site
	F19: Size
	F20: Lymph Node Involvement
Histopathological Features	
F21: Histological Type / Class	
F22: Differentiation (SCC Type)	
F23: Primary Tumor (T)	
F24: Regional Lymph Nodes (N)	
F25: Distant Metastasis (M)	
Primary Stage (Class/Target)	
One: Stage I	
Two: Stage II	
Three: Stage III	
Four: Stage IV	

B. SMOTE

Clinical data is imbalance in nature, therefore the data need to be preprocessed prior to the next stage of processes. The data set is unbalanced when at least one class have only a small number of instances (called the minority class) while other classes are a majority (with a large number of instances). The limitation of data collection often contributes to imbalance data set [19]. In this situation, classifiers of the majority class usually have good accuracy while the minority class(es) has/have very poor accuracy. In this study, Synthetic Minority Oversampling Technique (SMOTE) algorithm was applied to resolve the problem of imbalance data set during the preprocessing stage. SMOTE is running in a WEKA software environment under the supervised filter function, weka.filters.supervised.instance.SMOTE. The original oral cancer data set must fit entirely in memory. The amount of SMOTE and number of nearest neighbors is specified as Fig. 1.

doi:10.15849/icit.2015.0058

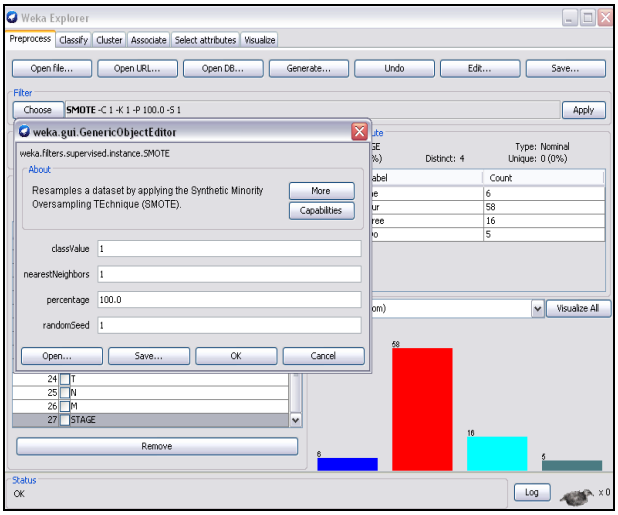

Fig. 1. SMOTE function in Weka software

C. *Features Selection*

Feature selection (FS) is the process of revealing and reducing unrelated, weakly relevant or redundant features or dimensions in a given data set. The objective of FS is to find the optimal subset. Following are the functions used for feature evaluation (FS) within this study:

- CfsSubsetEval. It evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them.
- CorrelationVariableEval. It evaluates the worth of features by measuring the correlation (Pearson's) between it and the class. Nominal features are considered as a value by value basis by treating each value as an indicator. An overall correlation for a nominal feature is arrived at via a weighted average.
- InfoGainVariableEval. It evaluates the worth of a feature by measuring the information gain with respect to the class.

All the features were searched using these algorithms:

- BestFirstForward or sequential forward features selection (SFFS). It searches the space of feature subsets by greedy hill climbing augmented with a backtracking facility.
- Ranker. Rank features by their individual evaluations. It is used in conjunction with features evaluators (ReliefF, GainRatio, Entropy, and others).
- LinearForwardSelection with floating forward selection or known as Sequential Backward Selection (SBFS). It is an extension of BestFirst. The search direction can be forward or floating forward selection (with optional backward search steps).

D. *An Integrated Diagnostic Model*

In this study, the integrated diagnostic model is proposed to diagnose OC data set. It integrates the preprocessing phases and features selection methods (see Fig. 2). The collected OC data are first introduced, as well as the case study with a number of instances and features. The data is preprocessed by scaling or standardizing them to reduce the level of dispersion between the features in the data set. After re-sampling of imbalance data set, the process proceeds to features selection in order to find the most relevant variables in the diagnosis. At this phase, FS techniques are used to select most relevant feature's model, and the various methods of that technique are employed. These models are validated by using the test validation data set. Four algorithms of machine learning are used at this stage to evaluate performance measure accuracy of FS model. Finally, the optimum result gives the best prediction technique or algorithm for that particular type of data set.

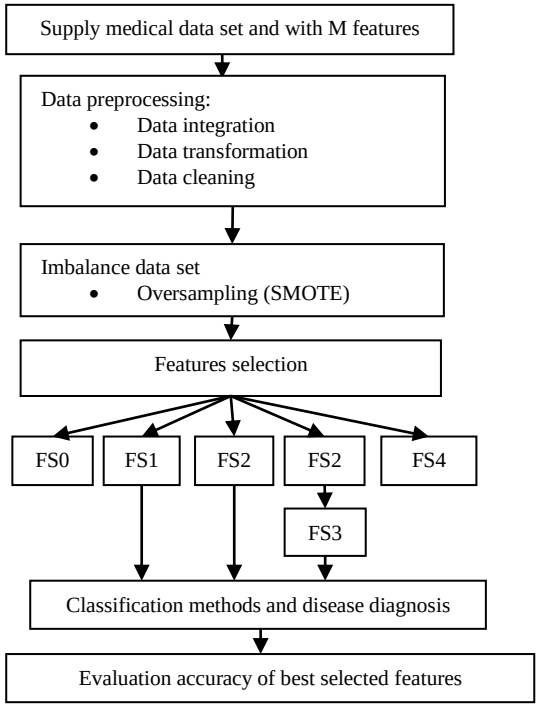

Fig. 2. An integrated diagnostic model for OC data set

III. Results and Discussion

A. *Balance Data set*

The original OC data set were categorized into four classes. There were 58 instances of the majority class (stage four), 16 for stage three and stage one and two falls under the category of minority class with the number of instances less than 10. In this study, for the training set 10-fold cross-validation is used. The minority class is over-sampled at 100%, 200%, 300%, and 400% of its original size. Table II shows the result of re-sample an imbalance OC data set using SMOTE. The result after over sampling showed the number of instances is a re-sample to 210 instead of 82 instances.

doi:10.15849/icit.2015.0058

TABLE II. BALANCED CLASS DISTRIBUTION FOR OC BY APPLYING SMOTE

Class Name	# of Instances	%	# of Instances with SMOTE	%
One	3	3.66	48	22.86
Two	5	6.09	40	19.05
Three	16	19.51	64	30.48
Four	58	70.73	58	27.62
Total	82		210	

Fig. 3 shows the class distribution of each minority class of OC data set, stage one (22.86%) and two (19.05%) are almost balance as majority class, stage three (30.48%) and four (27.62%).

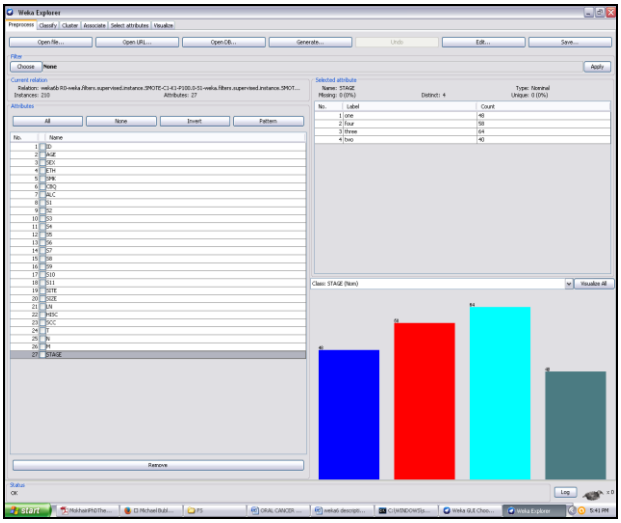

Fig. 3. Balance OC data set using SMOTE in WEKA software.

B. Optimum Features Selected

After loading the data set, the FS algorithms are applied to find the most significant features of the data set. It started with all features selected (FS0), cfsSubSetEval with Best First Forward (FS1), InfoGain Variable Evaluator combined Sequential Backward Selection or known as Linear Forward Selection with Floating Forward Selection (IGSBFS) (FS2), Correlation Variable Evaluator with Ranker (FS3), and hybrid FS3 with CfsSubset Evaluator with Linear Forward Selection (FS4). Table III shows the details of results for each FS method.

TABLE III. SELECTED ATTRIBUTES WITH FEATURES SELECTION METHODS

FS	Method	Selected attributes
FS0	No selected feature	25 attributes
FS1	cfsSubSetEval Best First Forward	2,3,8,9,15,16,17,18,19,20,21,22,23,24 (14 attributes)
FS2	CorrelationAttributeEval Ranker	20,23,21,22,16,19,24,8,2,15,7,17,3,18,5,1,13,9,11,6,25,10,14,4,12 (25 attributes) Remove 11 attributes
FS3	CfsSubsetEval LinearForwardSelection (forward)	20,23,21,22,16,19,24,8,2,15,17,3,18,9 (14 attributes)
FS4	(IGSBFS) InfoGainAttributeEval Ranker CfsSubsetEval LinearForwardSelection (floating forward selection)	23,21,19,24,20,18,22,16,8,5,1,7,3,2,17,13,5,9,11,12,14,4,6,10,25 Remove gain ratio=0 12,14,4,6,10,25 (6 attributes) Selected features = 23,21,19,24,20,18,22,16,8,5,1,7,3,2,17,13,5,9,11 (19 attributes) Optimum features = 23,21,19,24,20,18,22,16,8,15,3,2,17,9 (14 attributes)

The experiment of FS using WEKA software started with 25 features and 210 instances. It ended at FS4 with 14 optimal features namely 2, 3, 8, 9, 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24.

C. Accuracy Classification Performance

The performance measure of accuracy is considered in order to evaluate the efficiency of the FS methods. The measures are compiled by the following unit: Classification Accuracy (%) = (TP+TN) / (TP + FP + FN +TN). In this study, the evaluations are conducted in WEKA with 10 fold cross validation. Four different machine learning algorithms are used to classify the OC data set with four FS methods:

- Updateable Naive Bayes (NB). This is the updateable version of Naïve Bayes and using estimator classes.
- Multilayer Perceptron (MLP). A Classifier that uses backpropagation network to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time.
- SMO-Poly Kernel (E-1.0) (SVM). This implementation globally replaces all missing values and transforms nominal variables into binary ones. It also normalizes all features by default.
- K-Nearest neighbors classifier (lazy.IBk). K-nearest neighbors classifier can select appropriate value of K based on cross-validation. It can also do the distance weighting.

Table IV shows the result for the classifier without oversampling method, SMOTE. It started with select all features of OC data set, 25 features. Next feature selection phase, FS2 is also carrying on with 25 features. Finally, a classifier with 14 selected features from FS3 is generated. Using oversampling (SMOTE), the results for three FS methods with four classifiers show that the features selected by the integrated diagnostic model contributed to improved accuracy of the entire classification algorithm used for the OC data sets.

Table V demonstrates that FS with SMOTE outperforms FS without the implementation of SMOTE. The accuracy of OC data set for FS3 improves from 87.80% to 94.76% for NB,

doi:10.15849/icit.2015.0058

90.24% to 95.24% for MLP, 86.59% to 96.20% for SVM and 76.83% to 91.43% for KNN. Findings from Table VI are also shown that the highest classification accuracy performance using SVM algorithm, with accuracy of 96.19% with 14 optimal features selection namely 2, 3, 8, 9,15, 16, 17, 18, 19, 20, 21, 22, 23 and 24. The empirical comparison between fives FS methods for the entire classifier algorithm is as well performed as graph comparison as Fig 3. It shows the optimal features set from FS3 contribute the highest accuracy performance.

TABLE IV. PERFORMANCE ACCURACY FOR THREE SELECTED FEATURES SELECTION ON OC DATA SET WITHOUT SMOTE

Classification Accuracy Without SMOTE (%)			
Algorithm	**FS0**	**FS2**	**FS3**
NB	85.37 14.63	75.61 24.39	87.80 12.20
MLP	76.83 23.17	79.27 20.73	90.24 9.76
SVM	62.20 37.80	62.20 37.80	86.59 13.41
KNN	75.61 24.39	75.61 24.39	76.83 23.17

TABLE V. PERFORMANCE ACCURACY FOR THREE SELECTED FEATURES SELECTION ON OC DATA SET WITH SMOTE

Classification Accuracy With SMOTE (%)			
Algorithm	**FS0**	**FS2**	**FS3**
NB	91.90 8.10	91.91 8.10	94.76 5.24
MLP	94.29 5.71	93.81 6.19	95.24 4.76
SVM	93.33 6.67	93.33 6.67	96.20 3.80
KNN	86.19 13.81	86.19 13.81	91.43 8.57

TABLE VI. PERFORMANCE ACCURACY FOR FIVE FEATURES SELECTIONS ON OC DATA SET

Algorithm	**No. of Features**	**Accuracy (%)**			
		NB	**MLP**	**SVM**	**KNN**
FS0	25	91.90	94.23	93.33	86.19
FS1	14	94.76	94.76	92.38	90.95
FS2	25	91.90	93.81	93.33	86.19
FS3	14	94.76	95.24	96.19	91.43
FS4	14	94.76	94.76	92.38	90.95

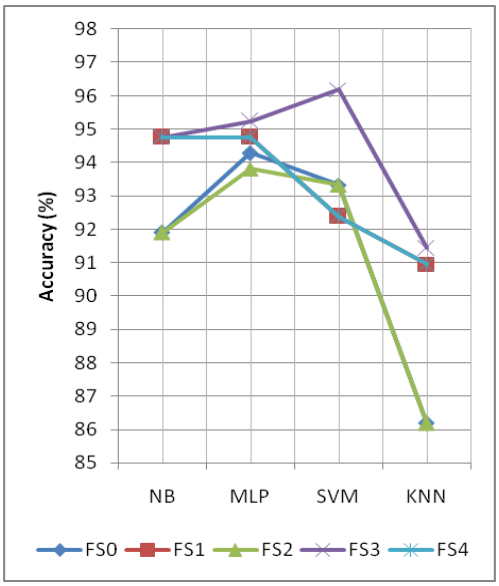

Fig. 4. Performance accuracy comparison between the fives features selection methods with NB, MLP, SVM and KNN algorithm.

IV. CONCLUSION

In the field of medical diagnosis, one of the main issues is accuracy in the diagnose of the patient disease. In order to generate the highest accuracy, it is important to reduce and select most related features. Thus, we investigate data reduction methods to be applied in the diagnosis of OC primary stage using machine learning classification methods. In this study, the integrated diagnostic model between preprocessing phases and hybrid FS method to diagnose OC primary stage demonstrated an increase in classification accuracy. It shows highest classification accuracy with 14 optimal features from a set of 25 features. The optimal feature subset was trained with four classification algorithms, Updatable Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine. Experimental results from this study present that a preprocessing technique before data selection greatly enhances the accuracy of classification. It is also noted that the classifier accuracy enhanced by applied by FS methods than the classifier accuracy without FS. These results clearly demonstrate the great potential of the proposed model for the diagnostic of clinical data.

ACKNOWLEDGMENT

This study has been supported in part of the Exploratory Research Grant Scheme (ERGS) 600_RMI/ERGS 5/3 (3/2011) under the Malaysia Ministry of Higher Education (MOHE) and Universiti Teknologi MARA (UiTM) Malaysia. The authors would like to acknowledge all contributors who have provided their assistance in the completion of the study and anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

doi:10.15849/icit.2015.0058

REFERENCES

[1] B. Neville, D. Damm, C. Allen, and J. Bouguot, "Differential diagnosis of oral and maxillofacial disease," in *Oral and Maxillofacial Pathology*, 3rd ed. China: Saunders Elsevier, 2009, Appendix, pp 917.

[2] J. Xie, J. Lei, W. Xie, X. Gao, Y. Shi, and X. Liu, "Novel hybrid feature selection algorithms for diagnosing erythemato-squamous diseases," in *Health Information Science*, J. He, et al., Eds. Berlin Heidelberg: Springer, 2012, ch. 21, pp. 173-185.

[3] B. Karlk and G. Harman. "Computer-aided software for early diagnosis of erythemato-squamous diseases," in *Electronics and Nanotechnology (ELNANO), IEEE XXXIII International Scientific Conference*, Kiev, Ukraine, 2013, pp. 276-279.

[4] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, vol. 32, pp. 71-83, October 2004.

[5] K. C. Tan, Q. Yu, C. M. Heng, and T. H. Lee, "Evolutionary computing for knowledge discovery in medical diagnosis," Artificial Intelligence in Medicine, vol. 27, pp. 129-154, February 2003.

[6] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, IBM New York, 2001, pp. 41-46.

[7] S. Mukherjee and N. Sharma, "Intrusion detection using naive bayes classifier with feature reduction," Procedia Technology, vol. 4, pp. 119-128, February 2012.

[8] F. Calle-Alonso, C. J. Pérez, J. P. Arias-Nicolás, and J. Martín, "Computer-aided diagnosis system: a bayesian hybrid classification method," Computer Methods and Programs in Biomedicine, vol. 112, pp. 104-113, October 2013.

[9] M. Wozniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," Information Fusion, vol. 16, pp. 3-17, March 2014.

[10] F. Baronti, F. Colla, V. Maggini, A. Micheli, A. Passaro, A. M. Rossi, and A. Starita, "Experimental comparison of machine learning approaches to medical domains: a case study of genotype influence on oral cancer development," in *European Conference on Emergent Aspects in Clinical Data Analysis (EACDA)*, Italy, 2005, pp. 81-86.

[11] L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, "An enhanced Support Vector Machine classification framework by using euclidean distance function for text document categorization," Applied Intelligence, vol. 37, pp. 80-99, July 2012.

[12] G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," Neuroscience and Biobehavioral Reviews, vol. 36, pp. 1140-1152, April 2012.

[13] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Systems with Applications, vol. 41, pp. 1476-1482, 2014.

[14] D. Mantzaris, G. Anastassopoulos, and A. Adamopoulos, "Genetic algorithm pruning of probabilistic neural networks in medical disease estimation," Neural Networks, vol. 24, pp. 831-835, October 2011.

[15] A. Ozcift and A. Gulten, "Genetic algorithm wrapped bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases," Digital Signal Processing: A Review Journal, vol. 23, pp. 230-237, January 2013.

[16] S. W. Chang, S. A. Kareem, A. Merican, and R. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," BMC Bioinformatics, vol. 14, pp. 170, May 2013.

[17] W. L. Tung and C. Quek, "GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data," Artificial Intelligence in Medicine, vol. 33, pp. 61-88, January 2005.

[18] A. E. Hassanien, H. M. Moftah, A. T. Azar, and M. Shoman, "MRI breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier," Applied Soft Computing Journal, vol. 14, pp. 62-71, January 2014.

[19] J. M. Malof, M. A. Mazurowski, and G. D. Tourassi, "The effect of class imbalance on case selection for case-based classifiers: an empirical study in the context of medical decision support," Neural Networks, vol. 25, pp. 141-145, january 2012.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Int. Res, vol. 16, pp. 321-357, June 2002.

[21] Q. Wang and W. Chen, "A combined SMOTE and cost-sensitive twin support vector machine for imbalanced classification," Journal of Computational Information Systems, vol. 10, pp. 5245-5253, June 2014.

[22] F. Mohd, Z. A. Bakar, N. M. M. Noor, Z. A. Rajion, and N. Saddki, "A hybrid selection methods based on HCELFS and SVM for the diagnosis of oral cancer staging," in *Advanced Computer and Communication Engineering Technology*, H. A. Sulaiman, et al., Switzerland: Springer, 2015, ch. 77, pp. 821-831.