Breast Cancer Prediction Using Genome Wide Single Nucleotide Polymorphism Data

Mohsen Hajiloo[1,2], Babak Damavandi[1,2], Metanat Hooshsadat[1,2], Farzad Sangi[1,2], John R. Mackey[3], Carol E. Cass[3], Russell Greiner[1,2*], and Sambasivarao Damaraju[4,5*]
[1]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[2]Alberta Innovates Centre for Machine Learning, University of Alberta, Edmonton, Alberta, Canada
[3]Department of Oncology, University of Alberta, Edmonton, Canada
[4]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, Canada
[5]PolyomX Program, Cross Cancer Institute, Alberta Health Services, Edmonton, Alberta, Canada
Email: {hajiloo@ualberta.ca, damavand@ualberta.ca, hooshsad@ualberta.ca, fsangi@ualberta.ca,
john.mackey@albertahealthservices.ca, carol.cass@albertahealthservices.ca, rgreiner@ualberta.ca,
sambasivarao.damaraju@albertahealthservices.ca}
[*] These authors contributed equally to this work and are joint corresponding authors

***Abstract*—This paper introduces and applies a Genome Wide *Predictive* Study (GWPS) to learn a model that predicts whether a new subject will develop breast cancer or not, based on her SNP profile. We applied a combination of a feature selection method (MeanDiff) and a learning method (K-Nearest Neighbours, KNN) to a dataset of 623 female subjects, including 302 cases of breast cancer and 321 apparently healthy controls from Alberta, Canada. The learning algorithm considered all the SNPs (506,836) from a whole genome scan with 100% call rate and with minor allele frequency of > 5%. The leave-one-out cross-validation (LOOCV) accuracy of this classifier is 59.55%. Random permutation test show that this result is significantly better than the baseline accuracy of 51.52%. Sensitivity analysis shows that our model is robust to the number of selected SNPs. We then used the only relevant publicly available breast cancer dataset (CGEMS breast cancer dataset with 1145 breast cancer cases and 1142 controls) to further validate our approach. We showed that our combination of MeanDiff and KNN leads to a LOOCV accuracy of 60.25%, which is significantly better than the CGEMS baseline of 50.06%. To better understand the challenge of this dataset, we systematically explored a large variety of other feature selection and learning algorithms. We found that none of the biologically naïve approaches to feature selection worked as well as our MeanDiff. We also considered many biologically-informed methods to select SNPs – using SNPs reported in the literature to be associated with breast cancer, SNPs associated with genes of KEGG's cancer pathways, and SNPs associated with breast cancer in the F-SNP database. However, those SNPs produced classifiers that were not even better than baseline. These negative findings suggest the challenge of our task. Finally, we identified several limitations that may hinder a more accurate predictive model for breast cancer susceptibility: Our study implicitly assumes that breast cancer is a homogenous phenotype, but it is not. Moreover, while our study does involve 623 samples, this is small relative to the number of features (SNPs) from a whole genome scan; we expect to achieve yet better results given a larger sample sizes. Furthermore, we anticipate developing better predictive models by incorporating other genetic information (such as point mutations and copy number variations) as well as environmental and lifestyle factors.**

I. Introduction

A genome wide association study (GWAS) compares the SNP profiles, over a wide range of SNPs, of two groups of participants: e.g., people with the disease (cases) versus people without the disease (controls). Each individual SNP whose values are significantly different between these groups (typically based on chi-square test between the values observed for the two groups) is said to be associated with the disease [1]. The database of Genotypes and Phenotypes (dbGaP) archives and distributes the results of studies that have investigated the interaction of a genotype and phenotype in GWASs [2]. While GWASs can help the researchers better understand diseases, genes and pathways, they are not designed to predict whether a currently undiagnosed subject is likely to develop the disease. This paper introduces Genome Wide *Predictive* Studies (GWPSs), which take the same input as a GWAS (a set of SNP arrays for individuals, each labelled as a case or a control) but outputs a classification model that can be used later to predict the class label of a previously undiagnosed person, based on his/her SNP profile. Here, we consider a way to learn a predictor ("who has breast cancer?"), for a dataset that specifies all available SNPs about each subject.

Our approach differs from research that attempt to learn predictors from only a pre-defined set of candidate SNPs. As an example of such a candidate SNP study, Listgarten et al. [3] applied a machine learning tool (support vector machine, SVM) to a pre-defined set of 98 SNPs, distributed over 45 genes of potential relevance to breast cancer, to develop a predictive model with 63% accuracy for predicting breast cancer. Ban et al. [4] applied a SVM to analyze 408 SNPs in 87 genes involved in type 2 diabetes (T2D) related pathways, and achieved 65% accuracy in T2D disease prediction. Wei et al. [5] studied type 1 diabetes (T1D) using genome wide scan of SNPs and reported 84% area under curve (AUC) using an SVM.

Our approach also differs from the conventional risk modeling/prediction studies. Those studies also begin with a small set of pre-defined features: they first sort the training subjects into a small set of bins, based on the values of these features – e.g., the Gail model uses 7 features to produce a small number of bins – and record the percentage in each bin with the phenotype (here breast cancer) [6-7]. Afterwards, to estimate the risk a new subject will face, this tool uses the subject's values for those relevant features to sort that subject into the proper bin, and returns the associated probability (called risk). Hence this approach bases its assessment on only a small number of pre-specified features. Note this might not be sufficient to usefully characterize the subjects, especially if the hand-picked features are not adequate. On the other hand, our machine learning (ML) approach lets the data dictate on the possible combination of features that are relevant. (While the ML model described in this paper returns a specific prediction for the individual – here breast cancer or not – there are other ML models that will return the probability that the individual will have the disease P(disease | feature_values), which is basically risk).

Our general goal is to develop a tool to help screen women, by predicting which of the apparently healthy subjects sampled in a population will eventually develop breast cancer. This cannot be done by gene expression-based microarray analyses, as those results require biopsies of tissues from organs or tumours, which means they are only relevant to individuals with suspect tissues; hence they are not effective at identifying individuals at risk in a general population, before the onset of the disease, and so cannot be used for our early detection.

II. Method

In general, a Genome Wide *Predictive* Study (GWPS) takes as input the SNP profiles of a set of N individuals (including both cases and controls) and outputs a classifier, which can later be used to predict the class label of a new individual, based on his/her SNP profile. Here, we used a dataset of N=623 subjects including 302 cases (with breast cancer) and 321 controls (disease free at the time of recruitment), accessed from a previous study on sporadic breast cancer wherein breast cancer predisposition in women is not related to mutations in the known high penetrance breast cancer genes (eg, BRCA) nor other genes of moderate penetrance, described in earlier studies. Briefly, our study subjects began with 348 cases (late onset of disease, i.e., of sporadic nature) and 348 controls (with no family history of breast cancer), both predominantly of Caucasian origin. Population stratification correction using EIGENSTRAT technique removed 73 subjects (46 cases and 27 controls) that did not co-cluster with Hapmap II Caucasian subjects. We then isolated germline DNA from peripheral blood lymphocytes and generated genotyping profiles using Affymetrix Human SNP 6.0 array platform (906,600 SNPs on each array). The study subjects provided informed consent and the study was approved by the Alberta Cancer Research Ethics Committee of the Alberta Health Services [8]. Following probe labelling, hybridization and scanning, the data was filtered by removing any SNP (1) that had any missing calls, (2) whose genotype frequency deviated from Hardy-Weinberg equilibrium (nominal p-value <0.001 in controls) or (3) whose minor allele frequency were less than 5%; this left a total number of 506,836 SNPs for analysis. For each SNP, we represented wild type homozygous, heterozygous and variant homozygous by 1, 2, and 3 respectively. Collectively, we view this as a labelled dataset, where the label for each subject is either case or control.

A trivial classifier, which just predicts the majority class (here control), will be 321/623 = 51.52% accurate. The challenge is producing a classifier that uses subject SNP data to produce predictions that are significantly more accurate. In particular, we explored tools that use the given labelled dataset to find the patterns that identify breast cancer (i.e., case versus control). Fortunately, the field of machine learning (ML) provides many such learning algorithms, each of which takes as input a labelled dataset, and returns a classifier. These systems typically work best when there are a relatively small number of features – typically dozens to hundreds – but they tend to work poorly in our situation, with over half-a-million features; here, they will invariably over-fit: that is, do very well on the training data as they find ways to fit the details of this sample, but in a way that does not translate to working well on the subjects that were not part of the training dataset. Note that our goal is to correctly classify such novel (that is, currently-undiagnosed) subjects. We therefore apply a pre-processing step to first reduce the dimensionality of the data, by autonomously identifying a subset of the most relevant SNPs (features). We then give this reduced dataset to a learning algorithm, which produces a classifier. We also discuss how to evaluate the classifier produced by this "feature-selection + learning" system.

A. *Feature Selection*

In our analysis, as we expect only a subset of the SNPs to be relevant to our prediction task, we focused on ways to select such a small subset of the features. In general, this involves identifying the features that have the highest score based on some criteria (which we hope corresponds to being most relevant to the classification task). In this study, we used the MeanDiff feature selection method, which first sorts the SNPs based on their respective MeanDiff values, which is the absolute value of the difference between mean values of this SNP over the cases and the controls:

$$\mathrm{MeanDiff}(\mathrm{SNP_i}, \mathrm{D}) = |\mu(\mathrm{i}, \mathrm{C}) - \mu(\mathrm{i}, \mathrm{H})| \qquad (1)$$

over the dataset D = C ∪ H where C is the set of subjects known to have cancer (each labelled as case) and H is the remaining healthy subjects (each labelled as control), and using Expr(i,j) as the value of the i'th SNP of subject j, $\mu(i,H) = \frac{1}{|H|}\sum_{j\in H} Expr(i,j)$ is the mean value of the i'th SNP over the subset H (the controls) and $\mu(i,C) = \frac{1}{|C|}\sum_{j\in C} Expr(i,j)$ is the mean value of the i'th SNP over the subset C (the cases). Note this MeanDiff($\mathrm{SNP_i}$, D) score will be 0 when $\mathrm{SNP_i}$ is irrelevant and presumably larger for SNPs that are more relevant to our prediction task.

B. *Learning*

To build a classifier, we use the very simple learning algorithm, K-Nearest Neighbors (KNN), which simply stores the (reduced) profiles for all of the training data [9]. To classify a new subject *p*, this classifier determines *p*'s k nearest neighbors, then assigns p the majority vote. (So if k=5, and *p*'s

5 closest neighbors include 4 controls and 1 case, then this classifier assigns *p* as control). Of course, we need to define distances to determine the nearest neighbors. As we are representing each patient as a m-tuple of the SNP values, we define the distance between two individuals $p = [p_1, ..., p_m]$ and $q = [q_1, ..., q_m]$ as the square of the Euclidean distance (aka L2 distance) as shown below.

$$\mathrm{d}(\mathrm{p},\mathrm{q}) = \sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathrm{p}_\mathrm{i} - \mathrm{q}_\mathrm{i})^2 \qquad (2)$$

C. *Evaluation*

Here we use two strategies to evaluate our classification algorithm: (1) by using Leave-One-Out Cross Validation (LOOCV) strategy and (2) by using an external hold-out (validation) dataset from the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer project [10].

III. RESULTS

Our LOOCV estimates the accuracy of this model to be 59.55%; with precision 50.40%, recall/sensitivity 61.92%, and specificity 57.32%. To test if this result is significantly more accurate than the baseline of 51.52%, we applied a permutation test [11]. Here, we permuted the labels in the original dataset randomly, which should destroy any signal relating the SNPs to the cancer/no-cancer phenotype. We then ran the KNN to build new classifiers on this new dataset, and ran the LOOCV process to estimate the accuracy of the new model. We repeated this "permute, learn, evaluate" process over 100 permutations. None of these accuracies (of the 100 models built over randomly permuted labelled datasets) exceeded the 59.55% accuracy of our model. This suggests that our result is significantly better than the baseline, with a confidence of more than 1 - 1/100 = 0.99 -- ie, the associated p-value is $p<0.01$. Furthermore, we measured the LOOCV accuracy of the classification model built using KNN on sets of SNPs with the top {500, 600, ..., 1500} MeanDiff scores and we realized that our model is fairly robust to the number of MeanDiff selected SNPs, when selecting more than 500 SNPs.

To test the effectiveness of our approach, we next explored ways to apply it to other datasets. Unfortunately, there are no other public datasets for this phenotype that use the same Affy 6.0 Platform. We did, however, consider applying our C_{623} = KNN(D_{623}) classifier on the CGEMS breast cancer dataset with 1145 breast cancer cases and 1142 controls genotyped on the Illumina I5 array platform. This dataset includes only 101 SNPs in common with the m=500 SNPs used by C_{623}. As this meant the CGEMS data was missing ~80% of the SNP values used by C_{623}, we obviously could not apply C_{623} directly on this dataset. As this CGEMS breast cancer dataset is the only available genome wide association study dataset on Caucasian population, we therefore had to design another experiment to use the external hold-out set to evaluate our approach, of the KNN learning method that involved the MeanDiff feature selection method. Here, we applied the same algorithm explained in the Methods section, KNN(.), but trained this method over D_{2287}, the 2287 subjects of CGEMS breast cancer dataset. We evaluated the performance of this model using the LOOCV method. LOOCV accuracy is 60.25% (which is significantly better than the baseline of 50.06%), with precision 59.39%, recall/sensitivity 59.65%, and specificity 59.11%.

Table 1 – 10-Fold Cross Validation Accuracy of Various Combinations of Statistical Feature Selection and Learning Methods

		Feature Selection Methods			
		Information Gain	**MeanDiff**	**mRMR**	**PCA**
Learning Methods	**Decision Tree**	50.88%	52.06%	51.20%	51.69%
	KNN	56.17%	58.71%	57.78%	51.36%
	SVM-RBF	55.37%	57.30%	56.18%	51.84%

This confirms that our approach and algorithm, is reproducible, as this exact system works effectively on a second, very different breast cancer dataset.

Hoping to further improve these results, we explored several techniques – both biologically naïve and informed – for both selecting features and for building the classifier itself. To select features, we considered biologically naïve methods such as information gain [12], minimum redundancy maximum relevance (mRMR) [13] and principal component analysis (PCA) [14]. We also applied other biologically naïve learning algorithms, including decision trees [12], and support vector machines (with RBF kernel) [15]. In all, we tried dozens of different combinations of the learning and feature selection algorithms (each with its own range of parameters values) – each of which proved to be computationally intensive (several CPU days). Table 1 shows the (10-fold cross validation) accuracy of 12 of these combinations.

We also used biological information related to cancer to inform feature selection – ie, use SNPs known to be relevant to breast cancer, rather than our biologically-naïve MeanDiff method: First, we analyzed 28 SNPs identified by recent GWASs as being highly associated with breast cancer. We trained a classifier over the 623 subjects, but using only these 28 SNPs; unfortunately the LOOCV of this classifier was just baseline. (We also noticed that none of these 28 SNPs appear in the list of 500 SNPs selected by our MeanDiff feature-selection algorithm.) Second, we tried using only the 12,858 SNPs associated with genes of KEGG's cancer pathways [16] recognized as hallmarks of cancer [17]; unfortunately, the classifier based on these features also did not perform better than baseline. Finally, we built a classifier using only the 1,661 SNPs associated with breast cancer in the F-SNP database [18]; this too had just baseline accuracy. These negative results show that the obvious approach of first using prior biological information to identify SNPs, and then learning a classifier using only those SNPs, does not work here. Recall that our feature selection method found the relevant SNPs itself; *n.b.*, it did not just use the SNPs considered significant by some earlier association test. (This demonstrates that the predictive power of our model does not depend on the SNPs that previous GWASs have reported to be statistically significant for breast cancer susceptibility.) Our feature selection method automatically deals with the redundancies of features – ie, SNPs that are highly correlated with one another. We are now exploring ways to use SNPs from common variants, anticipating that clinically useful models may emerge from integrating rarer variants and mutations in the genome as well as gene-environment interactions, using the machine learning approaches described.

IV. Discussions

Our studies, using MeanDiff within KNN, confirm that SNPs do carry information related to breast cancer genetic susceptibility, and that GWPSs are a promising tool for decoding and exploiting this information. While this approach is theoretically applicable for studying other cancer types and diseases, we list below some of the potential limitations that may make it difficult to produce more accurate breast cancer prediction models:

***Small Sample Size vs. Large Feature Size*:** As noted earlier, the number of subjects in this study is much less than the number of SNPs (a few hundred instances versus half a million features) can easily cause standard learning systems to over-fit – i.e., produce models that perform well on the training subjects but relatively poorly on new subjects distinct from the those training subjects.

Heterogeneity of Breast Cancer: Breast cancer is biologically heterogeneous [19]. Our current dataset ignores the differences between different subtypes by merging these them into a single label, case. We might be able to produce a more accurate predictor if we employed more detailed labelling of sub-cases, that sought a classifier that could map each subject to its specific molecular subtype.

SNPs are Only one Form of Genomic Alterations: While the heritable genetic basis for breast cancer occurs in SNPs, mutations, copy number variations (CNVs), and other chromosomal changes, this study considered only SNPs. We believe that augmenting the SNP data with such additional genetic information, could lead to more accurate breast cancer predictive models.

Breast Cancer is also influenced by Non-genetic Factors: Heritable factors are only part of the issue. Indeed, for many of diseases, the genetic component accounts for only 30-60% of the risk, with the remaining risk due to environmental and life style risk factors [20]. We anticipate a better predictive result from a comprehensive model that includes both genetic and non-genetic factors.

Acknowledgment

We thank Dr Badan Sehrawat and Mr. Yadav Sapkota for discussions and help with data access. We sincerely thank Dr Paula Robson from the Tomorrow Project of the Alberta Health Services, Alberta for discussions and GWAS studies. We acknowledge financial support for the interfaculty graduate studentship from the Faculty of Medicine and Dentistry and the Faculty of Science, University of Alberta (to MH); Natural Sciences and Engineering Research Council and Alberta Innovates Centre for Machine Learning (to RG); Alberta Cancer Foundation, Canadian Breast Cancer Foundation-Prairies/NWT region, Alberta Innovates-Health Solutions funding support for the Alberta Cancer Research Biorepository/CBCF-Tumor Bank (to SD and CEC); Canadian Breast Cancer Foundation-Prairies/NWT region for financial support through operating grants (to SD and JRM). We acknowledge dbGaP of NCBI for providing us access to the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer dataset.

References

[1] Manolio TA: **Genomewide association studies and assessment of the risk of disease**. *N Engl J Med* 2010, **363**:166–76.

[2] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39**(10)**:**1181-1186.

[3] Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B: **Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms.** *Clinical Cancer Research* 2004, **10:**2725-2737.

[4] Ban HJ, Heo JY, Oh KS, Park KJ: **Identification of Type 2 diabetes-associated combination of SNPs using Support Vector Machine**. *BMC Genet*ics 2010, 11:26.

[5] Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H: **From disease association to risk assessment: an optimistic view from genome-wide association studies on type-1 diabetes.** *PLoS Genetics* 2009, **5**(10)**:**e1000678.

[6] Bondy ML and Newman LA: **Assessing breast cancer risk: evolution of the Gail Model**, *Journal of National Cancer Institute* 2006, **98**(17): 1172-3.

[7] Decarli A, Calza S, Masala G, Specchia C, Palli D, Gail MH: **Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort**, *Journal of National Cancer Institute* 2006, **98**(23): 1686-1689.

[8] Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, Mackey J, Greiner R, Damaraju S: **Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility,** *Human Genetics* 2011, **130**(4):529-537.

[9] Cover TM, Hart PE: **Nearest neighbor pattern classification.** *IEEE Trans Inform Theory* 1967, **IT-13**:21–27.

[10] Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer,** *Nature Genetics* 2007, **39**(7):870-874.

[11] Good P: *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd edition. New York: Springer Series in Statistics; 2005.

[12] Quinlan AR: **Induction of decision trees.** *Machine Learning* 1986, **1**(1)**:** 81-106.

[13] Ding C, Peng H: **Minimum redundancy feature selection from microarray gene** expression **data.** *International Conference on Computational Systems Bioinformatics* 2003, 523-528.

[14] Jollife IT: *Principal Component Analysis*, Springer-Verlag, New York 1986.

[15] Vapnik V: *The Nature of Statistical Learning Theory,* Springer-Verlag, New York 1995.

[16] Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res.*2000, **28:** 27-30.

[17] Hanahan D, Weinberg RA: **The hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5)**:**646-667.

[18] Lee PH, Shatkay H: **F-SNP: computationally predicted functional SNPs for disease association studies.** *Nucleic Acids Res.* 2008, **36:** 820-824.

[19] Bertucci F, Birnbaum D: **Reasons for breast cancer heterogeneity**. *J Biol* 2008, **7**(2):6.

[20] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. *Nature* 2009, 461:747-753.