International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017) pp. 4893-4898

Oral Cancer Prediction Using Gene Expression Profiling and Machine Learning

Wafaa K. Shams[1] **and Zaw Z. Htike**[2]

[1]*Faculty of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*
[2]*Faculty of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*

Abstract

Oral premalignant lesion (OPL) patients have a high risk of developing oral cancer. In this study we investigate using machine learning techniques with gene expression profiling to predict the possibility of oral cancer development in OPL patients. Four classification techniques were used: support vector machine (SVM), Regularized Least Squares (RLS), multi-layer perceptron (MLP) with back propagation and deep neural network (DNN). Fisher discriminate analysis was used to select relevant features from the gene expression array. The results show high accuracy (96%) using DNN and 94% accuracy using SVM and MLP with one sample cross validation. Furthermore, we achieved the same results using 10-fold cross validation.

INTRODUCTION

Contributing studies have shown that patients with oral premalignant lesions (OPLs) are at risk to develop oral cancer, including oral squamous cell carcinoma (OSCC). These reports were based on medical variables [1-5] Unfortunately, medical markers for prediction of the development of oral OSCC ares still poor. Prediction of oral cancer is an important part of early treatment and improved long-term prognosis. The last study done by Saintigny at al [6], indicated that gene expression arrays on OPL patients can improve the predication of oral cancer, combined with the medical and histology variables. This study was based on traditional statistical methods. The results reported a misclassification rate of 16% using 29 significant features from the whole dataset. Further, machine learning methods also were used to predict oral cancer occurance from genomic and clinic pathologic data with an accuracy around 93%. [7] Other groups used machine learning techniques to study oral cancer stages based on medical variables [8-9]. Machine learning methods show promising results for the diagnosis and prognosis of cancer research [10-11], however research was limited due to the difficulties of acquiring data collected before cancer diagnosis. In this field, researchers have tried to test different types of medical data combined with various mathematical and statics methods to create a highly accurate model of cancer prognosis. As mentioned in Saintigny et al [6], gene expression profiling of patients with (OPL) has significant results compared to other medical markers. The objective of this study is to investigate the same data set using machine learning methods to predicate cancer development. The model is based on using four classifiers: Deep neural network (DNN) [12], support vector machine (SVM)[13], Regularized Least Squares [14], and multi-layer perceptron (MLP) with back propagation [15]. The gene array has a huge number of columns compare to small sample sizes previously investigated using fisher discriminate analysis (FD)[16] to extract relevant features with high discrimination between the two classes. Different feature selection methods have been used for reduction of medical dataset as well as for extraction of the relevant features from gene microarrays [17-18], even though the performance of these methods is based on the type of data set. In this study, the two classes are the OPL patients who developed cancer and the cancer free OPL patients.

MATERIALS AND METHODS

Figure 1 shows the classification model for predication cancer development for (OPL) patients.

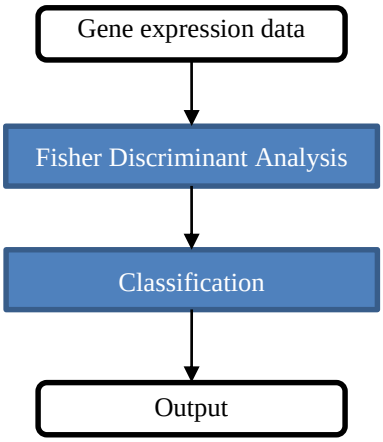

Figure 1. Classification model.

Dataset description

Sub-gene expression profiling of 86 patients with OPL is used in the present study. Of the patient set, 51 had developed oral cancer and 31 were cancer-free. The raw microarray data was generated at the University of Texas and it was deposits online at www.ncbi.nlm.nih.gov/geo. Details on the samples and description of the data can be found in Saintigny et al [6].

Fisher discriminative analysis

Feature selection is one of the main processing events used in gene recognition models. The gene expression array of OPL patients contain of a huge number of variables (features) and small numbers of raw samples (patients). Processing the data in this form is not a trivial task due to features that impart no important information for discrimination between the two classes. Therefore application of one of the featured method is important to extract the relevant results. In this study we applied the Fisher method Fisher discriminant analysis. This method is based on the statistical characteristic of the features to recognize differences between classes. The Fisher measure is defined as [16]

$$P_{12}(f) = \frac{|c_1(f) - c_2(f)|}{\sigma_1(f) + \sigma_2(f)} \quad (1)$$

where: *c1* and *c2* are the mean values for classes 1 and 2, respectively. The *σ1* and *σ2* are the standard deviations of classes 1 and 2. A $P_{12}(f)$ represents the discriminative ability of this feature.

Classification Procoess

Classification processes were done by learning with a sequence of training examples consisting of pairs (x_{ij} , y_i), where x_{ij} represents the raw variables within sample i and y_i is the associated label. The goal of the learning program is to build a classifier model, *f*, that accurately predicts the label of any unseen samples. The performance of a learning algorithm is measured in terms of the accuracy of the classifier.

Classification processes were done by learning with a sequence of training examples consisting of pairs (x_{ij} , y_i), where x_{ij} represents the raw variables within sample i and y_i is the associated label. The goal of the learning program is to build a classifier model, *f*, that accurately predicts the label of any unseen samples. The performance of a learning algorithm is measured in terms of the accuracy of the classifier.

Multi -layer perceptron (MLP)

In this study, multi-layer perceptron (MLP) with back propagation was employed [15]. The neural network contains two hidden layers with tangent-sigmoid function, and two output layers with a linear function. The tangent-sigmoid (tans*ig*) function that given by:

$$f(X_j) = \frac{1-e^{-2X_j}}{1+e^{-2X_j}} \quad (2)$$

where X_j is the input to the node of the hidden layer. Functional error is used to compute the mean square error (*E*) between expected value (target) of the training data set and the network output *Y*. The errors are then back propagated through the network, performing the descent algorithm with weights adjusted accordingly as:

$$W = \eta \frac{\delta E}{\delta W_{ij}} \quad (3)$$

where η is the learning constant.

The initial values of the weights were set randomly and the learning rate was set to 0.01. A 10-fold cross validation was used to evaluate the performance of the network and to set the optimal values for the number of nodes and the number of iterations.

Support vector machine (SVM)

For the support vector machine (SVM), the classification process is performed with separate labeled data with hyperplane and results in the maximum difference between them. This hyperplane has the maximum margin allowed, which determines the normal vector, /w/, which is given by

$$w = \sum_{i \in SV} \alpha_i y_i x_i \quad (4)$$

where *αi*'s are Langrange coefficients that maximize

$$\sum_i \alpha_i - \sum_{ij} \alpha_i \alpha_j y_i y_j \langle x_j . x_i \rangle \quad (5)$$

where $\langle x_j . x_i \rangle$ represents the kernel. In this study, we use SVM Matlab tool box with Gaussian kernel [13].

Regularized Least Squares (RLS)

Regularized Least Squares (RLS), also known as the Tikhonov regularization problem [14], minimize the given equation with a square loss function:

$$\sum (y_i - f(x_i))^2 + \gamma \|f\|_k \quad (6)$$

where *n* is the number of samples , X_i is the data samples of I of the training set, *Y* is the binary outcome. $\|f\|$ is the norm of *f* (the expected values) in Hilbert space defined by kernel K, λ is a regularization parameter that is computed from the kernel of the training data. The RLS is applied from MIT toolbox [19].

The optimal value of the regularized parameter (λ) is determined by leaving one sample out of cross validation for the training set, hence many λ values are used and the optimal one, which minimizes the validation error, was chosen.

Deep neural network (DNN)

The deep learning method utilizes unsupervised feature learning as a pre-training process, and the unsupervised feature is subsequently used with a set of labeled data to predict the classes [12]. The DNN network was designed to feed forward with back propagation, with a sigmoid function for hidden layers. In this study we used the Matlab deep learning toolbox. A 10-fold cross validation was used to adjust the number of nodes in each layer.

ClassificationScenario

To assess the performance of each of the classifiers after applying the learning mechanism, the data was divided into a training set to learn the classifier and a testing set (unseen data), which was not labelled. High performance was achieved when the classifier accurately maps out the unseen data. Two scenarios were used, the first one being to leave one sample and the second one being a 10-fold cross validation [20].

RESULTS AND DISCUSSION

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [2], [5]. The discussion can be made in several sub-chapters.

Figure 2 shows the average accuracy of classification methods applied to the gene expression array without performing Fisher discriminate analysis. The results illustrate the outcome of omitting one sample for cross validation to predicate patient's state. The accuracies are above 80% for all the methods, however DNN and SVM have the best accuracies, 84.3% and 83.5% respectively.

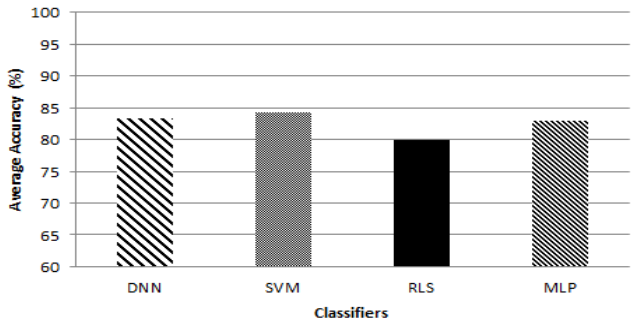

Figure 2. The average accuracy of classification methods.

The results in Figure 2, indicate that machine learning methods are capable of detecting genetic differences between the two classes. Figure 3, shows that the best classification accuracy occurs after applying Fisher discriminate analysis. The accuracy increased to above 90% for all classification methods. Cleary, DNN has the highest accuracy that is 96%, while SVM and MLP have a similar value, 94%. It is clear that classifier performance improves with the application of features selection methods.

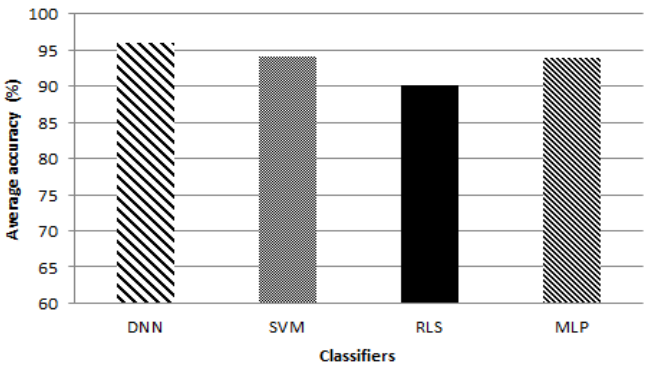

Figure 3. The average accuracy of classification methods with FD.

Table 1 shows the average accuracy, sensitivity, and specificity of the classification methods performance without FD compared to their performance with the FD approach. Clearly, detection in the sample set of patients with OSCC development has a high accuracy. This may be due to the difference in sample size between the two classes. Furthermore, the FD-DNN method showed the best performance for both classes.

Table 1. The accuracy, sensitivity and specificity of the classification methods.

Methods	Accuracy (%)	Sensitivity (%) Cancer development	Specificity (%) Cancer free
DNN	83.7	84.3	82.8
SVM	84.8	86.2	82.8
RLS	80.2	82.3	77.1
MLP	83.7	86.2	80
FD-DNN	96.5	98.1	94.2
FD-SVM	94.2	98.28	88.5
FD-RLS	90.07	96.02	85.7
FD-MLP	94.1	98.01	88.5

We examine the effects of feature selection number on the classification methods. The results are demonstrated in Figure 4. The accuracy of the classification process increased with increased number of features. Numbers were increased from 3 input features to 30 input features for most classifiers, then at 40 input features the accuracy of DNN, RLS, and MLP decreased. However it increases to reach 94% for SVM. The DNN has the highest accuracy among all classifiers.

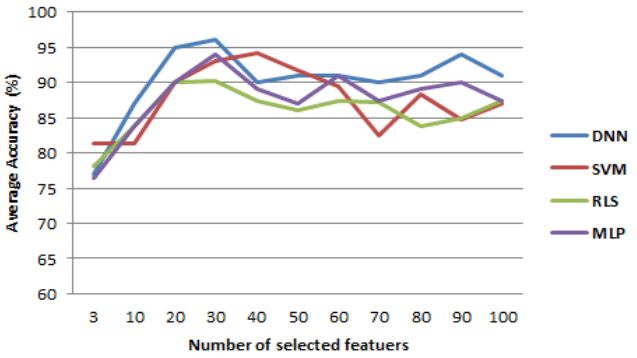

Figure 4. The average accuracy of classification methods at different n- input features.

Similarly, the performance of classification methods using 10-fold cross validation improves with input features in range of 10-30%. The high accuracy of 97% was achieved using DNN with 30 input features, as demonstrate in Figure 5. Overall, SVM and DNN showed the best performance.

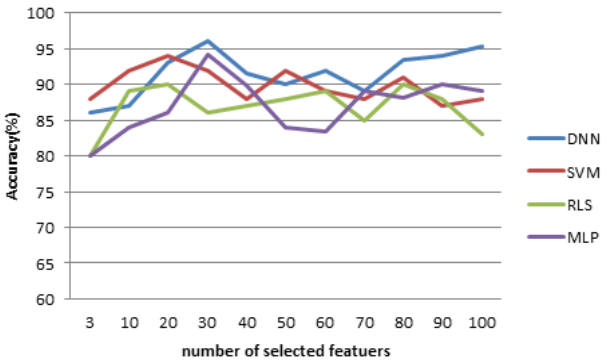

Figure 5. The average accuracy of classification methods at different n- input features using 10-fold cross validation.

Interestingly, the number of selected features has various effects on the performance of the classifiers. We choose features based on the highest value of FD. The selected range was from the maximum value of FD, 0.74, to 0.5 for 100 features. According to [6], 29 features had significant discrimination between the classes. Table 2 shows the selected gene array features with their FD values and the significant features that were reported by [6]. We investigated the performance of these features with the classification methods. FD features were referred to as Group 1 and features from [6] as Group 2 Figures 6 and 7 show the average classification accuracy of the different classification methods omitting one sample and with 10 fold cross validation, respectively. Obviously, the DNN method shows significant performance enhancement with Group 1 and Group 2, 94 % and 92%, respectively with omission of one sample. The same results have been achieved using 10 fold cross validation. However, RLS performs better with Group 2 than Group 1. Overall Group 1 has the best classification accuracy with the DNN and SVM methods

Table 2. The Probeset ID number of 29 selected gene expression profiling using FD and 29 selected gene that reported in [6].

Probeset ID	FD value	Probeset ID from [6]
8055474	0.74	8095441
7908777	0.74	8023314
8000480	0.73	7986442
7990827	0.71	8062842
8052667	0.69	8084002
8106401	0.67	7915846
8125415	0.67	8165709
8131385	0.669	8122200
8048173	0.66	8046408
8092638	0.6496	8153223
8096265	0.649	8172119
8099805	0.647	8061092
8169432	0.645	7927106
8162848	0.64	7948894
8175418	0.636	8083939
8035863	0.634	7939865
7920422	0.6247	7916777
7971359	0.6240	7964360
8086465	0.6192	8101762
8056041	0.619	7962489
8151215	0.614	7901361
8175391	0.613	8044682
8075633	0.598	8028950
8136215	0.597	7977480
7978905	0.586	8067983
8018966	0.584	8097743
8085283	0.583	8093957
8093957	0.579	8086536
8023261	0.576	8121943

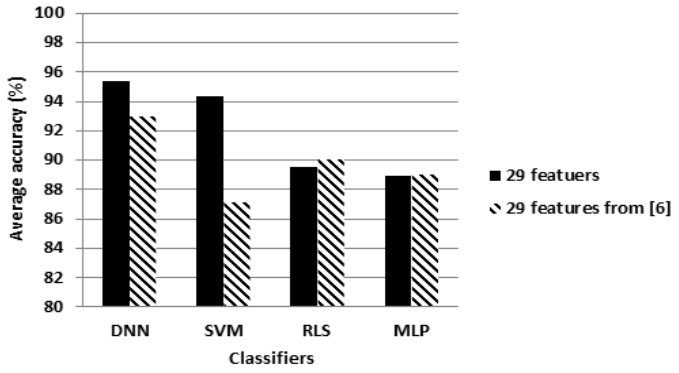

Figure 6. The average accurcay using 29 features from6 compared to FD featuers with leave one sample out.

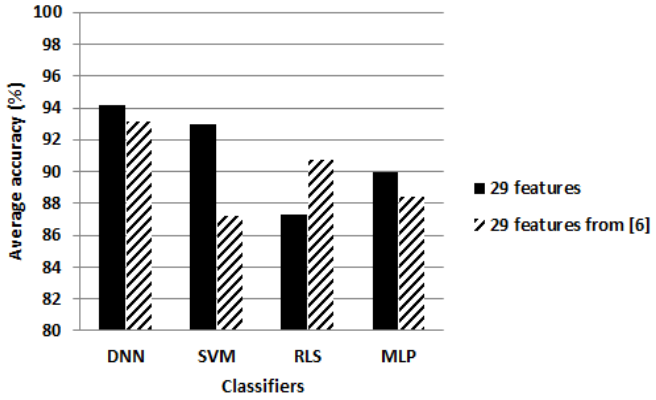

Figure 7. The average accurcay using 29 features from6 compared to FD featuers with 10 fold cross validation.

CONCLUSION

In this study, machine learning methods were investigated to predict the development of oral cancer in OPL patients. The results show the significantly accurate performance of FD with DNN compared to other classification methods, as well as compare performances with the significant features selected in [6]. Fisher discriminate analysis depends on the statistical characteristics of the data to discriminate among them. Implementation is not complex, making it useful in medical applications. Furthermore, the deep learning method showed accurate performance in discriminating and predicting the state of the medical data, which it is not an easy task. Most medical data are not linear and their classes are therefor closed.

ACKNOWLEDGEMENTS

This work was supported by the International Islamic University Malaysia under the Research Initiatives Grant Scheme (RIGS16-350-0514).

REFERENCES

[1] Lee JJ, Hong WK, Hittelman WN, Mao L, Lotan R, Shin DM, et al.Predicting cancer development in oral leukoplakia: ten years of translational research. Clin Cancer Res ,6:1702–10,2000.

[2] .Kim J, Raz D, Jablons D. Unmet need in lung cancer: can vaccines bridge the gap?Clin Lung Cancer ;9 Suppl 1:S6–12,2008

[3] Silverman S Jr, Gorsky M, Lozada F. Oral leukoplakia and malignant transformation.Afollow-up study of257patients. Cancer, 53:563–8,1984

[4] Mao L, Lee JS, Fan YH, Ro JY, Batsakis JG, Lippman S, et al. Frequent microsatellite alterations at chromosomes 9p21 and 3p14 in oralpremalignant lesions and their value in cancer risk assessment. Nat Med .2:682–5,1996

[5] Saintigny P, El-Naggar AK, Papadimitrakopoulou V, Ren H, Fan YH, Feng L, et al. DeltaNp63 overexpression, alone and in combination with other biomarkers, predicts the development of oral cancer in patients with leukoplakia. Clin Cancer Res,15:6284–91.2009

[6] Saintigny, P., Zhang, L., Fan, Y.-H., El-Naggar, A. K., Papadimitrakopoulou, V. A., Feng, L., . . . Mao, L.Gene expression profiling predicts the development of oral cancer. Cancer Prevention Research, 4(2), 218-229.2011

[7] Chang, S.-W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, *14*(1), 170.

[8] Tseng, W.-T., Chiang, W.-F., Liu, S.-Y., Roan, J., & Lin, C.-N. (2015). The Application of Data Mining Techniques to Oral Cancer Prognosis. *Journal of medical systems*, *39*(5), 1-7

[9] Mohd, F., Bakar, Z. A., Noor, N. M. M., Rajion, Z. A., & Saddki, N. (2015). A Hybrid Selection Method Based on HCELFS and SVM for the Diagnosis of Oral Cancer Staging *Advanced Computer and Communication Engineering Technology* (pp. 821-831): Springer

[10] Cruz JA, Wishart DS: Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2:59–78.2006

[11] Cai, Z., Xu, D., Zhang, Q., Zhang, J., Ngai, S.-M., & Shao, J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*. 2015

[12] Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*, 153.2007

[13] Vapnik, V. N. Statistical learning theory. Adaptive

and learning systems for signal processing, communications, and control. *Simon Haykin*.1998

[14] Evgeniou, T., Pontil, M., & Poggio, T. Regularization networks and support vector machines. *Advances in Computational Mathematics*, *13*(1), 1-50.2000

[15] Jemberie, A. A. *Information theory and artificial intelligence to manage uncertainty in hydrodynamic and hydrological models*: Taylor & Francis.2004

[16] Wiliński, A., & Osowski, S. Ensemble of data mining methods for gene ranking. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, *60*(3), 461-470.2012

[17] Latkowski, T., & Osowski, S. Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*, *42*(2), 864-872.2015

[18] Singh, R. K., & Sivabalakrishnan, M. Feature Selection of Gene Expression Data for Cancer Classification: A Review. *Procedia Computer Science*, *50*, 52-57.2015

[19] Tacchetti, A., Mallapragada, P. S., Santoro, M., & Rosasco, L. GURLS: a Toolbox for Regularized Least Squares Learning.2012

[20] Maimon, O. Z., & Rokach, L.*Data mining and knowledge discovery handbook*: Springer. 2005

[21] 21 Pirooznia, M., Yang, J., Yang, M. Q., & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, *9*(Suppl 1), S13,2008