Chapter 77
A Hybrid Selection Method Based on HCELFS and SVM for the Diagnosis of Oral Cancer Staging

Fatihah Mohd, Zainab Abu Bakar, Noor Maizura Mohamad Noor, Zainul Ahmad Rajion and Norkhafizah Saddki

Abstract A diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method is developed to diagnose the stage of oral cancer in patients. The hybrid feature selection method, named Hybrid Correlation Evaluator and Linear Forward Selection (HCELFS), combines the advantages of filters and wrappers to select the optimal feature subset from the original feature set. In HCELFS, Correlation Attribute Evaluator acts as filters to remove redundant features and Linear Forward Selection with SVM acts as the wrappers to select the ideal feature subset from the remaining features. This study conducted experiments in WEKA with ten fold cross validation. The experimental results with oral cancer data sets demonstrate that our proposed model has a better performance than well-known feature selection algorithms.

F. Mohd (✉) · N.M.M. Noor
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
21030 K. Terengganu, Terengganu, Malaysia
e-mail: mpfatihah@yahoo.com

N.M.M. Noor
e-mail: maizura@umt.edu.my

Z.A. Bakar
Faculty of Computer and Mathematical Sciences, Universiti Teknology MARA (UiTM),
Selangor 40450 Shah Alam, Malaysia
e-mail: zainab@tmsk.uitm.edu.my

Z.A. Rajion · N. Saddki
School of Dental Sciences, Universiti Sains Malaysia (USM), Kubang Kerian 16150,
Kelantan, Malaysia
e-mail: zainul@kck.edu.my

N. Saddki
e-mail: fizah@kb.usm.my

H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 315,
DOI 10.1007/978-3-319-07674-4_77

77.1 Introduction

Feature selection (FS) as preprocessing steps to machine learning in real world data, is very useful in reducing dimensionality, removing irrelevant data and noise to improve result. It could directly reduce and remove irrelevant number of the original features by selecting a subset that contributes to the optimum information for classification. The FS algorithms are divided into two categories: the filter methods and wrappers methods [1]. Both methods have their own abilities and advantages. The filter method contributes high computational efficiency compared to the wrapper method. The wrapper could achieve better results than the filter approach. In this study, we combined both methods to propose a hybrid algorithm to gain the optimum selected features.

Head and neck (HNC) cancer is one of the major cancers worldwide. One part of HNC is oral cancer with the incidence rising in every country. Early clinical cancer diagnosis is seen as an important element in reducing the mortality rate of this deadly disease. The process of clinical diagnosis begins with information gathering or eliciting data from a patient's history. It includes data collection from 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 [2]. Then, by doing a physical examination, the physician detects abnormalities by looking at, feeling, and listening to all parts of body. However, the patient's record is a collection of features and data that leads to problems for the diagnosis process. The challenge of applying computational solution to the data collected is in the conversion of data into an appropriate form, suitable for the diagnosis process [3]. Because of this, FS method is applied to reduce the irrelevant data and finally select the optimum features to diagnose the stage of oral cancer. This paper explains the development of a diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method in diagnosing the stage of oral cancer.

The remaining of the paper is organized as follows: related work is given in Sect. 77.2, while Sect. 77.3 gives a brief description about the FSA algorithms—Correlation Attribute Evaluator and LFS, SVM algorithm. Section 77.4 discusses the diagnostic model for oral cancer and Sect. 77.5 reports the results and discussion. The concluding remarks are given in Sect. 77.6 to address further research issues.

77.2 Related Work

Support Vector Machine (SVM) is an affective algorithm used in medical diagnosis for pattern recognition, machine learning and data mining. In the literature, there are some works related to medical diagnosis. Aruna et al. compared the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM RBF kernel, RBF neural networks, Decision trees J48 and Simple CART. The experiments conducted were found that SVM RBF Kernel produced highest result than other classifiers with respect to accuracy, sensitivity, specificity and precision [4]. Jaganathan et al. have proposed a feature selection method with improved F-score and SVM for breast cancer diagnosis and produced a classification accuracy of 95.565 %. This result is better than RBF Network (95.278 %) [5]. In other field, SVM is also applied in cyber-security. Maldonado and L'Huillier proposed an embedded approach for feature selection using SVM in phishing and spam classification. It outperforms other techniques in terms of classification accuracy by removing the features that affect on the generalization of the classifier by optimizing the Kernel function [6].

This related work is also focusing on the diagnosis of head and neck cancer using machine learning and data mining algorithm. For instance, Kawazu et al. [7] used neural network (NN) to predict lymph node metastasis of patients with oral cancer. They utilized histopathological data set of lymph nodes which saw an accuracy of 93.6 % in diagnosing patients. Boronti et al. produced four different results with three different methods such as SVM, Decision Trees (DTs), XCS and NN with accuracies of 75.5, 76.5, 79.2 and 71.3 % respectively [8]. In another study, they continued with other methods. They produced a classification result with DTs (70 %), XCS (79 %) and NN (78 %) [9]. Besides this, Exarchos et al. [10] employed a feature selection algorithm, Correlation-based Feature Subset selection (CFS) and the wrapper algorithm in order to omit redundant or possible irrelevant features and maintain the most informative and discriminatory ones. With the applications of Bayesian Networks, Artificial Neural Networks, SVM, DTs and Random Forests, the study produced an accuracy of (69.6 %), (66.1 %), (69.6 %), (66.1 %) and (58.9 %) respectively. However, with a hybrid model of Relief F-GA-ANFIS, Chang et al. produced a better classification accuracy with 93.81 % [11].

77.3 Materials and Methods

77.3.1 Features Selection Algorithm

In this study, feature selection for high-dimensional data are conducted in WEKA with tenfold cross validation. The main idea of feature selection functions are used to find the most significant attributes by removing features with little or no predictive information. The functions used for attribute evaluation (feature selection) within this study are as follows:

Correlation Attribute Evaluator. This algorithm evaluates the worth of an attribute by measuring the correlation between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

CFS Subset Evaluator. This algorithm evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

The Correlation Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. It measures subsets of features on the basis of the hypothesis, "*A good feature subset is one that contains features highly correlated with* (*predictive of*) *the class, yet uncorrelated with* (*not predictive of*) *each other*". The following equation gives the merit of a feature subset S consisting of k features:

$$Merit_{sk} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

where $Merit_{Sk}$ is the heuristic "merit" of a feature subset S containing k features, r_{cf} is the average value of all feature-classification correlations ($f \in S$), and r_{ff} is the average value of all feature-feature correlations. The numerator of (1) can be thought of as providing an indication of how predictive of the class a set of features are; the denominator of how much redundancy there is among the features [12].

All the attributes were searched using these algorithms:

Ranker. Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc.).

Linear Forward Selection (*forward*). Linear Forward Selection, a technique to reduce the number of attributes expansions in each forward selection step. This function is extension of Best First. It takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected. The search uses either the initial ordering to select the top k attributes, or performs a ranking (with the same evaluator the search uses later on). This algorithm starting from the empty set, sequentially add the feature x^{+} that results in the highest objective function $J(Y_k + x^{+})$ when combined with the features Y_k that have already been selected.

1. Start with the empty set $\mathrm{Y}_0 = (\emptyset)$
2. Select the next best feature $\mathrm{X} += \arg\max\,[\mathrm{J}(\mathrm{Y}_\mathrm{k} + \mathrm{x})]\mathrm{x} \notin \mathrm{Y}_\mathrm{k}$
3. Update $\mathrm{Y}_{\mathrm{k+1}} = \mathrm{Y}_\mathrm{k} + \mathrm{x}^{+}$; $\mathrm{k} = \mathrm{k} + 1$
4. Go to 2

77.3.2 Oral Cancer Dataset

The study obtained a record review of oral cancer patients from the Otorhinolaryngology Clinic at Hospital Universiti Sains Malaysia (HUSM) in Kelantan respectively. The dataset is made up of 27 parameters and a primary tumor stage as attributes for the diagnosis of the patients. The study was conducted after the

Table 77.1 Description of the datasets

Attributes no.	Attributes name
1.	Age
2.	Gender
3.	Ethnicity
4.	Smoking
5.	Chewing betel quid
6.	Alcohol
7.	S1
8.	S2
9.	S3
10.	S4
11.	S5
12.	S6
13.	S7
14.	S8
15.	S9
16.	S10
17.	S11
18.	Site
19.	Size
20.	Lymph node
21.	Histological
22.	SCC
23.	T
24.	N
25.	M
26.	Stage (class label): Stage I, Stage II, Stage III, Stage IV

obtainment of the required approvals from the Research and Ethics Committee (Human), Universiti Sains Malaysia, No.236.4.(4.4) [13]. Number of instances was 210, and 27 features with patient_id was named as label and stage was named as class label. The numerical variables were analysed through the corresponding ranges of their values. Age was divided into five groups (group 1: below 30 years old; group 2: 30–39 years old; group 3: 40–49 years old: group 4: 50–59 years old and group 5: 60 years old and above) [14, 15]. The oral cancer regions included in this study were the tongue, buccal mucosa, palate, floor of mouth, maxilla, lip, cheek, mandible, tonsil, parotid gland, oropharynx and other unspecified parts. The details of the attributes found in this dataset for features selection listed in Table 77.1.

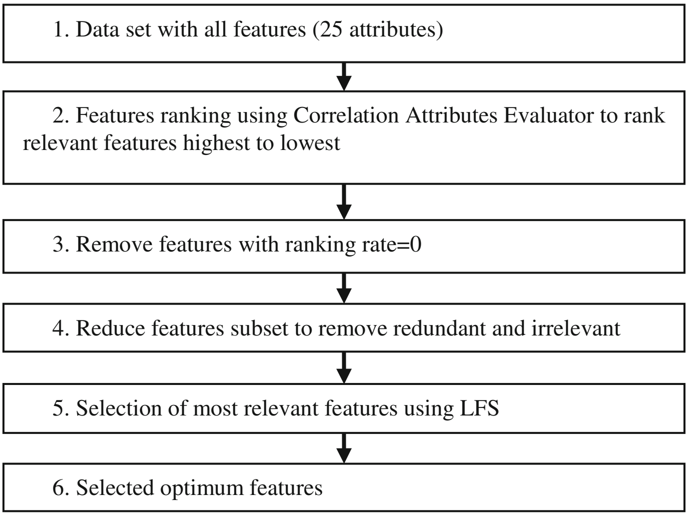
1. Data set with all features (25 attributes)
2. Features ranking using Correlation Attributes Evaluator to rank relevant features highest to lowest
3. Remove features with ranking rate=0
4. Reduce features subset to remove redundant and irrelevant
5. Selection of most relevant features using LFS
6. Selected optimum features

Fig. 77.1 Stage in hybrid correlation evaluator and linear forward selection (HCELFS) algorithm

77.4 Diagnostic Model: Hybrid Correlation Evaluator and Linear Forward Selection

A hybrid feature selection method, named HCELFS is proposed in this study. This FS method, hybrid Correlation Attribute Evaluator with ranker and CFS Subset Evaluator with Linear Forward Selection was applied for oral cancer diagnosis. It combines the advantages of both methods to select the optimal features subset from the original feature set. In the diagnostic model, the first step included the Correlation Attribute Evaluator method filtering the relevant features which resulted in reduced features subset. From this subset CFS Subset Evaluator with Linear Forward Selection (LFS) searched for the most relevant features resulting in optimum feature set used for diagnosing the cancer stage. Figure 77.1 shows the stage in the algorithm.

77.5 Results and Discussion

The experiments of features selection against oral cancer data set are conducted in WEKA with tenfold cross validation. Algorithm started with 25 features and 210 instances. With Correlation Ranking Filter, the algorithm ranked 25 features namely 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 7, 17, 3, 18, 5, 1, 13, 9, 11, 6, 25, 10, 14, 4 and 12 (see Table 77.2). We removed 1 feature namely 12 with ranking rate 0. With the resultant 24 features, the subset method then remove the redundant and

Table 77.2 Correlation ranking filter for oral cancer data set with 25 attributes

Correlation ranking filter	Ranked attributes (%)
20: Lymph node	0.4602
23: T	0.4319
21: Histological	0.3715
22: SCC	0.3611
16: S10	0.3561
19: Size	0.3558
24: N	0.3349
8: S2	0.3203
2: Gender	0.2751
15: S9	0.2660
7: S1	0.2644
17: S11	0.2345
3: Ethnicity	0.2323
18: Site	0.2297
5: Betel quid	0.1966
1: Age	0.1884
13: S7	0.1810
9: S3	0.1496
11: S5	0.1465
6: Alcohol	0.1042
25: M	0.1042
10: S4	0.0905
14: S8	0.0599
4: Smoking	0.0455
12: S6	0
Selected 25 attributes:	
20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 7, 17, 3, 18, 5, 1, 13, 9, 11, 6, 25, 10, 14, 4 and 12	

irrelevant 10 features namely, 7, 5, 1, 13, 11, 6, 25, 10, 14 and 4. This algorithm ended with 14 features namely 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 17, 3, 18 and 9 as optimum features set.

Table 77.3 summarize hybrid the features selection methods experimented in this study. It started with no features selection (FS0), Correlation Attribute Evaluator with Ranker (FS1), and then combined FS1 with CfsSubset Evaluator with Linear Forward Selection (FS2).

Table 77.3 Selected attributes with hybrid feature selection methods

FS	Method	Selected attributes
FS0	No features selection	All attributes
FS1	CorrelationAttributeEval	Ranked attributes: 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)
	Ranker	Remove ranting value = 0 (attribute 12)
FS2	CfsSubsetEval	Remove irrelevant attributes = 10 attributes 7, 5, 1, 13, 11, 6, 25, 10, 14, 4
	LinearForwardSelection (forward)	Optimum features: 14 attributes 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 17, 3, 18, 9

Table 77.4 Accuracy performance for classification algorithms on oral cancer data set

Algorithm	FS0	FS1	FS2
Updateable Naïve Bayes	91.9048	91.9048	94.7619
	8.0952	8.0952	5.2381
MLP	94.2857	93.8095	95.2381
	5.7143	6.1905	4.7619
Lazy-IBK	86.1905	86.1905	91.4286
	13.8095	13.8095	8.5714
SMO- poly kernel (E-1.0)	93.3333	93.3333	96.1905
	6.6667	6.6667	3.8095

In order to evaluate the efficiency of the FS methods, performance measure of accuracy were considered. The measures are compiled by the Classification Accuracy (%) = (TP + TN)/(TP + FP + FN + TN). In this study, four different machine learning algorithms were used to classify the oral cancer data set with three features selection methods and optimum features selected by the proposed hybrid algorithm, Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest neighbors (KNN), and SVM. NB classifier using estimator classes. MLP classifier uses backpropagation 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. K-Nearest neighbor classifier (lazy.IBk) can select appropriate value of K based on cross-validation. It can also do distance weighting. SVM or SMO-Poly Kernel (E-1.0) implemented globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Table 77.4 shows the results for the classifier. The empirical comparison shows that the features selected by the hybrid algorithm also improved the accuracy of the entire classifier algorithm used for the oral cancer data set. Table 77.5 shows the classification accuracies of our method and other classifiers from literature for the head and neck data set.

Table 77.5 Classification accuracies of this study method and other classifiers from literature

Author, year	Method	Accuracy (%)
Kawazu et al. 2003 [7]	Neural networks (NN)	93.6
Baronti et al. 2005 [8]	Support vector machines (SVM)	75.5
	Decision trees (C4.5)	76.5
	XCS	79.2
	NN	71.3
Tung and Quek 2005 [16]	FS based on wrapper	Above 90 %
	Monte Carlo evaluative selection (MCES)	
	Classification using	
	SVM with polynomial kernel, K-Nearest Neighbor (K-NN) classifier, artificial neural network (ANN) and the GenSoFNN-TVR(S) network	
Baronti and Starita, 2007 [9]	Naive Bayes (NB)	69.4
	C4.5	70
	NN	78
	XCS	79
	Hypothesis classifier systems (HCS)	83.8
Xie et al. 2010 [17]	Improved F-score and sequential forward floating search (IFSFFS)	100 (best)
	SVM	97.58 (avg)
Exarchos et al. 2011 [10]	Bayesian networks (BNs)	69.6
	ANN	66.1
	SVM	69.6
	Decision trees (DTs)	66.1
	Random forests (RFs)	58.9
Chang et al. 2013 [11]	1. Pearson's correlation coefficient (CC) and Relief-F as the filter approach	93.81 %
	2. Genetic algorithm (GA) as the wrapper approach	
	3. CC-GA and ReliefF-GA as the hybrid approach.	
	Hybrid model of ReliefF-GA-ANFIS	
Calle-Alonso et al. 2013 [18]	Combines pairwise comparison, Bayesian regression and K-NN	97.74
This study	FS on erythemato squamous disease data set, combine	98.64
	1. CorrelationAttributeEval and Ranker	
	2. CfsSubsetEval and LinearForwardSelection (forward)	
	3. SMO- Poly Kernel (E-1.0)	
This study	FS on oral cancer data set, combine	96.19
	1. CorrelationAttributeEval and Ranker	
	2. CfsSubsetEval and LinearForwardSelection (forward)	
	3. SMO- Poly Kernel (E-1.0)	

77.6 Conclusion

In this study, it is noted that a diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method to diagnose the stage of oral cancer showed an increased of classification accuracy. The hybrid feature selection method, named HCELFS, combines Correlation Attribute Evaluator which acts as a filter and SBFS which acts as the wrapper to select the ideal feature subset from the remaining features.

The experimental results with oral cancer data sets demonstrate that the new hybrid feature selection method has a better performance than well-known feature selection algorithms. It obtained optimal classification accuracy with 14 features from a set of 25 features. The optimal feature subset obtained were then trained with various data mining algorithms such as Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest neighbors (KNN), and SVM to diagnose the stage of oral cancer. One direction for future studies is to consider proposing a hybrid algorithm with various dataset and other data mining classifier.

Acknowledgments This study has been supported in part 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, technical members at Hospital Universiti Sains Malaysia (HUSM) who have helped and greatly assisted in the completion of the study. Dr. Zailani Abdullah, data mining expertise from SIAM, UMT and anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

References

1. Alpaydin, E.: Introduction to machine learning. MIT Press, Cambridge (2004)
2. Neville, B., Damm, D., Allen, C., Bouguot, J.: Oral and maxillofacial pathology. Saunders Elsevier, St. Louis (2009)
3. Poolsawad, N., Kambhampati, C., Cleland, J.G.F.: Feature selection approaches with missing values handling for data mining—a case study of heart failure dataset. World Acad. Sci. Eng. Technol. **60**, 828–837 (2011)
4. Aruna, S., Rajagopalan, S.P., Nandakishore, L.V.: Knowledge based analysis of various statistical tools in detecting breast cancer. In: First International Conference on Computer Science, Engineering and Applications (CCSEA), pp. 37–45. Chennai, India (2011)
5. Jaganathan, P., Rajkumar, N., Kuppuchamy, R.: A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification. Int. J. Mach. Learn. Comput. **2**, 741–745 (2012)
6. Maldonado, S., L'Huillier, G.: SVM-based feature selection and classification for email filtering. In: Latorre Carmona, P., Sánchez, J.S., Fred A.L.N. (eds.) Pattern Recognition—Applications and Methods, pp. 135–148. Springer, Heidelberg (2013)
7. Kawazu, T., Araki, K., Yoshiura, K., Nakayama, E., Kanda, S.: Application of neural networks to the prediction of lymph node metastasis in oral cancer. Oral Radiol. **19**, 35–40 (2003)
8. Baronti, F., Colla, F., Maggini, V., Micheli, A., Passaro, A., Rossi, A.M.: 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), pp. 81–86. Pisa, Italy (2005)

9. Baronti, F., Starita, A.: Hypothesis testing with classifier systems for rule-based risk prediction evolutionary computation. Mach. Learn. Data Min. Bioinf. **4447**, 24–34 (2007)
10. Exarchos, K., Goletsis, Y., Fotiadis, D.: Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans. Inf. Technol. Biomed. **16**(6), 1127–1134 (2011)
11. Chang, S.W., Abdul Kareem, S., Merican, A., Zain, R.: Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinf. **14**, 170 (2013)
12. Hall, M.A.: Correlation-Based Feature Selection for Machine Learning. The University of Waikato, Hamilton, New Zealand (1999)
13. Bakar, Z.A., Mohd, F., Noor, N.M.M., Rajion, Z.A.: Demographic profile of oral cancer patients in east coast of peninsular Malaysia. Int. Med. J. **20**, 362–364 (2013)
14. Mohd, F., Bakar, Z.A., Noor, N.M.M., Rajion, Z.A.: Data preparation for pre-processing on oral cancer dataset. In: 13th International Conference on Control, Automation and Systems (ICCAS), p. 324. Gwangju, Korea (2013)
15. Razak, A.A., Saddki, N., Naing, N.N., Abdullah, N.: Oral cancer presentation among Malay patients in hospital Universiti Sains Malaysia, Kelantan. Asian Pac. J. Cancer Prev **10**, 1131–1136 (2009)
16. Tung, W.L., Quek, C.: GenSo-FDSS: a neural-fuzzy decision support system for pediatric all cancer subtype identification using gene expression data. Artif. Intell. Med. **33**, 61–88 (2005)
17. Xie, J., Xie, W., Wang, C., Gao, X.: A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythemato-squamous diseases. J. Mach. Learn. Res. Workshop Conf. Proc. **11**, 142–151 (2010)
18. Calle-Alonso, F., Pérez, C.J., Arias-Nicolás, J.P., Martín, J.: Computer-aided diagnosis system: a Bayesian hybrid classification method. Comput. Methods Programs Biomed. **112**, 104–113 (2013)