Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer

Aniket Bochare, Aryya Gangopadhyay, Yelena Yesha, Anupam Joshi and Yaacov Yesha*

University of Maryland Baltimore County,
1000 Hilltop Circle,
Baltimore, Maryland 21250, USA
E-mail: aniketb1@.umbc.edu
E-mail: gangopad@.umbc.edu
E-mail: yeyesha@.umbc.edu
E-mail: joshi@.umbc.edu
E-mail: yayesha@cs.umbc.edu
E-mail: yayesha1@gmail.com
*Corresponding author

Mary Brady

National Institute of Standards and Technology,
100 Bureau Drive,
Gaithersburg, Maryland 20899, USA
E-mail: mary.brady@nist.gov

Michael A. Grasso

University of Maryland, School of Medicine,
655 West Baltimore Street,
Baltimore, Maryland 21201,USA
E-mail: mgrasso@umem.org

Napthali Rishe

Florida International University,
11200 SW 8th St,
Miami, Florida 33174, USA
E-mail: rishen@cis.fiu.edu

Abstract: 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 nine best SNPs using various feature selection algorithms and evaluate binary classifiers performance. 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 nine best SNPs, feature selection resulted in removing age from the set of features to be used for cancer risk assessment.

Keywords: breast cancer; classification; single nucleotide polymorphism; SNP; genome-clinical; domain knowledge; medical informatics; feature selection.

Reference to this paper should be made as follows: Bochare, A., Gangopadhyay, A., Yesha, Y., Joshi, A., Yesha, Y., Brady, M., Grasso, M.A. and Rishe, N. (2014) 'Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer', *Int. J. Medical Engineering and Informatics*, Vol. 6, No. 2, pp.87–99.

Biographical notes: Aniket Bochare graduated from the University of Maryland, Baltimore County in 2012 and is currently working as a Software Engineer at Cerner Corporation. He completed his Master in Computer Science. His research focused on using genetic data to find disease outcome using supervised machine learning algorithms. He was awarded a research fellowship from IBM Canada in 2011 to conduct this research under the supervision of his advisor, Dr. Yelena Yesha.

Aryya Gangopadhyay is a Professor and the Chair of Information Systems at the University of Maryland Baltimore County (UMBC). His research interests are in the areas of databases and data mining. Currently, he is focused on privacy preserving data mining, spatio-temporal data mining, and data analytics for health informatics. His research has been funded by grants from NSF, NIST, US Department of Education, Maryland Department of Transportation, and other agencies. He has published five books and nearly 100 research articles. He holds a PhD in Computer Information Systems from Rutgers University.

Yelena Yesha is a tenured Professor at the Computer Science and Electrical Engineering Department of the University of Maryland, Baltimore County (UMBC). She received her BSc in Computer Science and in Applied Mathematics from York University, Toronto, Canada, in 1984, and her MSc and PhD from The Ohio State University in 1986 and 1989, respectively. She has published ten books as author or editor, and more than 120 papers in prestigious refereed journals and refereed conference proceedings, and has been awarded external funding in a total amount exceeding 21 million dollars.

Anupam Joshi is the Oros Family Professor of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County (UMBC), the Director of the UMBC Center for Cybersecurity, and in charge of the UMBC Cyberscholars program. He obtained his BTech in Electrical Engineering from IIT Delhi in 1989, and Masters and PhD in Computer Science from Purdue University in 1991 and 1993 respectively. He has published over 175 technical papers, and filed and been granted several patents. His research has been supported by US Government agencies and by industry.

Yaacov Yesha is a Professor at the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. He received his PhD in Computer Science in 1979 from the Weizmann Institute of Science. He has received substantial external research funding from US Government agencies and industry. He was a program committee member of several conferences, a program Vice Chair for the Seventh International Conference on Parallel and Distributed Computing Systems, 1994, and a Chair of two workshops at IBM CASCON 2007.

Mary Brady is the Group Manager at the Information Technology Laboratory of the US Department of Commerce National Institute of Standards and Technology. She received her MSc in Computer Science from George Washington University in 1990, and a BSc in Computer Science and Mathematics from Mary Washington College in 1985.

Michael A. Grasso is an Assistant Professor of Emergency Medicine, Internal Medicine, and Computer Science at the University of Maryland School of Medicine. He earned his Medical degree from George Washington University and his PhD in Computer Science from the University of Maryland Baltimore County. He is the Director of the University of Maryland Clinical Informatics Group. He has over 20 years of experience in clinical informatics with an emphasis on clinical decision support, big data, emergency preparedness, personalised medicine, and patient safety.

Naphtali Rishe is an author of three books and 300 papers in journals and proceedings on databases, software engineering, geographic information systems, internet and life sciences. He is an awardee of over $55 million in research grants by government and industry, including NASA, NSF, IBM, DoI, USGS, and DoT. He is the Director of the NSF International FIU-FAU-Dubna Industry-University Cooperative Research Center for Advanced Knowledge Enablement (I/UCRC). He is the inaugural FIU Outstanding University Professor. He is the Principal of the TerraFly project, which has been extensively covered by worldwide press, including the New York Times, USA Today, NPR, Science and Nature journals, and FOX TV News.

1 Introduction

According to Fletcher et al. (2012), each year in USA 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 (Fletcher et al., 2012).

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 single nucleotide polymorphisms (SNPs) in early disease prediction. This will assist doctors in assessing the likelihood of developing breast cancer and in deciding whether to 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 three 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.

1.1 SNPs and personalised medicine

According to Kong and Choo, a 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 personalised drugs to individuals depending on their genetic makeup (Kong and Choo, 2007).

Engle et al. (2006) address the contribution of SNPs to cancer development. Onay et al. (2006) highlight the fact that SNPs belonging to certain genes increases the susceptibility to breast cancer. 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 (http://www.snpedia.com) and use domain knowledge of SNPs to come up with an improved prediction model.

Khoury and Yang (1998) have shown that in complex diseases the disease susceptibility may vary with gene-environmental interactions and genes originating from diverse demography. 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 (2011) focuses on the importance of combinations of SNPs instead of a single SNP in the development of type II diabetes. Due to cumulative effect of SNPs, we selected a set of risk associated SNPs for determining genomic risk of an individual. We used three 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 nine 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.

2 Background and related work

2.1 Genetics and breast cancer

SNPedia provides references to different studies conducted by researchers on breast cancer patients and SNPs associated with breast cancer (http://www.snpedia.com/index.php/Breast cancer). Therefore, we hope that by delving into the genomic patterns of a population one can detect risks at an early stage and assist physicians.

2.2 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-neighbour (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 (Witten and Frank, 2005).

2.3 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 sanitisation 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 (Yu et al., 2010). Niyogi et al. (1998) discusses incorporating domain knowledge by using virtual examples (Poggio and Vetter, 1992) in the learning task.

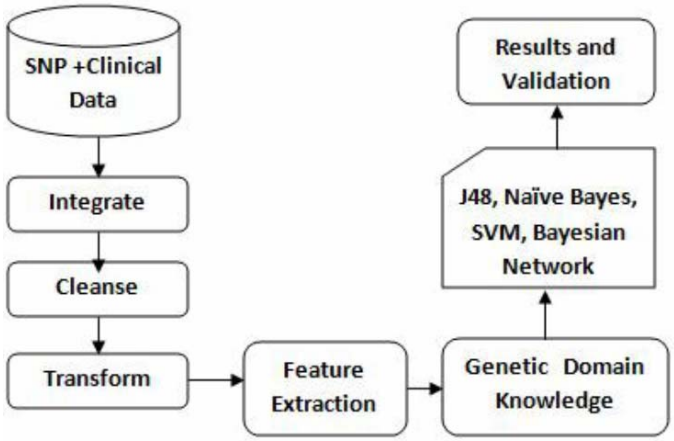
We know that traditional machine learning algorithms do not take into consideration the knowledge about data for training classifiers. Sun, et al. point out that combining prior domain knowledge with training set aids in machine learning. They have demonstrated a novel approach of combining domain knowledge into SVM for better efficiency (Sun and DeJong, 2005). 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 (Yu et al., 2010). 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.

3 Clinical-genomic data

3.1 Data description

In our experiments we have used the 'Nurses Health Study (NHS)-GCEMS Stage-1' breast cancer data from dbGaP (http://www.ncbi.nlm.nih. gov/gap). The dataset contains post menopausal women of European ancestry out of which 1,145 are cases and 1,142 controls.

Figure 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 (http://www.cs.waikato.ac.nz/ml/weka/) and MATLAB http://www.mathworks.com/products/matlab/ for our experiments. Figure 1 shows the framework of our project.

4 Classification and validation

4.1 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. Kharya (2012) 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 *NB:* NB 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:

$$P(class = Y \mid data = X) = P(data = X \mid class = Y) * P(class = Y) / P(data = X) \quad (1)$$

NB has been used before in predicting the risk of breast cancer susceptibility from multiple SNPs. Listgarten et al. (2004) demonstrate 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 (Burnside et al., 2006).

4 *SVM:* Cortes and Vapnik (1995) 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. 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 (Ban et al., 2010). We used an open source SVM library called LibSVM (http://www.csie.ntu.edu.tw/cjlin/libsvm/) 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.

4.2 Validation and accuracy

1 *Ten-fold cross validation:* The testing and validation was carried out using ten-fold cross validation (Witten and Frank, 2005).

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 (Wray et al., 2010). It is a measure used in machine learning to predict binary classifier's performance.

4.3 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.

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

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

5 Methods and procedures

5.1 Classification using naive SNP selection from literature

Table 1 shows 17 SNPs considered for our experiments which were obtained from SNPedia. Table 2 shows the results obtained after performing classification using 17 SNPs as features. This method is called naive approach since the SNPs were neither prioritised nor assigned any weights. We have used 10-fold cross validation to test our model.

5.2 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. Ustunkar et al. (2011) highlight the importance of selecting a subset of the available SNPs for conducting association studies. 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.

Table 1 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
TCF712	rs12255372	T
LSP1	rs3817198	T
CCNE1	rs997669	A
CDKNB1	rs34330	T
2q35	rs13387042	A

Table 2 Classification results using naïve SNP selection approach

Classification algorithm	*Accuracy*	*ROC*
J48-decision tree	52.30%	0.538
NB	55.05%	0.557
LibSVM (radial basis)	53.41%	0.530
LibSVM (polynomial)	52.95%	0.530
Bayesian network	54.27%	0.566

We used three techniques named filtered attribute evaluation, gain ratio attribute evaluation and information gain attribute evaluation available in WEKA to extract nine most informative SNPs from the dataset for binary classification. Table 3 shows the nine SNPs obtained using three feature selection techniques and Table 4 shows the results of binary classification using these nine SNPs. We have used ten-fold cross validation in method I and method II.

Table 3 Attributes ranked based on information gain, gain ratio and filtered attribute evaluation technique

Information gain	*Gain ratio*	*Filtered attribute evaluation*
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 3 results also indicate that rs2420946, rs1219648 and rs2981579 are the top 3 SNPs which appear in all the three 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 three feature selection results.

Table 4 Classification results using feature selection technique

Classification algorithm	*Accuracy*	*ROC*
J48-decision tree	54.92%	0.559
NB	56.39%	0.571
LibSVM (radial basis)	56.94%	0.573
LibSVM (polynomial)	54.57%	0.562
Bayesian network	55.83%	0.571

5.3 Classification using domain knowledge addition

We designed the following algorithm for adding domain knowledge:

Algorithm 1: Add Virtual Instances to the Original Dataset.

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

1: Assign '0' – norisk, '1' – mediumrisk, '2' – highrisk SNP.

2: 'SNP'_X = (0; 1; 2) {depends on number of risk alleles}.

3: **for** SNP_1 to SNP_n **do**

4: Let $C1_X$ & $C2_X \leftarrow$ row-count and $R1_X$ & $R2_X \leftarrow$ risk-value when $SNP_X = 1$ & $SNP_X = 2$ respectively.

5: Add total of $C1_X * (1 - R1_X)$ & $C2_X * (1 - R2_X)$ virtual instances where $SNP_X = 1$ & $SNP_X = 2$ respectively.

6: Add a random row vector V_i as follows:

< $SNP_1 \ldots SNP_n$, 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.

$$X_1 + R1_X * X_1 = C1_X \tag{4}$$

$$X_2 + R2_X * X_2 = C2_X \tag{5}$$

Solve for X_1 and X2. 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.

7: **end for**

We repeat this procedure for all the 11 SNPs selected from SNPedia which have risk associated values. Eight 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 trained 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 ten trials are shown in Table 5. By comparing Tables 2 and 5 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 5 is around ±2.5% across ten trials.

Table 5 Classification results using both domain knowledge and feature extraction

Classification algorithm	*Accuracy*	*ROC*
J48-decision tree	60.56%	0.591
NB	60.12%	0.574
LibSVM (radial basis)	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 is not 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 five algorithms used to evaluate the performance of binary classifier. Table 6 shows the values of sensitivity (SN) and specificity (SP) for all the three methods.

Table 6 Sensitivity and specificity results across three methods using ten fold cross validation

Sensitivity and specificity comparison						
Classification algorithm	*Method I*		*Method II*		*Method III*	
	SP	*SN*	*SP*	*SN*	*SP*	*SN*
J48-decision tree	0.566	0.536	0.579	0.531	0.301	0.798
NB	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 trade-off between sensitivity and specificity in case of J48, NB and Bayesian network classifiers. This fact, and the fact that using domain knowledge resulted in improved accuracy, are both important.

6 Conclusions

In this study, three analytical methods were compared across four classification algorithms. Validation tests were performed to evaluate the classifier's performance using ten-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 ten 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.

Acknowledgements

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. This material is based in part upon work supported by the National Science Foundation under Grant Nos. CNS-0821345, CNS-1126619, HRD-0833093, IIP-1338922, IIP-0829576, CNS-1057661, IIS-1052625, CNS-0959985, OISE-1157372, IIP-1237818, IIP-1330943, IIP-1230661, IIP-1026265, IIP-1058606, IIS-1213026.

References

Ban, H-J., Heo, J.Y., Oh, K-S. and Park, K-J. (2010) 'Identification of type 2 diabetes-associated combination of SNPs using support vector machine', *BMC Genetics*, Vol. 11, No. 26, doi:10.1186/1471-2156-11-26 [online] http://www.biomedcentral.com/1471-2156/11/26 (accessed 8 October 2013).

Burnside, E.S., Rubin, D.L., Fine, J.P., Shachter, R.D., Sisney, G.A. and Leung, W.K. (2006) 'Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience', *Radiology*, Vol. 240, No. 3, pp.666–673.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, No. 3, pp.273–297.

dbGaP, *The Database of Genotypes and Phenotypes (dbGaP)* [online] http://www.ncbi.nlm.nih.gov/gap (accessed 8 October 2013).

Engle, L.J., Simpson, C.L. and Landers, J.E. (2006) 'Using high-throughput SNP technologies to study cancer', *Oncogene*, Vol. 25, No. 11, pp.1594–1601, doi:10.1038/sj.onc.1209368.

Fletcher, S.W., Hayes, D.F. and Dizon, D.S. (2012) *Patient Information: Risk Factors for Breast Cancer (Beyond the Basics)* [online] https://www.fshealth.com/info/21198/risk-factors-for-breast-cancer/ (accessed 7 October 2013).

Kharya, S. (2012) *Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease*, CoRR abs/1205.1923.

Khoury, M.J. and Yang, Q. (1998) 'The future of genetic studies of complex human diseases: an epidemiologic perspective', *Epidemiology*, Vol. 9, pp.350–354, Office of Genetics and Disease Prevention, Centers for Disease Control and Prevention [online] http://www.ncbi.nlm.nih.gov/pubmed/9583430 (accessed 8 October 2013).

Kong, W. and Choo, K.W. (2007) 'Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support vector machine', *Frontiers Biosci.*, Vol. 12, No. 5, pp.1610–1614.

LIBSVM – A Library for Support Vector Machines [online] http://www.csie.ntu.edu.tw/cjlin/libsvm/.

Listgarten, J., Damaraju, S., Poulin, B. et al. (2004) 'Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms', *Clin. Cancer Res*., Vol. 10, pp.2725–2737 [online] http://clincancerres.aacrjournals.org/content/10/8/2725.long (accessed 13 October 2013), published online 20 April 2004.

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

McCarthy, M.I. (2011) 'Dorothy Hodgkin Lecture 2010. From hype to hope? A journey through the genetics of type 2 diabetes', *Diabetic Medicine*, Vol. 28, No. 2, pp.132–140, doi: 10.1111/j.1464-5491.2010.03194.x.

Niyogi, P., Girosi, F. and Poggio, T. (1998) 'Incorporating prior information in machine learning by creating virtual examples', *Proceedings of the IEEE*, Vol. 86, No. 11, pp.2196–2209, November, DOI: 10.1109/5.726787.

Onay, V.U., Briollais, L., Knight, J.A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I.L. and Ozcelik, H. (2006) 'SNP-SNP interactions in breast cancer susceptibility', *BMC Cancer*, Vol. 6, p.114 [online] http://www.biomedcentral.com/1471-2407/6/114 (accessed 13 October 2013).

Poggio, T. and Vetter, T. (1992) *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.

SNPedia – Information [online] http://www.snpedia.com (accessed 13 October 2013).

SNPs Associated with Breast Cancer [online] http://www.snpedia.com/index.php/Breast cancer (accessed 13 October 2013).

Sun, Q. and DeJong, G. (2005) 'Explanation-augmented SVM: an approach to incorporating domain knowledge into SVM learning', *Proceedings of the 22nd International Conference on machine learning (ICML'05)*, ACM, New York, NY, USA, pp.864–871, http://doi.acm.org/10.1145/1102351.1102460.

Ustunkar, G., Ozogur-Akyuz, S., Weber, G. and Son, Y.A. (2011) 'Analysis of SNP-complex disease association by a novel feature selection method', *Operations Research Proceedings 2010*, Springer, Berlin, Heidelberg, pp.21–26.

WEKA, The University of Waikato [online] http://www.cs.waikato.ac.nz/ml/weka/.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco.

Wray, N.R., Yang, J., Goddard, M.E. and Visscher, P.M. (2010) 'The genetic interpretation of area under the ROC curve in genomic profiling', *PLoS Genet*., Vol. 6, No. 2, p.e1000864, doi:10.1371/journal.pgen.1000864.

Yu, T., Simoff, S. and Jan, T. (2010) 'VQSVM: a case study for incorporating prior domain knowledge into inductive machine learning', *Neurocomput*, August, Vol. 73, Nos. 13–15, pp.2614–2623, http://dx.doi.org/10.1016/j.neucom.2010.05.007.