A Model Fitting Approach for Prediction of Oral Cancer Second Primary Tumor

Andjelković Ćirković, Bojana; Costea, Daniela Elena; and Filipović, Nenad

Abstract: ***The oral squamous cell carcinoma (OSCC) is a common malignant head tumor exhibiting quite aggressive nature, often leading to unfavorable prognosis. The significant problem for this disease is the high incidence rate of second primary tumors which has considerable impact on the survival period. In this paper, we propose an intelligent model fitting approach for the development of reliable data mining model that will be able to predict occurrence of the second primary tumors for OSCC patients with high accuracy. Clinical and immunohistochemical data for 95 patients are analyzed. Distribution of patients with the second primary tumors required the resolving of problem of imbalanced classification. In order to overcome this difficulty, we propose the approach that hybridizes the advantages of intelligent genetic algorithm (GA) and artificial neural network (ANN). Data level transformations and classification algorithm design are performed simultaneously. The automatic feature selection and ANN parameter tuning are optimized by GA in order to advance the performance of classification. The results showed that 14 features have the best prognostic value for the second primary tumors occurrence. The optimized ANN classifier has better sensitivity and specificity in comparison to alternative approaches. The optimal configuration of classifier parameters as well as selected the most important input attributes empowers this model with the clinical utility.***

Index Terms: ***Artificial neural network, Feature selection, Genetic algorithm, Second primary tumors***

1. INTRODUCTION

THE oral squamous cell carcinoma is the predominant neoplasm of head. Annually, more than 0.3 million new patients are diagnosed with oral cancer worldwide [1]. Despite the advances made in treatment modalities and initial treatment of patients, the prognosis of OSCC is still poor and recurrence rates remain quite high due to aggressive local invasion and metastasis of OSCC [2]. Additionally, patients with OSCC are at increased risk of the development of second primary malignancy, which is defined as second malignancy that occurs either simultaneously or after the diagnosis of an index tumor. From the aspect of "field cancerization" hypothesis, the large areas of head and neck mucosa are affected by carcinogen exposure, resulting in a wide field of premalignant disease that gives rise to multiple independent primary tumors. At present, there is also the evidence that second primary tumors can share some or even all genetic markers with the index tumor, indicating that both tumors have arisen from a common clonal progenitor cell [3]. Due to the aggressive nature of OSCC and high rates of local and regional relapses, early identification of potential second primary tumors can be proven very beneficial for the prognosis of the patient [4] and the subsequent adjustment of the follow-up treatment. Among the aims of this study, the one was to identify a limited subset of clinical and molecular factors that are highly correlated with the occurrence of the second primary tumors at patients with OSCC using data mining approach.

The classifiers based on artificial neural network algorithm (ANN) represent promising tool in medical research due to their ability to identify complex and nonlinear patterns between input data and output variables. These classifiers can be modeled for diagnostic [5], [6] as well as prognostic [7]-[10] problems in various clinical domains. However, it is still a challenge to efficiently construct an ANN classifier which can provide accurate prediction of the unseen new samples. This so-called generalization ability depends on two tasks, feature selection and models parameter optimization [10], [12]. The selection of feature subset influences the appropriate classifiers' parameters and vice versa [13]. Therefore, obtaining the optimal feature subset and ANN parameters must occur simultaneously. Additionally, classifiers may suffer from unbalanced datasets, i.e. when at least one class is represented by a small number

Manuscript received June 19, 2017. This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant III41007 and OI174028 and international COST Action 16122 and SCOPES project JRP/IP IZ73Z0_152454.

B. A. Cirkovic is with the Faculty of Engineering, University of Kragujevac, Serbia (e-mail: abojana@kg.ac.rs) – corresponding author.

D. E. Costea is with the Department of Clinical Medicine, University of Bergen, Norway (e-mail: Daniela.Costea@uib.no).

N. Filipovic is with the Faculty of Engineering, University of Kragujevac, Serbia and with the BioIRC, Bioengineering R&D Center, Kragujevac, Serbia (e-mail: fica@kg.ac.rs).

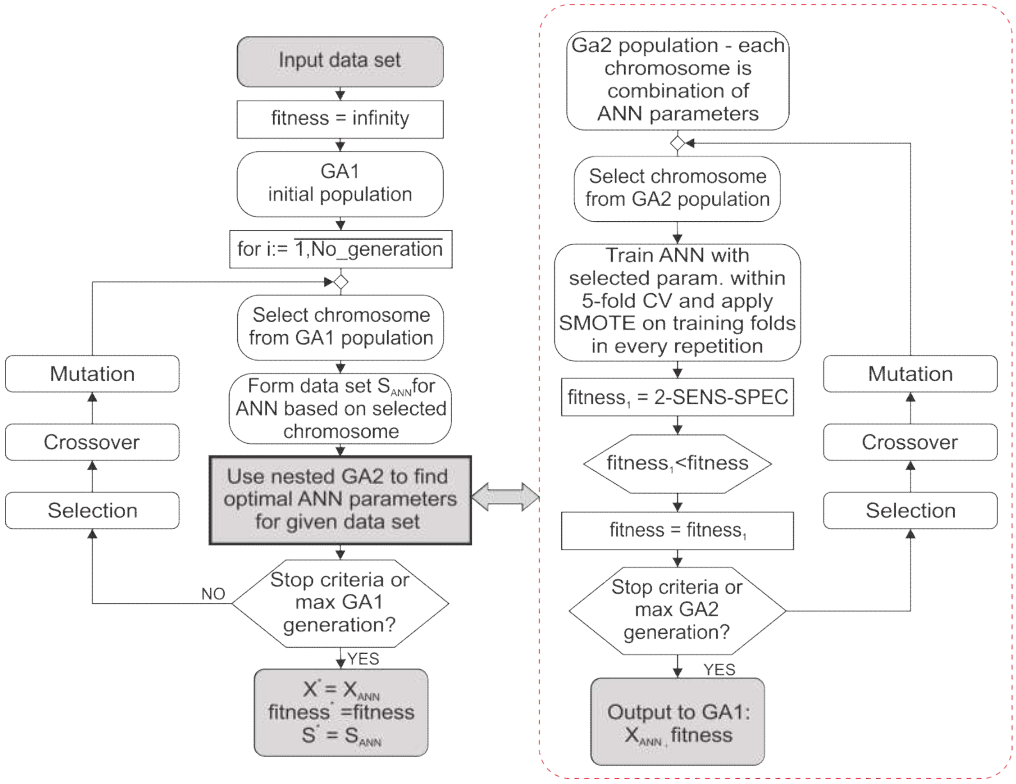

Figure 1. Overall procedure workflow. XANN is the vector of ANN parameters; X* is the optimal combination of ANN parameters; S* is the optimal subset of features

of examples. In many real classification problems, large amount of data is generated with skewed distribution and, in medicine, right decision making for minority class examples is often of key importance. Thus, it is crucial to develop the classifier which will be capable of achieving high accuracy on majority (high specificity) as well as minority class samples (high sensitivity). Various techniques in the machine learning field are developed to deal with this problem on data and algorithms level.

Genetic algorithms (GA) [14] have the potential to generate both the optimal feature subset and ANN parameters at the same time keeping the accuracy balanced for both classes. By defining the fitness function to include metrics for the minority and majority classes' performances, we can combine different objectives into a single fitness value. Our research objective was to perform optimization of the ANN parameters and feature subset simultaneously based on evolutionary principles taking into account the problem of the unbalanced dataset. In this manner, we contribute to the identification of a limited subset of factors that are highly correlated with the occurrence of OSCC second primary tumors, thus, formulating the disease profile.

The rest of paper is structured as follows. Section 2 describes the data set used in the experiment and methodology. Results and discussion are presented in Section 3, while conclusions and directions for further research are described in Section 4.

2. METHODS

2.1 Classification Dataset

The study involves patients with OSCC. Based on literature, all patients were described with immunohistochemical and clinical parameters obtained within the 5-year study. Basic statistics of data is presented in the Appendix. The database was developed for the research project purposes and all patient information was anonymized. The patient population with complete prognostic information consisted of 95 records and 23 attributes which represent the combination of values with different nature.

In order to make this database more convenient for processing, some data transformation was made. The attributes T, N, and Stage, according to their nature, were transformed into numeric values so that T and Stage values fall within the range 1 – 4, and values for N attribute within the range 0 - 3. Additionally, the rest of nominal attributes were binarized [15]. The final dataset that represented the input to machine learning consisted of 95 instances described by 30 attributes.

2.2 Machine Learning Methodology

An Artificial Neural Network (ANN) constructed

as a multilayer perceptron is an algorithm inspired by human brain and by the way the brain processes information. Each unit called node or neuron in the ANN is highly interconnected with other units. It receives input from some other nodes or from an external source and computes an output. Each input or connection has an associated weight (**w**), which is assigned on the basis **b** according to its relative importance to other inputs. The node applies an activation function **f** to the weighted sum of its inputs $\mathbf{x_i}$ in order to calculate the output **s** (Equation 1):

$$s = f\left(\sum_{i}^{n} x_i w_i + b\right) \quad (1)$$

The main role of this function is to make ANN non-linear allowing it to perform complex mapping of input data space. The question is how to learn a multilayer perceptron to understand the intelligence hidden in the data. Two aspects of this problem can be examined: determining the network structure as well as the learning of the connections weights. For defined network structure, supervised back-propagation learning is relatively simple algorithm for correctly setting of weights and biases. However, the identification of appropriate network structure requires the configuration of a number of parameters among which are a number of nodes in the hidden layer, the type of activation function, the learning algorithm, etc. According to the literature [16], a multilayer perceptron with a single hidden layer is the most suitable for the purposes of binary classification.

The optimal ANN topology and other network parameters vary depending on the selected subset of attributes (features). We proposed simultaneously selection of features and ANN parameters optimization using GA method following the procedure described in **Figure 1**. The proposed algorithm is developed under the Matlab 2013 environment.

The implementation of the proposed algorithm is based on wrapper approach for feature selection. For the input data set which consists of N features, a feature subset can be represented as n-dimensional vector of ones and zeros, so that value '1' or '0' indicates if the feature at that particular location is selected or not. The GA starts the optimization from the initial population which in our case represented the random combination of features. It did not involve the additional processes such as the statistical screening on the original dataset which ensured avoiding the biased result. For every combination of inputs nested GA was started in order to find the best structure and parameters of ANN. Classifier was evaluated using 5-fold cross validation procedure. This means that a dataset was split into 5 disjunctive subsets and, in 5 iteration steps procedure, 4 folds were used for training, and the remaining one was used for testing. It is important to emphasize that in this process of training the classifier, training data was balanced using the SMOTE algorithm [17] in order to tackle the problem of unbalanced dataset. This algorithm represents synthetic oversampling technique for minority class based on k-NN approach to create new instances similar to the already available ones. Test instances were processed without changing its structure.

ANN parameters that we considered for optimization were: number of nodes in the hidden layer (in range 5 - 60), number of epochs (50 - 1000), activation function for nodes in the hidden layer, as well as for output layer (tansig, logsig, purelin), learning algorithm (traingd, traingda, traingdm, traingdx) and, depending on the selected algorithm, learning rate (0.001 – 0.3) and momentum constant (0.1-0.9). Within the procedure for ANN optimization, a percentage for oversampling the minority class was optimized as well, taking values from 100% - 800%.

The fitness function was based on the performance of ANN in a way that ensures that accuracy of both class samples, sensitivity (Sens) and specificity (Spec) will be taken into account. The fitness function was assessed by the following formula (Equation 2):

$$fitness = 2 - Sens - Spec \quad (2)$$

Details about GA are:

- integer coding GA;
- tournament selection;
- crossover: scattered;
- crossover probability: 0.8;
- population size 20 individuals;
- number of generation 100.

3. RESULTS AND DISCUSSION

Following the previously described methodology, in this section we present results. The best achieved value for *fitness is 0.192 (Sens = 0.889, Spec = 0.919)* for 14 selected attributes shown in Table 1 and following ANN parameters: learning algorithm – gradient descent with momentum and adaptive learning rate back-propagation, activation function for nodes in the hidden as well as output layer – tansig, number of nodes in the hidden layer - 26, number of epochs - 346, learning rate – 0.051, momentum constant – 0.719 and percentage of artificially created data – 400%.

Additionally, we studied the influence of each attribute according to selection frequency in each generation. Higher frequency indicates more important rank of features in comparison to

Table 1. Selected features

Feature No	Name
f2	WPI
f5	LoxL4tumor
f7	LoxL4stroma
f12	Sex
f13	Site_BUCCAL MUCOSA
f14	Site_ALVEOLUS
f15	Site_TONGUE
f16	Site_Other
f17	T
f20	Thickness
f21	Bone
f22	Skin
f24	LVI
f26	Therapy_CTRT

others (**Figure 2**).

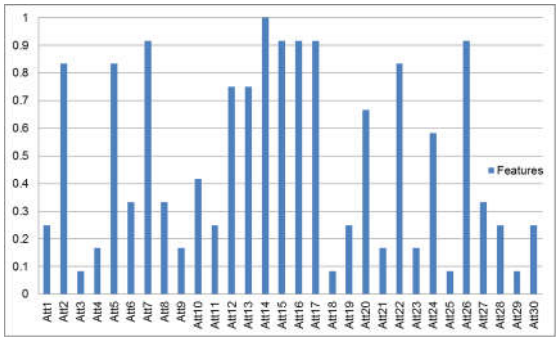

Figure 2. Relative frequency of features during the proposed procedure

We compared the proposed method with three state-of-the-art algorithms for feature selection: ReliefF [18], mRMR [19] and Gain ratio [15], widely used for extraction of useful features from data. The ReliefF algorithm is based on the idea that useful features should differentiate between instances from different classes and have similar values from instances from the same class. It randomly choses an instance from the dataset, finds its nearest neighbor from the same and opposite class, and updates relevance score for each feature by comparing the values of nearest neighbors to the sampled instance. The mRMR algorithm maintains the features that have high correlation with the class attribute and low inter-correlation among themselves. Gain ratio (GR) normalizes the information gain score of splitting on an attribute by the entropy of this attribute.

For evaluation of these feature selection approaches we used the following classifiers: Support vector machine (SVM), Decision Tree (DT) and Artificial Neural Network (ANN). The classifiers were tested within the 5-fold cross validation procedure. In every iteration step of this procedure, only training fold is used for oversampling the minority class with SMOTE algorithm and selection of features. Accuracy (AC), Sensitivity (Sens), Specificity (Spec) and area under ROC curve (AUC) were used to measure the performances of the classifiers. Results are presented in **Table 2**.

In comparison to classifiers from **Table 2**, proposed *ANN_GA method achieved the following results: Ac = 0.916, Sens = 0.889, Spec = 0.919, AUC = 0.903.* Also, from this table, we can clearly conclude that our proposed method outperforms standard classification and feature selection algorithms. The ROC curve of the proposed method ANN_GA is depicted in **Figure 3**.

Table 2. Performances of classifiers ANN, SVM and DT using best k ranked features

		ANN				SVM				DT			
		Ac	Sens	Spec	AUC	Ac	Sends	Spec	AUC	Ac	Sens	Spec	AUC
k=20	ReliefF	0.779	0.333	0.826	0.694	0.768	0.222	0.826	0.524	0.811	0.000	0.895	0.578
	MRMR	0.738	0.200	0.791	0.381	0.791	0.400	0.894	0.500	0.831	0.300	0.837	0.457
	Gain Ratio	0.758	0.333	0.802	0.601	0.779	0.333	0.826	0.579	0.811	0.000	0.895	0.578
k=15	ReliefF	0.779	0.333	0.826	0.672	0.800	0.444	0.837	0.641	0.811	0.000	0.895	0.578
	MRMR	0.728	0.100	0.792	0.589	**0.780**	**0.600**	**0.918**	**0.500**	0.841	0.200	0.803	0.500
	Gain Ratio	0.747	0.000	0.826	0.629	0.684	0.222	0.733	0.477	0.811	0.000	0.895	0.569
k=14	ReliefF	**0.747**	**0.444**	**0.779**	**0.689**	0.716	0.333	0.756	0.545	0.789	0.000	0.872	0.557
	MRMR	0.728	0.100	0.792	0.612	0.814	0.500	0.850	0.500	0.810	0.200	0.884	0.647
	Gain Ratio	0.768	0.111	0.837	0.570	0.726	0.333	0.767	0.550	0.789	0.000	0.872	0.548
k=10	ReliefF	0.663	0.111	0.721	0.538	0.674	0.333	0.709	0.521	0.800	0.000	0.884	0.525
	MRMR	0.758	0.100	0.825	0.740	0.791	0.400	0.955	0.500	0.831	0.000	0.931	0.796
	Gain Ratio	0.705	0.111	0.767	0.575	0.758	0.222	0.814	0.518	0.779	0.000	0.860	0.547

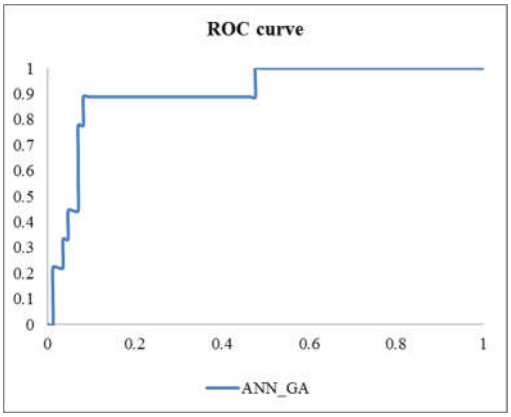

Figure 3. ROC curve of proposed ANN_GA method

4. CONCLUSION

We have presented the intelligent method for prediction of second primary cancer in OSCC patients. The method is based on hybrid approach which combines Genetic Algorithm and Artificial Neural Network, tailored to select the most useful features as well as the optimal ANN parameters. The method is tested on the dataset with real patients' data which suffer from OSCC and showed significant improvement regarding the ability to predict the occurrence of second primary tumor in comparison to other classification methods.

APPENDIX

The database of OSCC patients is described in the following table (Table 3)[1].

Table 3. OSCC database

No	Variables	Statistics	No	Variables	Statistics
f1	p16-HPVstatus (Binary)	0 (89) 1 (6)	f13	Site (Nominal)	Buccal mucosa (36) Alveolus (20) Tongue (31) Other (8)
f2	WPI (Numeric)	Min: 1 Max 5 Mean: 3.28 StdDev: 0.85	f14	T (Nominal)	T1 (25) T2 (25) T3 (8) T4 (37)
f3	Ki67% (Numeric)	Min: 3.1 Max: 67.8 Mean: 21.30 StdDev: 14.26	f15	N (Nominal)	N0 (60) N1 (15) N2B (16) N2C (4)
f4	p53(Numeric)	Min: 0 Max: 2 Mean: 0.64 StdDev: 0.50	f16	Stage (Nominal)	I (17) II (17) III (12) IVA (49)
f5	LoxL4tumor (Binary)	0 (32) 1 (63)	f17	Thickness (Numeric)	Minimum: 2 Maximum: 45 Mean: 13.35 StdDev: 9.28
f6	aSMAstroma (Binary)	0 (39) 1 (56)	f18	Bone (Binary)	1 - Involved (29) 0 - NA (66)
f7	LoxL4stroma (Binary)	0 (61) 1 (34)	f19	Skin (Binary)	1 - Involved (8) 0 - NA (87)
f8	FVIII (Numeric)	Min: 4 Max: 159 Mean: 69.72 StdDev: 39.70	f20	PNI (Binary)	1 – Yes (17) 0 – No (78)
f9	D2-40 (Numeric)	Min: 0 Max: 82 Mean: 25.62 StdDev: 16.64	f21	LVI (Binary)	1 –Yes (1) 0 – No (94)
f10	Foxp3 (Numeric)	Min: 0 Max: 3 Mean: 2.32 StdDev: 0.90	f22	PNE (Binary)	1 – Yes (25) 0 – No (70)
f11	Age (Numeric)	Min: 24 Max: 77 Mean: 55.57 StdDev: 26.66	f23	Therapy (Nominal)	CTRT (14) PORT (14) RT (27) NK (21) Without (19)
f12	Sex (Binary)	0 - Male (79) 1 - Female (16)	f24	Second Primary	0 – No (86) 1 – Yes (9)

[1] After binarization of nominal features, feature f13 from the Table 3 is transformed into 4 binary features f13 – f16 and feature f23 is transformed into 5 binary features f26 – f30. Features f14 – f23 from this table consequently changed the order to f17-f25.

REFERENCES

[1] Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A. “Global Cancer Statistics, 2012”, *CA: A Cancer Journal for Clinicians*, 2015, 65, 87-108

[2] Wang, B., Zhang, S., Yue, K., & Wang, X.-D. “The recurrence and survival of oral squamous cell carcinoma: a report of 275 cases”, Chinese Journal of Cancer, 2013, 32(11), 614–618. http://doi.org/10.5732/cjc.012.10219

[3] Tabor MP, Brakenhoff RH, Ruijter-Schippers HJ, et al. “Multiple head and neck tumors frequently originate from a single preneoplastic lesion”, *Am J Pathol*, 2002;161:1051–60

[4] Forastiere, R. Weber, and K. Ang, "Treatment of head and neck cancer," *N Engl J Med*, 2008, vol. 358, pp. 1076; author reply 1077-8

[5] C.-L. Chang, M.-Z. Hsu. “The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer Original Research Article”, *Expert Systems with Applications*, 36(7) (2009) 10663-10672.

[6] H. Kato, M. Kanematsu, X. Zhang, M. Saio, H. Kondo, S. Goshima, H. Fujita. “Computeraided diagnosis of hepaticfibrosis: preliminary evaluation of MRI texture analysis using thefinite difference method and an artificial neural network”, *AJR Am J Roentgenol*, 189 (2007) 117–122.

[7] A. Cucchetti, M. Vivarelli, N.D. Heaton, S. Phillips, F. Piscaglia, L. Bolondi, G. La Barba, M.R. Foxton, M. Rela, J. O'Grady, A.D. Pinna. “Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease”, *Gut.* 56 (2007) 253–258.

[8] A. Das, T. Ben-Menachem, G.S. Cooper, A. Chak, M.V. Jr. Sivak, J.A. Gonet, R.C. Wong. ”Prediction of outcome in acute lower-gastrointestinal haemorrhage based on an artificial neuralnetwork: internal and external validation of a predictive model”, *Lancet*, 362 (2003) 1261–1266.

[9] R. Mofidi, M.D. Duff, K.K. Madhavan, O.J. Garden, R.W. Parks. “Identification of severe acute pancreatitis using an artificial neural network”, *Surgery*. 141 (2007) 59–66.

[10] F. Piscagli, A. Cucchetti, S. Benlloch, M. Vivarelli, J. Berenguer, L. Bolondi, A.D. Pinna, M. Berenguer, Prediction of significantfibrosis in hepatitis C virus infected liver transplant recipients by artificialneural network analysis of clinical factors, Eur. J.Gastroenterol.Hepatol. 18 (2006) 1255–1261.

[11] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, “Particle swarm optimization for parameter determination and feature selection of support vector machines,”*Expert Systems with Applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[12] C.-L. Huang, “ACO-based hybrid classification system with feature subset selection and model parameters optimization,” *Neurocomputing*, vol. 73, no. 1–3, pp. 438–448, 2009.

[13] H. Fröhlich, O. Chapelle “Feature selection for support vector machines by means of genetic algorithms**”,** *Proceedings of the 15th IEEE international conference on tools with artificial intelligence*, Sacramento, CA, USA (2003) pp. 142–148

[14] Goldberg D. „Genetic Algorithms in Search, Optimization and Machine Learning“, New York: Addison Wesley, 1989

[15] Kononenko I and Kukar M. “Machine Learning and Data Mining: Introduction to Principles and Algorithms”, *Horwood publ*. 2007.

[16] G.P. Zhang. ”Neural networks for classification: a survey”, *IEEE Transactions on Human-Machine Systems*. (2000); 30(4) 451-462. doi: 10.1109/5326.897072

[17] N. Chawla, K. Bowyer, L. Hall, andW. Kegelmeyer, “SMOTE: Syntheticminority over-sampling technique”, *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[18] Robnik-Sikonja M, Kononenko I. “An adaptation of Relief for attribute estimation in regression”, *Fourteenth International Conference on Machine Learning,* 1997; 296-304.

[19] Peng, H. Long, F. Ding, C. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2005, vol. 27, no. 8, 1226-1238