Integrating Domain Knowledge in Supervised Machine Learning to Assess the Risk of Breast Cancer

Aniket Bochare, Aryya Gangopadhyay, Yelena Yesha,
Anupam Joshi, Yaacov Yesha
University of Maryland Baltimore County,
Baltimore, USA
(aniketb1, gangopad, yeyesha, joshi, yayesha)@.umbc.edu

Michael A. Grasso
University of Maryland, School of Medicine,
Baltimore, USA
mgrasso@umem.org

Mary Brady
National Institute of Standards and Technology,
Gaithersburg, USA
mary.brady@nist.gov

Napthali Rishe
Florida International University,
Florida, USA
rishen@cis.fiu.edu

***Abstract*— Breast cancer is the most common form of cancer in women. Breast cancer comprises 22.9% of invasive cancers in women and 16% of all the female cancers. Our study takes into consideration data of postmenopausal women of European descent and their single nucleotide polymorphism (SNP) information to assess the risk of developing breast cancer. We used various supervised machine learning and data mining techniques to generate a model for predicting risk of breast cancer in post menopausal women using genomic data, family history, and age. In this paper we propose an approach to select 9 best SNPs using various feature selection algorithms and evaluate binary classifiers performance. We evaluated the performance of binary classifier by adding the domain knowledge of 11 SNPs into the training set and performing classification based on most informative features obtained from feature selection technique. We have also designed an algorithm to incorporate domain knowledge into our machine learning model. Our observations revealed that the machine learning model generated using both the domain knowledge and the feature selection technique performed better compared to the naive approach of classification. It is also interesting to note that, in addition to selecting 9 best SNPs, feature selection resulted in removing age from the set of features to be used for cancer risk assessment, and the machine learning model generated using both feature selection and domain knowledge provided improved performance without using age for prediction, compared with the naïve method that did use both age and family history among the features. We could observe improvement in the accuracy and sensitivity values using domain knowledge learning which could be beneficial for initial screening.**

***Keywords*- Breast Cancer, SNP, Genome-Clinical, Classification, Domain Knowledge, Feature Selection**

I. Introduction

According to Fletcher [1] each year in United States about 210,000 women are diagnosed with breast cancer. The risk of developing breast cancer varies from person to person depending on risk factors. Breast cancer can also occur in women with no observable signs of risk factors. Moreover, the risk of getting breast cancer is higher for women with strong family history. In addition, a breast cancer gene increases the likelihood of getting breast cancer more than any other risk factors. There are many environmental and clinical factors such as older age, family history, race, radiation exposure, density of breast, nulliparity, breast feeding, hormone replacement therapy, weight, etc which increase a person's risk of developing breast cancer [1].

Cancer is a complex and a deadly disease, and its detection in early stages could help to improve the probability of survival. Therefore it is imperative to research the contribution of SNPs in early disease prediction. This will assist doctors in assessing the likelihood of developing breast cancer and in deciding whether order further testing.

In this study we have used various data mining and supervised machine learning techniques for generating a prediction model capable of distinguishing between cases and controls for initial screening. We present statistical analysis of 3 different methods named *Naive SNP Selection Approach*, *Feature Selection Approach* and *Domain Knowledge Integration Approach*. From our observation we could conclude that addition of domain knowledge of SNPs in machine learning procedures was beneficial.

A. SNPs and Personalized Medicine

According to Kong, et al. a single nucleotide polymorphism (SNP) is a location in the human genome which differs from one person to another and may affect the functions of the gene in which it is found. Researchers are trying to understand SNPs due to varying susceptibility of individuals to various diseases. Much attention was received by SNPs as genetic markers since different patients responded differently to various drugs. Hence, researchers are exploring SNPs to provide personalized drugs to individuals depending on their genetic makeup [2].

Engel, Simpson, and Landers address the contribution of SNPs to cancer development [5]. Onay, et al. highlight the fact that SNPs belonging to certain genes increases the susceptibility to breast cancer [3]. Our goal is to use breast cancer associated SNPs as genetic markers for classifying an individual as case or control. But it has been observed in the past that the use of SNPs only as features to develop a prediction model has not yielded satisfactory performance. In this paper we have identified 22 SNPs from SNPedia [6] and use domain knowledge of SNPs to come up with an improved prediction model.

Khoury M. J., et al. have shown that in complex diseases the disease susceptibility may vary with gene-environmental interactions and genes originating from diverse demography [7]. Therefore, we need to select features wisely for diagnosis of such diseases. Hence, we chose 17 SNPs for classification algorithms after initial filtering and pre-processing.

McCarthy, et al. focus on the importance of combinations of SNPs instead of a single SNP in the development of Type II Diabetes [8]. Due to cumulative effect of SNPs, we selected a set of risk associated SNPs for determining genomic risk of an individual. We used 3 different methods in our experiments and considered the cumulative effect of these SNPs to generate a prediction model. In the naive SNP selection approach, we cumulatively used 17 SNPs for classification. In the second method we used feature selection to extract 9 most informative SNPs and used them cumulatively for generating a classification model. In the third method we used 11 SNPs which had risk values associated with them in SNPedia to incorporate domain knowledge into the model.

II. Background and Related Work

A. Genetics and Breast Cancer

SNPedia provides references to different studies conducted by researchers on breast cancer patients and SNPs associated with breast cancer [9]. Therefore, we hope that by delving into the genomic patterns of a population one can detect risks at an early stage and assist physicians.

B. Bioinformatics and Medicine

Bioinformatics is a field where various data mining and machine learning techniques are used for disease prognosis and drug interactions. Many classification approaches have been proposed earlier for disease prediction, such as: Decision Trees (J48), k Nearest-Neighbor (kNN), Naive Bayes (NB), Random forest (RF) and Support vector machine (SVM). These methods had been applied in various fields such as: decisions involving judgment, screening images, load forecasting, marketing and sales and medical diagnosis [10].

C. Domain Knowledge and Machine Learning

According to Yu, et al. auxiliary information about a learning task which can be obtained from credible sources or domain experts is called domain knowledge. They explain that prior domain knowledge helps in selection, initial sanitization and pre-processing tasks involved in machine learning. This is not limited to removal of noise or redundancy but also transforming data using domain knowledge for inputting into our machine learning system. Adding virtual samples to the training set has gained much attention in recent times and it is important when there are not enough training examples to learn from [11]. Partha Niyogi, et al. discusses incorporating domain knowledge by using virtual examples [23] in the learning task [12].

We know that traditional machine learning algorithms do not take into consideration the knowledge about data for training classifiers. Sun, et al. point out that combining prior domain knowledge with training set aids in machine learning. They have demonstrated a novel approach of combining domain knowledge into support vector machine for better efficiency [13]. There are various ways to integrate domain specific information depending on its context and type. Positive domain knowledge is said to have occurred when the use of domain specific information results in a more accurate hypothesis compared to the use of just training examples [11]. We designed an algorithm to incorporate the risk associated with each SNP into the training set for integrating domain knowledge in our model and add virtual examples based on some rules.

III. Clinical-Genomic Data

A. Data Description

In our experiments we have used the `Nurses Health Study (NHS)-GCEMS Stage-1' breast cancer data from dbGaP. The dataset contains post menopausal women of European ancestry out of which 1145 are cases and 1142 controls.

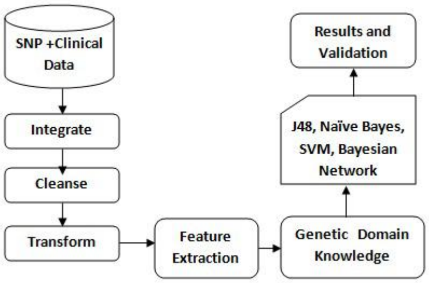


Fig. 1 Framework

The dataset contains mostly genomic information, and also age and family history. MySQL database is used to handle database. We have used the popular machine learning tools WEKA [14] and MATLAB [15] for our experiments. Figure 1 shows the framework of our project.

IV. Classification and Validation

A. *Classification Algorithms*

1) Decision trees: Decision trees are classification trees used in statistics and machine learning to predict a target value of a class based on the attributes or feature space. The leaves represent the classification and the branches represent sets of features which lead to classification. [16] Demonstrates use of the C4.5 decision tree algorithm for classifying patients based on genes and clinical information. J48, a java open source implementation of C4.5 algorithm available in WEKA, was used in our experiments.

2) Naive Bayes: Naive Bayes classifier is used to solve a classification problem based on a probabilistic framework that classifies new samples assuming conditional independence among features under consideration. To classify a new sample, one uses Bayes Rule:

$$P\ (class=Y\mid data=X) = P\ (data=X\mid class=Y)^{*}\ P(class=Y)\ /P\ (data=X) \qquad (1)$$

Naive Bayes has been used before in predicting the risk of breast cancer susceptibility from multiple SNPs. [17] demonstrates accuracy of 56% as compared to the baseline of 50%.

3) Bayesian Networks: Bayesian Network is a probabilistic model of relationships and predictions. Bayesian networks are used widely in the medical field to support prognosis and diagnosis by experts for predicting the outcome of an unknown event. We chose Bayesian Network for our experiments because it analyses dependencies among all the variables through relationships. Bayesian Networks are very powerful and were used in the medical domain for diagnosis and treatment of breast cancer in the past [24].

4) *Support Vector machine - SVM*: Cortes and Vapnik developed the SVM, a supervised learning approach which helps to predict the labels of the test samples from set of positive and negative training samples [18]. SVM attempts to establish a maximum margin for finding the best hyper plane to separate positive and negative samples in Euclidean space. The problem of handling real world data where samples are not linearly separable is handled by choosing a kernel [19]. We used an open source SVM library called LibSVM [20] for our experiments. In this research we have used polynomial kernel of degree 3 and Radial basis function (RBF) for classifying unlabelled instances into cases and controls.

B. *Validation and Accuracy*

1) *10-fold cross validation:* The testing and validation was carried out using 10-fold cross validation [10].

2) *Receiver Operating Characteristics:* Area under ROC is used widely in machine learning and data mining. ROC plots the true positive (TP) rate against the false positive (FP) rate [21]. It is a measure used in machine learning to predict binary classifier's performance.

C. *Sensitivity and Specificity*

The sensitivity (SN) and specificity (SP) are two of the statistical measures used to evaluate a binary classifier. Sensitivity is used in machine learning to measure the proportion of positive instances classified correctly by the model. Similarly, specificity is a proportion of negative instances classified correctly by the model. The following formulae can be used to calculate the specificity and sensitivity.

$$Sensitivity = TP/\ (TP+FN) \qquad (2)$$

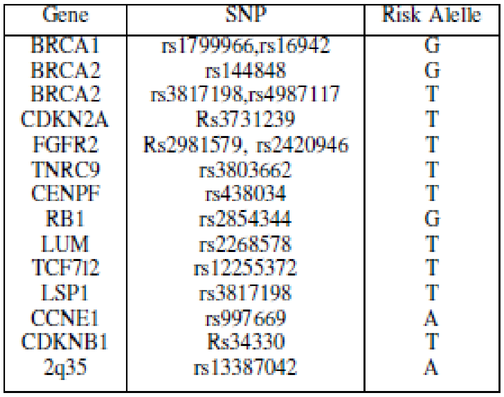
$$Specificity = TN/\ (TN+FP) \qquad (3)$$

V. Methods and Procedures

A. *Classification Using Naive SNP Selection from literature*

Table I shows 17 SNPs considered for our experiments which were obtained from SNPedia. Table II shows the results obtained after performing classification using 17 SNPs as features. This method is called naive approach since the SNPs were neither prioritized nor assigned any weights. We have used 10-fold cross validation to test our model.

TABLE I
BREAST CANCER ASSOCIATED SNPS



Gene	SNP	Risk Alelle
BRCA1	rs1799966,rs16942	G
BRCA2	rs144848	G
BRCA2	rs3817198,rs4987117	T
CDKN2A	Rs3731239	T
FGFR2	Rs2981579, rs2420946	T
TNRC9	rs3803662	T
CENPF	rs438034	T
RB1	rs2854344	G
LUM	rs2268578	T
TCF7l2	rs12255372	T
LSP1	rs3817198	T
CCNE1	rs997669	A
CDKNB1	Rs34330	T
2q35	rs13387042	A

TABLE II
CLASSIFICATION RESULTS USING NAÏVE SNP SELECTION APPROACH

ClassificationAlgorithm	Accuracy	ROC
J48 -Decision Tree	52.30%	0.538
Naive Bayes	55.05%	0.557
LibSVM (Radial Basis)	53.41%	0.530
LibSVM (Polynomial)	52.95%	0.530
Bayesian Network	54.27%	0.566

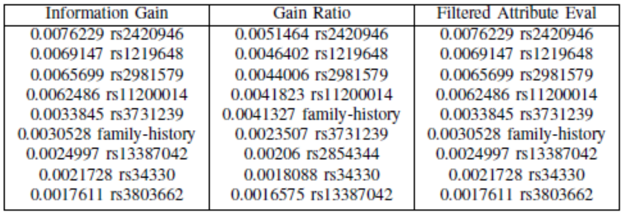
B. *Classification Using Feature Selection*

Feature selection and feature extraction are dimensionality reduction techniques which are mostly used to preprocess the data. They help to reduce the number of features under consideration by eliminating irrelevant ones. Many times it is necessary to narrow down the number of features under consideration for efficient classification. The removal of irrelevant features helps to improve classification accuracy in most of the cases. G. Ustunkar, et al. highlight the importance of selecting a subset of the available SNPs for conducting association studies [22]. Hence, we provide data of 17 SNPs, family history, and age as input data to feature selection techniques for selecting a subset of informative features. Feature Selection (FS) approach is used here to find a subset of the features to improve data quality and remove noisy data.

We used 3 techniques named Filtered Attribute Evaluation, Gain Ratio Attribute Evaluation and Information Gain Attribute Evaluation available in WEKA to extract 9 most informative SNPs from the dataset for binary classification. Table III shows the 9 SNPs obtained using 3 feature selection techniques and table IV shows the results of binary classification using these 9 SNPs. We have used 10-fold cross validation in method I and method II .

TABLE III

ATTRIBUTES RANKED BASED ON INFORMATION GAIN, GAIN RATIO AND FILTERED ATTRIBUTE EVALUATION TECHNIQUE



Information Gain	Gain Ratio	Filtered Attribute Eval
0.0076229 rs2420946	0.0051464 rs2420946	0.0076229 rs2420946
0.0069147 rs1219648	0.0046402 rs1219648	0.0069147 rs1219648
0.0065699 rs2981579	0.0044006 rs2981579	0.0065699 rs2981579
0.0062486 rs11200014	0.0041823 rs11200014	0.0062486 rs11200014
0.0033845 rs3731239	0.0041327 family-history	0.0033845 rs3731239
0.0030528 family-history	0.0023507 rs3731239	0.0030528 family-history
0.0024997 rs13387042	0.00206 rs2854344	0.0024997 rs13387042
0.0021728 rs34330	0.0018088 rs34330	0.0021728 rs34330
0.0017611 rs3803662	0.0016575 rs13387042	0.0017611 rs3803662

Table III results also indicate that rs2420946, rs1219648 and rs2981579 are the top 3 SNPs which appear in all the 3 feature selection techniques. According to SNPedia, these SNPs are really significant markers in European women for breast cancer. It was observed that family-history is also important attribute for assessing risk since it appeared in all the 3 feature selection results.

TABLE IV

CLASSIFICATION RESULTS USING FEATURE SELECTION TECHNIQUE

Classification Algorithm	Accuracy	ROC
J48 -Decision Tree	54.92%	0.559
Naive Bayes	56.39%	0.571
LibSVM(Radial Basis)	56.94%	0.573
LibSVM(Polynomial)	54.57%	0.562
Bayesian Network	55.83%	0.571

C. *Classification Using Domain Knowledge Addition*

We designed the following Algorithm for adding domain knowledge:

Algorithm 1: Add Virtual Instances to the Original Dataset.

Require: <SNP_1…SNPn, age, familyhistory, case> {Original Training Set}. n > 0 {Risk associated SNPs}. $R1_X$ & $R2_X$ {risk values for medium & high risk SNPs from SNPedia}.

```
Assign '0'- norisk, '1'- mediumrisk, '2' - highrisk SNP.
`SNP'$_X$ = (0; 1; 2) {depends on number of risk alleles}.
for $SNP_1$ to $SNP_n$ do
    Let $C1_X$ & $C2_X$ <-row-count and $R1_X$ & $R2_X$<-risk-value when $SNP_X$ = 1 & $SNP_X$ = 2 respectively.
    Add total of $C1_X$ *(1- $R1_X$) & $C2_X$ *(1- $R2_X$) virtual instances where $SNP_X$ = 1 & $SNP_X$ = 2 respectively.
    Add a random row vector $V_i$ as follows:
      < $SNP_1$…$SNP_n$, age, familyhistory > where $SNP_X$ = 1 & $SNP_X$ = 2 queried from the original training set.
      Let $X_1$ and $X_2$ be number of *virtual controls* assigned to class-label of row-vector $V_i$ where $SNP_X$ = 1 & $SNP_X$ = 2 respectively.
          $X_1 + R1_X * X_1 = C1_X$     (4)
          $X_2 + R2_X\ X_2 = C2_X$     (5)
      Solve for $X_1$ and $X_2$. The number of *virtual cases* assigned to the class-label of row-vector $V_i$ are $R1_X * X_1$ & $R2_X * X_2$ where $SNP_X$ = 1 & $SNP_X$ = 2 respectively.
end for
```

We repeat this procedure for all the 11 SNPs selected from SNPedia which have risk associated values. 8 SNPs out of these 11 SNPs were selected, which overlap with the SNPs obtained using feature selection technique along with family-history for classification purpose. We train the classifier using the combination of original and virtual training samples. The validation is conducted on randomly selected 20% test samples from original dataset. The mean results of classification after addition of virtual instances to the dataset across 10 trials are shown in table V. By comparing table II and V we can see around 6-8% increase in prediction accuracy which shows that domain knowledge was helpful. The deviation in the accuracy seen in table V is around ±2.5% across 10 trials.

TABLE V

CLASSIFICATION RESULTS USING BOTH DOMAIN KNOWLEDGE AND FEATURE EXTRACTION

Classification Algorithm	Accuracy	ROC
J48 -Decision Tree	60.56%	0.591
Naive Bayes	60.12%	0.574
LibSVM(RadialBasis)	58.93%	0.53
LibSVM(Polynomial)	59.79%	0.535
Bayesian Network	59.85%	0.588

From our observations we can conclude that although there is an improvement in accuracy, there isn't significant improvement in ROC area using domain knowledge integration to the machine learning model. In this experiment, we also calculated values for statistically important parameters like specificity and sensitivity for all the 5 algorithms used to evaluate the performance of binary classifier. The table VI shows the values of sensitivity (SN) and specificity (SP) for all the 3 methods.

TABLE VI

SENSITIVITY AND SPECIFICITY RESULTS ACROSS 3 METHODS USING 10FOLD CROSS VALIDATION

Sensitivity and Specificity Comparison						
	Method I		Method II		Method III	
Classification Algorithm	SP	SN	SP	SN	SP	SN
J48 -Decision Tree	0.566	0.536	0.579	0.531	0.301	0.798
Naive Bayes	0.579	0.519	0.603	0.523	0.227	0.853
LibSVM(Radial Basis)	0.531	0.530	0.609	0.533	0.145	0.922
LibSVM(Polynomial)	0.664	0.414	0.750	0.332	0.097	0.946
Bayesian Network	0.548	0.549	0.563	0.556	0.226	0.855

Comparing specificity and sensitivity values of Method III with Method I or Method II, we can observe a marked difference in the sensitivity and specificity values. This demonstrates that addition of virtual instances or domain knowledge into the model helps to increase the sensitivity of a test. Breast cancer is a deadly disease and a highly sensitive test is considered very important. Along with an increase in sensitivity there is a simultaneous decrease in specificity using the method III prediction model. The results obtained using domain knowledge model demonstrates a balanced tradeoff between sensitivity and specificity in case of J48, Naive Bayes and Bayesian Network classifiers. This fact, and the fact that using domain knowledge resulted in improved accuracy, are both important.

VI. CONCLUSION

In this study 3 analytical methods were compared across 4 classification algorithms. Validation tests were performed to evaluate the classifier's performance using 10-fold cross validation for method I and method II. Percentage split method was used to validate the classifier developed using method III. The methods were evaluated based on performance parameters like area under ROC, accuracy and statistically important attributes like specificity and sensitivity.

The initial method I called naive SNP selection was explored and the results obtained were unsatisfactory. Secondly, we could see marginal improvement in the accuracy by carrying out binary classification using just feature selection technique. We observed improvement in performance using domain knowledge of 11 SNPs in the prediction model. We could see around 6-8% increase in the accuracy and marginal improvement in the area under ROC values using domain knowledge. These experiments were conducted across 10 iterations and the observed deviation in the accuracy was around ±2.5%. Interestingly, in addition to improved accuracy, high sensitivity and lower specificity values were observed in the model developed using domain knowledge. For example, when J48 Decision Tree was used, the model developed using domain knowledge had improved accuracy (60.56%), 0.798 sensitivity, and 0.301 specificity, which can be useful for initial screening. Hence, we can conclude that the model generated using domain information of SNPs can be helpful for assessing the risk of breast cancer in European women.

VII. ACKNOWLEDGMENT

The authors of this paper would like to thank The National Institute of Standards and Technology (NIST) for funding the project. We would also like to thank Database of Genotypes and Phenotypes (dbGaP) for providing access to the breast cancer data for this research.

VIII. REFERENCES

[1] Suzanne W Fletcher, Daniel F Hayes, Don S Dizon Patient information: Risk factors for breast cancer (Beyond the Basics), http://www.uptodate.com/contents/patient-information-risk-factors-for-breast-cancer-beyond-the-basics.

[2] Kong, W. and Choo, K. W.,"Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support

vector machine,” Frontiers Biosci., v12, pp. 1610-1614, 2007

[3] VU Onay, L Briollais, JA Knight, E Shi, Y Wang, S Wells, H Li, I Rajendram, IL Andrulis, H Ozcelik, SNP-SNP interactions in breast cancer susceptibility, BMC Cancer 6 (2006), p. 114, http://www.ncbi.nlm.nih.gov/pubmed/16672066

[4] dbGaP,The database of Genotypes and Phenotypes (dbGaP) - http://www.ncbi.nlm.nih.gov/gap

[5] L J Engle1, C L Simpson, and J E Landers, Using high-throughput SNP technologies to study cancer, Oncogene (2006) 25, 1594–1601. doi:10.1038/sj.onc.1209368.

[6] SNPedia - Information http://www.snpedia.com

[7] Khoury M. J., Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective, Office of Genetics and Disease Prevention, Centers for Disease Control and Prevention, Epidemiology, 9: 350-354, 1998 - http://www.ncbi.nlm.nih.gov/pubmed/9583430.

[8] McCarthy, M. I. (2011), Dorothy Hodgkin Lecture 2010. From hype to hope? A journey through the genetics of Type 2 diabetes. Diabetic Medicine, 28: 132140. doi: 10.1111/j.1464-5491.2010.03194.x

[9]SNPs Associated with Breast Cancer - http://www.snpedia.com/index.php/Breast cancer

[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2 ed. San Francisco: Morgan Kaufmann, 2005.

[11] Ting Yu, Simeon Simoff, and Tony Jan. 2010. VQSVM: A case study for incorporating prior domain knowledge into inductive machine learning. Neurocomput. 73, 13-15 (August 2010), 2614-2623. DOI=10.1016/j.neucom.2010.05.007 http://dx.doi.org/10.1016/j.neucom.2010.05.007

[12] Niyogi, P.; Girosi, F.; Poggio, T ,”Incorporating prior information in machine learning by creating virtual examples,” Proceedings of the IEEE , vol.86, no.11, pp.2196-2209, Nov 1998 doi: 10.1109/5.726787

[13] Qiang Sun and Gerald DeJong. 2005. Explanation-Augmented SVM: an approach to incorporating domain knowledge into SVM learning. In Proceedings of the 22nd international conference on Machine learning (ICML’ 05). ACM, New York, NY, USA, 864-871. DOI=10.1145/1102351.1102460 http://doi.acm.org/10.1145/1102351.1102460.

[14] WEKA The University of waikato - http://www.cs.waikato.ac.nz/ml/weka/

[15] Matlab - http://www.mathworks.com/products/matlab/

[16] Shweta Kharya: Using data mining techniques for diagnosis and prognosis of cancer disease, CoRR abs/1205.1923: (2012)

[17] Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin, et al. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clin Cancer Res 2004; 10:2725-2737. Published online April 20, 2004.

[18] Cortes C, Vapnik V. Support-vector networks. Machine Learn-ing 1995;20(3):273297.

[19] H.-J. Ban, J. Y. Heo, K.-S. Oh, and K.-J. Park, "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine," BMC Genetics, vol. 11, p. 26, 2010.

[20] LIBSVM - A Library for Support Vector Machines, URL - http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[21] Wray NR, Yang J, Goddard ME, Visscher PM (2010) The Genetic Interpretation of Area under the ROC Curve in Genomic Profiling. PLoS Genet 6(2): e1000864. doi:10.1371/journal.pgen.1000864

[22] G. Ustunkar, S. Ozogur-Akyuz, G. Weber, and Y. A. Son, Analysis of SNP-Complex Disease Association by a Novel Feature Selection Method, in Operations Research Proceedings 2010, Springer Berlin Heidelberg, 2011, pp. 21-26.

[23] T. Poggio and T. Vetter, Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries, Artificial Intell. Lab., MIT, Cambridge, MA, A.I.Memo no. 1347, 1992.

[24] Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. Radiology 2006;240:666-673.