Classification of Serous Ovarian Tumors Based on Microarray Data Using Multicategory Support Vector Machines

Jee Soo Park, Soo Beom Choi, Jai Won Chung, Sung Woo Kim, Deok Won Kim, *Life member, IEEE**

Abstract— **Ovarian cancer, the most fatal of reproductive cancers, is the fifth leading cause of death in women in the United States. Serous borderline ovarian tumors (SBOTs) are considered to be earlier or less malignant forms of serous ovarian carcinomas (SOCs). SBOTs are asymptomatic and progression to advanced stages is common. Using DNA microarray technology, we designed multicategory classification models to discriminate ovarian cancer subclasses.**

To develop multicategory classification models with optimal parameters and features, we systematically evaluated three machine learning algorithms and three feature selection methods using five-fold cross validation and a grid search. The study included 22 subjects with normal ovarian surface epithelial cells, 12 with SBOTs, and 79 with SOCs according to microarray data with 54,675 probe sets obtained from the National Center for Biotechnology Information gene expression omnibus repository.

Application of the optimal model of support vector machines one-versus-rest with signal-to-noise as a feature selection method gave an accuracy of 97.3%, relative classifier information of 0.916, and a kappa index of 0.941. In addition, 5 features, including the expression of putative biomarkers SNTN and AOX1, were selected to differentiate between normal, SBOT, and SOC groups. An accurate diagnosis of ovarian tumor subclasses by application of multicategory machine learning would be cost-effective and simple to perform, and would ensure more effective subclass-targeted therapy.

I. Introduction

Ovarian cancer, the primary cause of death due to gynecological malignancies, is the fifth leading cause of cancer death in women in the United States. Serous ovarian carcinoma (SOC) is the most common histological subtype [1]. Serous borderline ovarian tumors (SBOTs), a subtype of ovarian-surface epithelial stromal tumors, are considered to be an earlier or less malignant form of SOC with a better prognosis [2]. However, because early stage ovarian cancer is mostly asymptomatic, understanding of the etiology is poor, and biomarkers for the disease are unreliable, most patients are diagnosed at the advanced stage [3]. The identification of useful molecular biomarkers of early stage ovarian cancer would improve screening and diagnosis, potentially improving the prognosis of the disease.

Based on a comparison of gene expression levels, microarrays can simultaneously analyze tens of thousands of genes on a genomic scale, providing useful biological, diagnostic, and prognostic information [4]. The application of machine learning and statistical techniques to a microarray data set can be used in a variety of class discovery or class prediction biomedical problems, including those relevant to tumor classification.

DNA microarray technology with machine learning has emerged as a promising tool for accurate diagnosis and classification. Machine learning for DNA microarrays involves using tumor gene expression data to select a discriminative set of genes related to classification, termed the learned classifier. A new input data is then screened using the machine-learned classifier [5]. For cancer classification, several machine learning techniques have been developed on the basis of gene expression profiling data. Ovarian cancer classification would greatly benefit from this technology given its lack of characteristic clinical symptoms.

The sequential progression of SBOTs to SOCs has not been well established, but such carcinomatous changes and their associated molecular events are being thoroughly investigated [3]. In this study, we developed a multicategory classification model for a reliable and discriminative diagnosis of ovarian cancer by testing three machine learning algorithms for DNA microarray. Furthermore, novel biomarkers for classifying ovarian tumor subclasses were identified.

II. Materials and Methods

A. Data Acquisition

We collected 113 raw DNA microarray data sets on ovarian tumor subclasses from the gene expression omnibus repository at the National Center for Biotechnology Information. The data sets originated from Affymetrix HG_U133 Plus 2.0 GeneChips comprised of 54675 probe sets, representing 20599 well-characterized human genes.

We included 22 subjects with normal ovarian surface epithelial cells, 12 with SBOTs, and 79 with SOCs [6]-[9]. Since SOC is the most common histological ovarian cancer subtype [1], we chose to analyze SOC and SBOT (Table I).

For machine learning, the raw DNA microarray data format was converted into the MATLAB data format. In

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2012R1A2A2A03045612).

J. S. Park is with Dept. of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: sampark@yuhs.ac).

S. B. Choi is with Brain Korea 21 PLUS Project for Medical Science, Yonsei University, Seoul, Korea (e-mail: plains7@yuhs.ac).

J. W. Chung is with the Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Korea (e-mail: chjw0915@yuhs.ac).

S. W. Kim is a senior research engineer with Future IT R&D Laboratory, LG Electronics Advanced Research Institute, Seoul, Korea (e-mail: sungwooloo.kim@lge.com).

*D. W. Kim is a Professor at Dept. of Medical Engineering, Yonsei University College of Medicine, Seoul, Korea (corresponding author; phone: 82-2-2228-1916; fax: 82-2-364-1572; e-mail: kdw@yuhs.ac).

TABLE I NUMBER OF DATASETS IN THE NORMAL, SBOT, AND SOC GROUPS

Reference	Normal	SBOTs	SOCs
Elgaaen BV et al [6]	4	4	4
King ER et al [7]	6	8	35
Bowen NJ et al [8]	12	-	12
Wu Y et al [9]	-	-	28
Total	22	12	79

Normal = normal ovarian surface epithelial cells, SBOTs = serous borderline ovarian tumors, SOCs = serous ovarian carcinomas.

particular, we performed a robust multi-array average procedure including $\log_2$ transformation and quantile normalization. The data were divided randomly into training and testing sets. The training set, comprising 66.7% (76 subjects) of the overall dataset, was used to construct models using three machine learning algorithms. The testing set, comprising 33.3% (37 subjects) of the overall dataset, was used to assess the model's ability to categorize subjects into the normal, SBOTs, and SOCs groups (Fig. 1).

B. Feature selection

Feature selection was necessary to reduce the high dimensionality of the datasets. Regardless of the classification methods, many microarray-based studies suggest that gene selection is vital for achieving a high level of generalization [10]. Feature selection is used to identify genes that might be informative for prediction by statistical and machine learning methods [5]. All gene probe sets were ranked by the feature selection method according to the calculated weight of each gene. We used three feature selection methods: (1) the ratio of genes between-categories to within-category sums of squares (BW); (2) the signal-to-noise (S2N) scores applied in a one-versus-one (S2N-OVO) model; and (3) the S2N scores applied in a one-versus-rest (S2N-OVR) model. These were chosen because they are widely used in multicategory classifications and represent different approaches for gene ranking and selection. Dudoit et al. proposed the BW sum of squares across all paired classes for multicategory classification [11]. S2N is calculated by dividing the difference of the means of two groups by the sum of the standard deviations of those two groups.

C. Machine learning

We used three multicategory support vector machines (MC-SVMs) based on the binary SVM method: SVM one-versus-one (SVM-OVO), SVM one-versus-rest (SVM-OVR), and directed acyclic graph SVM (DAGSVM). Binary SVMs are learning and pattern recognition algorithms developed with the goal of separating classes by a function that is computed from available examples. The goal is to find a hyper plane that maximizes the separation or margin between two classes. To solve multicategory problems using machine learning, classification methods use combinations of binary classifiers.

The SVM-OVO method involves construction of binary SVM classifiers for all class pairs. In other words, for every class pair, a binary SVM problem is solved, and then the

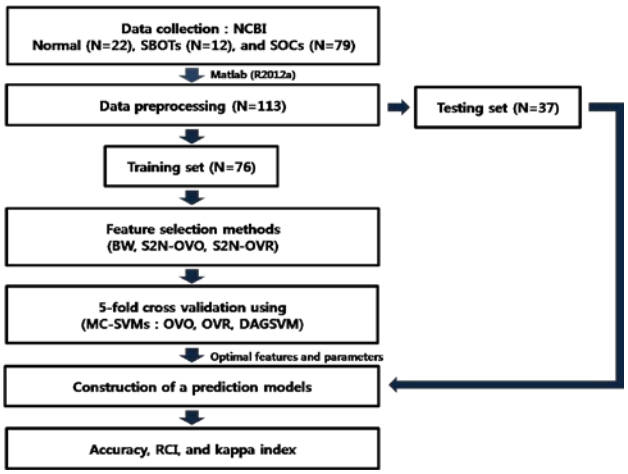

Figure 1. Flow diagram for serous ovarian tumor classification

decision function assigns an instance to a class that has the largest number of votes (the so-called Max Wins strategy). The SVM-OVR method constructs k binary SVM classifiers, where k is the number of classes. The combined OVR decision function chooses the class of a sample that corresponds to the maximum value of k binary decision functions specified by the furthest hyper plane. In DAGSVM, the training phase of the algorithm is similar to that in the OVO approach using multiple binary SVM classifiers. However, the testing phase of DAGSVM requires construction of a rooted binary decision-directed acyclic graph using classifiers [10].

D. Model selection and validation

After determining the order of the variables using the feature selection methods, we identified the optimal variables with which to construct classification models by increasing the number of variables in the order of their importance using sequential forward selection (SFS) as the wrapper method [12].

A grid search, in which a range of parameter values were tested using the 5-fold cross validation strategy, was applied (Fig. 1). For use with the MC-SVM method, we chose a radial basis function among the available kernel functions according to the recommendation of a practical guide [13]. The parameter values included a penalty parameter (C) and scaling factor (σ) for MC-SVMs. The best classification model was chosen and employed for prediction.

The diagnostic ability of the model based on accuracy, relative classifier information (RCI), and the kappa index for the testing set was determined. RCI, a parameter of an entropy-based measure of classifier performance, can be measured by the difference in the prior and posterior uncertainties. The kappa index is a statistical measure of inter-rater agreement or inter-annotator agreement for categorical items. We used MATLAB 2012a (Mathworks Inc. Natick, MA) to analyze machine learning and SPSS 20.0 (SPSS Inc., Chicago, IL) for statistical analysis.

III. Result

Table II shows the top five ranked genes for each feature selection method as applied to the training set. Among the ranked genes, sentan, cilia apical structure protein (SNTN) and aldehyde oxidase 1 (AOX1) were included in the five genes used by the best performing SVM model, SVM-OVO. Moreover, these two genes were simultaneously selected by at least two different feature selection methods among the top five ranked genes.

Table III summarizes the overall classification performance of the three MC-SVM models when each of the three feature selection methods was applied to the testing set. SVM-OVR with S2N-OVO as the feature selection method gave the best accuracy of 97.3%, an RCI of 0.916, and a kappa index of 0.941, using only five features. One, two, and three misclassifications out of 37 testing data sets gave accuracies of 97.3%, 94.6%, and 91.9%, respectively. The S2N-OVO model also required the fewest number of features (5) for the best performance; the BW and S2N-OVR used 10 and 14–20 features, respectively.

The optimal model of SVM-OVR using the S2N-OVO feature method were found using a radial basis function with a penalty parameter (C) of 1 and scaling factor (σ) of 0.01. Additionally, although there were not shown in table III, for the training set, the accuracy of the cross-validation when using the optimal parameters and features with the S2N-OVO method were 96.1%, 98.7%,and 97.3% for SVM-OVO, SVM-OVR, and DAGSVM, respectively; suggesting that the constructed models were not over-fitted.

Fig. 2 shows the gene expression rates of SNTN and AOX1 between the three data groups. The mean (standard deviation) SNTN expression rates were 4.2 (1.7), 8.2 (0.6), and 3.2 (1.4), in the normal, SBOT, and SOC groups, respectively. The mean AOX1 expression rates were 7.8 (1.2), 5.4 (0.8), and 5.0 (1.3), respectively. Intergroup differences were determined using repeated measures analysis of variance with the Bonferroni correction. The normal group showed high-level expression of AOX1, whereas SBOTs and SOCs showed low-level expression. AOX1 expression could be used to differentiate the normal group from the cancerous groups ($p < 0.001$), but could not be used to distinguish SBOTs from SOCs ($p = 0.783$). In contrast, SNTN was highly expressed in SBOTs whereas normal and SOCs groups showed low-level gene expression. In addition, SNTN expression was significantly different between all groups ($p < 0.01$), especially the SBOT and SOC groups ($p = 1.66 \times 10\text{-}19$). Therefore, we hypothesized that combinatorial assessment of the expression of both genes in a sample would accurately differentiate between normal cells and early and advanced subtypes.

TABLE II The top five ranked genes for each feature selection method used in the training set (N = 76)

Gene symbol		
BW	***S2N-OVO***	***S2N-OVR***
ITLN1	**SNTN**	PGR
LHX9	LOC100506777	ABCA8
MUC1	**AOX1**	FLRT2
SNTN	CAPS	**AOX1**
AOX1	TPPP3	ITLN1

BW = ratio of genes between-categories to within-category sums of squares, S2N = signal-to-noise scores, OVO = one-versus-one, OVR = one-versus-rest.

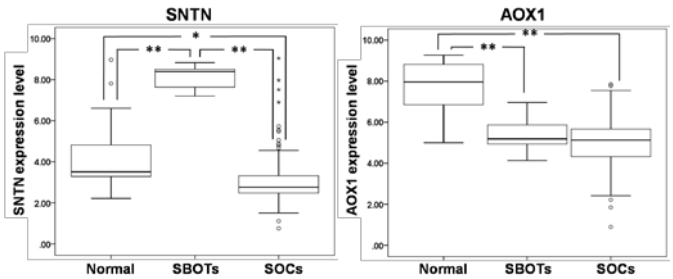

Figure 2. Box plots of the two gene expression levels for the normal (n=22), SBOTs (n=12), and SOCs (n=79) (a) SNTN and (b) AOX1 with a $\log_2$ scale.

IV. Discussion and Conclusion

Using three machine learning algorithms for multicategory classification of DNA microarray data, we determined an optimal approach to support a clinical diagnosis of ovarian tumor subclasses. This is the first study to construct a multicategory classification diagnosis model for ovarian cancer, which is particularly useful for detecting asymptomatic early-stage ovarian cancer and determining the prognosis of patients at risk of developing advanced stage ovarian cancer. The best accuracy of our model for subclass diagnosis was 97.3%, with a good discriminative ability. This model would be helpful in screening patients with the risk of developing advanced-stage ovarian cancer.

The current diagnostic tools for ovarian cancer are generally limited to biochemical tests, which require human and material resources and carry the risk of an inaccurate diagnosis. With the current model, ovarian tumors could be quickly classified into subclasses with a single biopsy using computation analysis of gene expression, providing a time and

TABLE III Performance of each machine learning method with each feature selection method when applied to the testing set

Algorithm		Feature selection methods (n = 37)											
		BW				***S2N-OVO***				***S2N-OVR***			
		Acc (%)	***RCI***	***Kappa***	***No. of features***	***Acc (%)***	***RCI***	***Kappa***	***No. of features***	***Acc (%)***	***RCI***	***Kappa***	***No. of features***
MC-SVM	OVO	91.9	0.611	0.804	10	94.6	0.711	0.874	5	94.6	0.781	0.879	14
	OVR	94.6	0.781	0.879	10	**97.3**	**0.916**	**0.941**	**5**	97.3	0.916	0.941	20
	DAGSVM	91.9	0.611	0.804	10	94.6	0.711	0.874	5	94.6	0.781	0.879	14

BW = ratio of genes between-categories to within-category sums of squares, S2N = signal-to-noise scores, OVO = one-versus-one, OVR = one-versus-rest, ACC = accuracy, RCI = relative classifier information, Kappa = kappa index, MC-SVM = multicategory support vector machine, DAGSVM = directed acyclic graph SVM.

cost-effective accurate diagnosis.

The SVM-OVR classification model, when used with S2N-OVO, yielded the highest accuracy in discriminating ovarian tumor subclasses using only five features. The fewer the features needed for selection, the more efficient is the model. The use of fewer features reflect a reduction in the high-dimensional microarray data, minimizing the need for high-throughput arrays and the synthesis of a large number of polymerase chain reaction primer sets that would drastically increase screening costs.

The high performance of SVM models is attributable to their efficiency in finding a unique optimal solution, their flexibility in incorporating multiple data types, and their ability to model nonlinear patterns. Moreover, SVMs perform well in various areas of biological analysis, and are well suited to high-dimensional data such as microarray data. MC-SVMs based on binary SVMs can also accurately classify a gene expression data set. In MC-SVMs, however, when the number of classes increases, the complexity of the overall classifier also increases.

Although further studies are necessary for validation of the SNTN and AOX1 genes identified as ovarian cancer biomarkers, the preliminary results show great promise, as has been seen with the oncotype molecular test for breast cancer, which uses a 21-gene molecular signature to diagnose breast cancer types [14]. SNTN and AOX1 were included in the five genes used by the best performing SVM model and simultaneously selected by at least two different feature selection methods. Changes in SNTN expression have been associated with pathological and cancerous phenotypes. King et al. [7] demonstrated differential gene expression in high-grade SOCs compared with low-grade SOCs and SBOTs. They found that 122 genes were upregulated in the SBOTs and low-grade SOCs compared with high-grade SOCs, and that SNTN was one of the 20 most commonly upregulated genes [7]. SNTN encodes sentan, an apical structure protein found in the cilia lining the female reproductive tract [15]. This could explain why SNTN was identified as a biomarker for early ovarian cancer in the current study. The AOX1 gene has not been much investigated and it should be further study.

Clinicians can use this machine learning model to objectively differentiate ovarian tumor subclasses with high accuracy. This accurate diagnostic support tool can reduce not only the cost and time for diagnosis, but also the risk of progression to more advanced stages. Moreover, if used, it would minimize the application of excessive treatment that may not be appropriate for SBOTs, such as radical ovariectomy and chemotherapy. Early identification of ovarian cancers will allow monitoring of disease progression to an advanced stage and will provide further information on the carcinomatous changes and underlying molecular events that occur. Improved insight into the molecular characteristics of the subclasses of ovarian tumors would eventually lead to more individualized and effective treatments [9].

One of the limitations to this study was the small sample size, because the prevalence of ovarian borderline tumors is very low. Future studies will be directed towards determining the DNA methylation status of these biomarker genes in ovarian cancer subtypes, and further experimental verification of SNTN and AOX1 as clinically useful biomarker genes.

In conclusion, using multicategory machine learning algorithms for DNA microarray data, we identified two biomarker genes SNTN and AOX1 that are likely involved in the pathogenesis and progression of ovarian tumors. Using this information, we generated a cost-effect, accurate diagnostic method that can be applied in the clinic to determine ovarian cancer subtypes, permitting patient-tailored therapy to improve the prognosis of at risk patients.

REFERENCES

[1] J. D. Seidman, I. Horkayne-Szakaly, M. Haiba, C. R. Boice, and R. J. Kurman et al., "The histologic type and stage distribution of ovarian carcinomas of surface epithelial origin," *Int. J. Gynecol. Pathol.,* vol. 23, no. 1, pp. 41-44, Jan. 2004.

[2] E. Banks, V. Beral, and G. Reeves, "The epidemiology of epithelial ovarian cancer: a review," *Int. J. Gynecol. Cancer*, vol. 7, no. 6, pp. 425-438, 1997.

[3] Y. L. Choi, S. Y. Kang, J. S. Choi, Y. K. Shin, and S. H. Kim et al., "Aberrant hypermethylation of RASSF1A promoter in ovarian borderline tumors and carcinomas," *Virchows Arch*., vol. 448, no. 3, pp. 331-336, Nov. 2006.

[4] Z. J. Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," *Artif. Intell. Med.*, vol. 42, no. 1, pp. 81-93, Jan. 2008.

[5] S. B. Cho, and H. H. Won, "Machine learning in DNA microarray analysis for cancer classification." *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, vol. 19, pp. 189-198, 2003.

[6] B. V. Elgaaen, O. K. Olstad, L. Sandvik, E. Ødegaard, and T. Sauer et al., "ZNF385B and VEGFA Are Strongly Differentially Expressed in Serous Ovarian Carcinomas and Correlate with Survival," *PloS one*, vol. 7, no. 9, pp. e46317, Sep. 2012.

[7] E. R. King, C. S. Tung, Y. T. Tsang, Z. Zu, and G. T. Lok et al., "The anterior gradient homolog 3 (AGR3) gene is associated with differentiation and survival in ovarian cancer," *Am. J. Surg. Pathol.*, vol. 35, no. 6, pp. 904-912, Jun. 2011.

[8] N. J. Bowen, L. D. Walker, L. V. Matyunina, S. Logani, and K. A. Totten et al., "Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells," *BMC Med. Genomics*, vol. 2, no. 71, Dec. 2009.

[9] Y. H. Wu, T. H. Chang, Y.F. Huang, H. D. Huang, and C. Y. Chou, "COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer," *Oncogene*, Aug. 2013. doi:10.1038/onc.2013.307.

[10] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics,* vol. 21, no. 5, pp. 631-643, Mar. 2005.

[11] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 77-87, Mar. 2002.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, Aug. 2007.

[13] C. Hsu, C. Chang, and C. Lin, "A practical guide to vector classification," Dept. Comput. Sci., National Taiwan Univ., Taipei, Taiwan, 2003.

[14] J. A. Sparano, and S. Paik, "Development of the 21-gene assay and its application in clinical practice and clinical trials," *J. Clin. Oncol.*, vol. 26, no. 5, pp. 721-728, Feb. 2008.

[15] A. Kubo, A. Yuba-Kubo, S. Tsukita, S. Tsukita, and M. Amagai, "Sentan: a novel specific component of the apical structure of vertebrate motile cilia," *Mol. Biol. Cell*, vol. 19, no. 12, pp. 5338-5346, Dec. 2008.