Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 34

11-1-2016

# Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach

Hansapani S. Rodrigo
*University of South Florida*, sarasepa@mail.usf.edu

Chris P. Tsokos
*University of South Florida*, ctsokos@usf.edu

Taysseer Sharaf
*University of Michigan-Dearborn*, tsharaf@umich.edu

# Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach

**Cover Page Footnote**
Data collection for this work was supported by a NCI-funded Breast Cancer Surveillance Consortium co-operative agreement (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040). The authors also wish to acknowledge the constructive suggestions of Dr. M.H. Gail, Senior Investigator at National Cancer Institute, for his constructive and helpful suggestions in improving the present study.

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 563-579.
doi: 10.22237/jmasm/1478003520


ISSN 1538 − 9472

# Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach

**Hansapani S. Rodrigo**
University of South Florida
Tampa, FL

**Chris P. Tsokos**
University of South Florida
Tampa, FL

**Taysseer Sharaf**
University of Michigan
Dearborn, MI

In the current study, we have exemplified the use of Bayesian neural networks for breast cancer classification using the evidence procedure. The optimal Bayesian network has 81% overall accuracy in correctly classifying the true status of breast cancer patients, 59% sensitivity in correctly detecting the malignancy and 83% specificity in correctly detecting the non-malignancy. The area under the receiver operating characteristic curve (0.7940) shows that this is a moderate classification model.

*Keywords:* Multi-Layer perceptron, classification, breast cancer, Bayesian

## Introduction

Early detection of breast cancer can reduce the deadly threat to life. Including the well-known "Gail model" (Gail et al., 1989), a number of other statistical models have been proposed to assess the risk of being diagnosed with breast cancer (Claus, Risch, & Thompson, 1993; Domchek et al., 2003; van Asperen et al., 2004). However, these models imposed some limitations in their use of risk prediction (Amir et al., 2003; Euhus, Leitch, Huth, & Peters, 2002).

The objective of the current study is to develop a better statistical model to correctly classify the malignant breast cancer patients with their demographic factors and previous mammogram results using a multi-layer perceptron (MLP), a type of feedforward neural network. Although there exist several other models based on neural networks with the same intention, few of them have make use of the evidence approach with automatic relevance determination (ARD) prior for

*Ms. Rodrigo is a graduate student in the Department of Mathematics and Statistics. Email her at: sarasepa@mail.usf.edu. Mr. Tsokos is a Professor in the Department of Mathematics and Statistics. Email him at: ctsokos@usf.edu. Sharaf is an Assistant Professor of Mathematics and Statistics. Email him at: taysseer.sharaf@sru.edu.*

network regularization. We have selected the optimal network based on the model evidence (or cost function) as oppose to the classical minimum square error.

In order to train MLPs, we have considered two different approaches. In the first approach, a MLP is trained in the standard setting without incorporating any prior probabilities in their weight structure, where the later approach is based on Bayesian evidence procedure and the posterior probabilities of malignancy (Hung, Shanker, & Hu, 2002) have been obtained. These probabilities have been used as an initial measure for risk of diagnosing with incident breast cancer.

The advantage of neural networks over the other models is that, it is a self-learning model which is free of statistical assumptions. This allows neural network process to be considered as a generalization of existing statistical methodologies.

MLPs are used in a wide variety of fields including pattern recognition, cognition and decision making (Ayer et al., 2010; Floyd, Lo, Yun, Sullivan, & Kornguth, 1994; Orr, 2001; Wu et al., 1993), where they learn by examples through training algorithms. Training can be supervised, where both inputs and their corresponding outputs are fed to the network, or can be unsupervised, where training data consist of only the inputs. During the training process, the weights and the biases of the network are continuously adjusted to minimize the error between the network's output and the target outputs (Haykin, 1999). This process leads weights and biases of the network to learn the knowledge or information about the problem.

In the Bayesian approach, the uncertainty about the weight parameters is estimated from data itself and represented by a probability distribution (Bishop, 1995). Apart from capturing the uncertainties and providing a natural interpretation on regularization techniques, Bayesian approach has some other useful aspects. Automatic relevance determination process is one of them, which can be used to identify the relative importance of different inputs. This method also allows making predictions by combining several networks (network committees) in order to obtain improved performance.

## Multi-Layer Perceptron (MLP)

MLPs are a popular class of feedforward networks which represent a multivariate non-linear function mapping between a set of input and output variables (Bishop, 1995). These networks are organized as several interconnected layers. Each layer is a collection of artificial neurons (nodes) where connections among the layers have not formed any loops, hence the name feedforward. Data have been fed

through the input layer, and then they pass through the hidden layer, and final outcome is given by the output layer.

The complexity of a MLP is directly proportional to the number of hidden nodes. It has been shown that a network with one hidden layer accompanied by sufficient number of hidden nodes is capable of approximating any continuous function (Hornik, Stinchcombe, & White, 1989). Therefore, we have considered a MLP with one hidden layer (Figure 1) and the final outcome is given by (1).

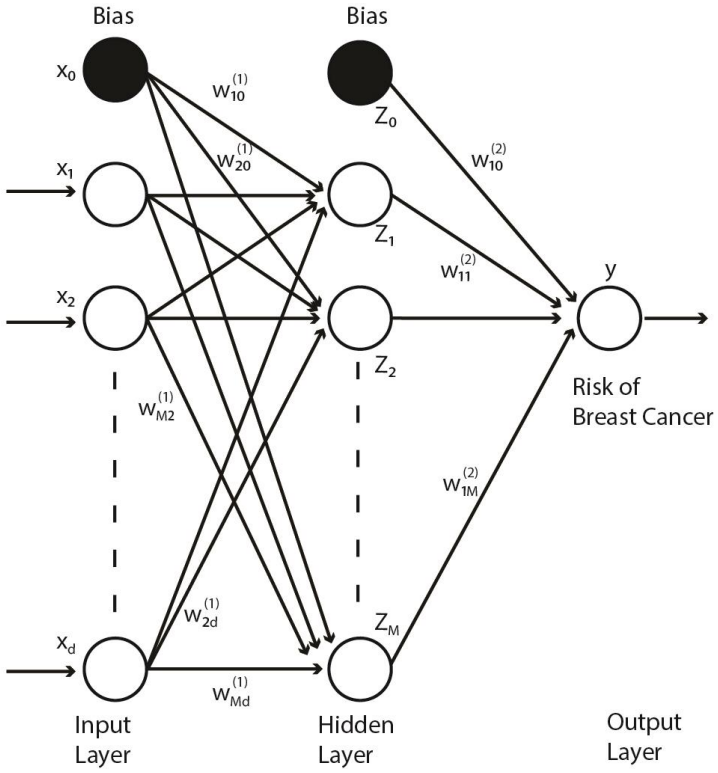


**Figure 1.** A multi-layer perceptron network (MLP)

$$y(x;w) = g(a) = g\left(\sum_{j=0}^{M} w_{1j}^{(2)} h\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i^{\circ}\right)\right) \tag{1}$$

During the training process, the goal is to minimize the difference between the actual and network predictions by adjusting the weights (including biases) using some optimization algorithms. A well trained MLP is capable of making

reasonable predictions to unseen data, which is known as generalization. This is achieved by incorporating the regularization techniques like weight decay (Bishop, 1995). Next, we discuss some theory related to MLP for a two-class classification problem.

### Two-Class Classification Problem

For a two class classification, logistic sigmoid is selected as the activation function in the output layer. This is the activation function "*g*" in (1), and has the form of

$$y(x;w)=\frac{1}{1+\exp(-a)} \tag{2}$$

In the Bayesian context, the $y(x; w)$ can be interpreted as the probability of membership in class $C_1$ given the input vector $x$. The probability of membership of class $C_2$ is then given by $(1 - y(x; w))$.

### MLP with Maximum Likelihood (Standard Network)

Network training (minimizing the difference between the actual and network predictions) can be done in two ways, using conventional maximum likelihood and Bayesian approaches. In maximum likelihood, a single set of most likely values for the weights are found whereas in Bayesian, a probability distribution for weights is obtained to represent the uncertainty in the weight estimation.

For a set of training data $\{x^n, t^n\}$ which are independent and identically distributed, the likelihood can be written as in (3) (Assuming the data are coming from a Bernoulli distribution). $G(D| w)$ is the negative logarithm of the likelihood which is defined as the cross entropy error function as given in (4).

$$P(D|w)=\prod_n p(t^n|x^n,w)=\prod_n y(x^n,w)^{t^n}\left(1-y(x^n,w)\right)^{1-t^n} \tag{3}$$

$$G(D|w)=-\sum_n\left\{t^n\ln y(x^n:w)+(1-t^n)\ln\left(1-y(x^n:w)\right)\right\} \tag{4}$$

Instead of maximizing the likelihood (since it is a monotonically decreasing function), it is more convenient to minimize the cross-entropy. When training the standard MLP in our analysis we have used this error function. The predictions on

new data are made using the optimal set of weights through the maximum likelihood method.

### MLP with Bayesian Techniques

In training a MLP, weights are adjusted whenever a new data point is presented to the network. A probability distribution which contains the degree of confidence associated with each different weight can be used to represent this uncertainty. The choice of prior distribution and about the corresponding posterior distribution will be discussed shortly.

### Network Regularization and Gaussian Prior

Smooth network mapping can be obtained by introducing network regularization techniques. This will lead for better generalization. In the simplest setting we have used a weight decay regularizer $E_w$ of the form (5).

$$E_w = \frac{1}{2}\|w\|^2 \tag{5}$$

As smaller weights (i.e a smaller $E_w$) are preferred for network weights, we have generated the weights from a zero mean Gaussian prior (6) initially.

$$P(w) = \frac{1}{Z_w(\alpha)}\exp\left(-\frac{\alpha}{2}\|w\|^2\right) = \frac{1}{Z_w(\alpha)}\exp\left(-\alpha E_w\right) \tag{6}$$

where $Z_w = \left(\frac{2\pi}{\alpha}\right)^{w/2}$ and, $\alpha$ is the inverse variance of the distribution which is known as the hyper-parameter of the prior distribution. As a part of Bayesian learning we can optimize the hyper-parameter $\alpha$ (evidence procedure).

### Posterior Distribution of Weights

The posterior probability distribution for weights can be determined according to the Bayes' theorem by incorporating the above prior (6) and the data likelihood (3),

$$P(w|D) = \circ \frac{1}{Z_s} \exp\left(-\left(G(D|w) + \alpha E_w\right)\right) = \frac{1}{Z_s} \exp\left(-S(w)\right) \quad (7)$$

where $Z_s$ is the normalization constant and $S(w)$ is the regularized cost function. The most probable weight vector $w_{MP}$ is found by maximizing the posterior, or minimizing the regularized cost function. From the second order Taylor series expansion of $S(w)$ around its minimum $w_{MP}$, we can obtain the following approximation.

$$S(w) \approx S(w_{MP}) + \frac{1}{2}(w - w_{MP})^T A(w - w_{MP}) \quad (8)$$

Where $A$ denotes the Hessian matrix of the regularized cost function. This leads to the Gaussian approximation to posterior distribution as given in (9) where $Z_s^*$ is the normalization constant.

$$P(w|D) = \circ \frac{1}{z_s^*} \exp\left(-S(w_{MP}) - \frac{1}{2}\Delta w^T A \Delta w\right) \quad (9)$$

Using the above posterior distribution, obtain the network predictions for the probability that a new input vector $x^*$ belongs to class $C_1$ as in (10). Although this prediction is not directly achievable, we can use marginalized predictions to obtain the results as suggested by (MacKay, 1992):

$$P(C_1|x^*, D) = \int P(C_1|x^*, w) P(w|D) dw = \int y(x, w) P(w|D) dw \quad (10)$$

## The Evidence Procedure

Prior to finding the above $w_{MP}$, it is needed to find the most probable hyper-parameter $\alpha_{MP}$, which maximizes the posterior probability of weights in Bayesian setting (MacKay, 1996) .This $\alpha_{MP}$ is obtained using the evidence $p(D|\alpha)$, by integrating the product of data likelihood and the prior distribution of the weights as given in (11).

$$p(D|\alpha) = \int p(D|w)\, p(w|\alpha)\, d\alpha \quad (11)$$

After several modifications, the logarithm of the evidence can be represented as in (12). The first term is the negative value of the regularized cost function, and the next two terms are the Occam factors that represent the ratio of posterior volume to prior volume. A network with higher number of hidden nodes has a large prior volume and thus, has a small Occam factor. Hence, these Occam factors act to penalize complex models and the evidence represents a trade-off between the accuracy and the complexity (MacKay, 1992).

$$\log E\upsilon = -S(\boldsymbol{w}) + \log(OCC_w) + \log(OCC_\alpha) \tag{12}$$

Periodically re-estimate $\alpha$ according to (13), in order to get the greatest log evidence value where $\gamma$ represents the effective number of weights whose values are controlled by the data rather than by the prior. Using that $\alpha_{MP}$ we can calculate the $w_{MP}$ (Thodberg, 1996). More details regarding this can be find in (Bishop, 1995).

$$\alpha^{New} = \frac{\gamma}{2E_w} \tag{13}$$

## The Automatic Relevance Determination

In the Bayesian setting, we can associate a separate hyper-parameter to each input variable which represents the inverse variance of the prior distribution of the weights fanning out from that input (Nabney, 2002). Optimal values for these hyper-parameters are obtained using the evidence procedure. So the weights connected to irrelevant inputs are automatically set to small values and this is known as the ARD approach.

## Committees

We can form a committee of networks to improve the prediction accuracies by combining several networks with different architectures. These networks can have different numbers of hidden nodes and/or they can be trained with different random initializations.

The simplest form of a committee, which involves taking the average predictions of the outputs of the $L$ networks, is given by (14). This will improve the accuracy of the predictions over an individual network output (Nabney, 2002).

$$y_{COM}(x) = \frac{1}{L}\sum_{i=1}^{L} y_i(x) \tag{14}$$

# Methodology

## Implementation of MLPs

***Study Population*** The data for this study are taken from Breast Cancer Surveillance Consortium (Barlow et al., 2006) for the period 1996 to 2002. The participating registries have obtained annual approvals from its institutional review board.

The data sample contains the information on menopausal type, age, breast density, ethnicity (Hispanic), body mass index (BMI), age at first birth, personal or family history of breast cancer, prior breast procedures, results of the last mammogram, type of menopause and current hormone therapy for each white woman. These women were aged from 35 to 84 years, and more details are available in Table 1.

## Implementation of the Standard and Bayesian MLPs

Training and testing data sets were created by partitioning the whole data sets each with 75% and 25% of data. A random sample out of the non-malignant group in the training set is selected and merged that with the malignant group in order to obtain a balanced training set. Table 2 represents the composition of data.

Different MLPs were trained using both standard and Bayesian approaches with varying number of hidden nodes from 1 to 25. For all of these MLPs, a logistic sigmoid activation function and scaled conjugate gradient (SCG) training algorithm were used. SCG is selected as it is a faster training algorithm compared to other algorithms (Penny & Roberts, 1999).

The standard MLP is trained using 10 fold cross-validation method and without any weight regularization. In 10 fold cross-validation, the training set is divided into 10 distinct segments, where 9 of those are used to train the network while the remaining segment is used for validation. This process is repeated for each of the 10 possible choices of the segments which are omitted from the training process and the validation errors (cross-entropy error) are averaged over all 10 results. The best network (with the corresponding hidden nodes) in this approach is the one with the smallest average cross-entropy in the validation data set (Kline & Berardi, 2005).

**Table 1.** Details of the Study Population

| | | Malignant (%) | | Not Malignant (%) | | Total | |
|---|---|---|---|---|---|---|---|
| | Total | 1053 | 6.47 | 15218 | 93.53 | 16271 | 100.00 |
| 1 | Menopausal Type ( $X_1$) | | | | | | |
| | Premenopausal | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |
| | Postmenopausal | 826 | 78.44 | 12336 | 81.06 | 13162 | 80.89 |
| 2 | Age Group ( $X_2$) | | | | | | |
| | 35-39 | 6 | 0.57 | 496 | 3.26 | 502 | 3.09 |
| | 40-44 | 72 | 6.84 | 788 | 5.18 | 860 | 5.29 |
| | 45-49 | 137 | 13.01 | 2355 | 15.48 | 2492 | 15.32 |
| | 50-54 | 168 | 15.95 | 2695 | 17.71 | 2863 | 17.60 |
| | 55-59 | 150 | 14.25 | 1872 | 12.30 | 2022 | 12.43 |
| | 60-64 | 141 | 13.39 | 1663 | 10.93 | 1804 | 11.09 |
| | 65-69 | 131 | 12.44 | 1533 | 10.07 | 1664 | 10.23 |
| | 70-74 | 96 | 9.12 | 1477 | 9.71 | 1573 | 9.67 |
| | 75-79 | 93 | 8.83 | 1343 | 8.83 | 1436 | 8.83 |
| | 80-84 | 59 | 5.60 | 996 | 6.54 | 1055 | 6.48 |
| 3 | Breast Density ( $X_3$) | | | | | | |
| | Almost entirely fat | 31 | 2.94 | 2575 | 16.92 | 2606 | 16.02 |
| | Scattered fibroglandular densities | 405 | 38.46 | 5319 | 34.95 | 5724 | 35.18 |
| | Heterogeneously dense | 506 | 48.05 | 4993 | 32.81 | 5499 | 33.80 |
| | Extremely dense | 111 | 10.54 | 2331 | 15.32 | 2442 | 15.01 |
| 4 | Hispanic ($X_4$) | | | | | | |
| | No | 1026 | 97.44 | 12476 | 81.98 | 13502 | 82.98 |
| | Yes | 27 | 2.56 | 2742 | 18.02 | 2769 | 17.02 |
| 5 | BMI ($X_5$) | | | | | | |
| | 10-24.99 | 432 | 41.03 | 4969 | 32.65 | 5401 | 33.19 |
| | 25-29.99 | 326 | 30.96 | 4404 | 28.94 | 4730 | 29.07 |
| | 30-34.99 | 181 | 17.19 | 3304 | 21.71 | 3485 | 21.42 |
| | 35 or more | 114 | 10.83 | 2541 | 16.70 | 2655 | 16.32 |
| 6 | Age at First Birth ( $X_6$) | | | | | | |
| | Age<30 | 692 | 65.72 | 7654 | 50.30 | 8346 | 51.29 |
| | Age 30 or greater | 154 | 14.62 | 3412 | 22.42 | 3566 | 21.92 |
| | Nulliparous | 207 | 19.66 | 4152 | 27.28 | 4359 | 26.79 |
| 7 | Number of first degree relatives with breast cancer ( $X_7$) | | | | | | |
| | Zero | 763 | 72.46 | 8515 | 55.95 | 9278 | 57.02 |
| | One | 252 | 23.93 | 5077 | 33.36 | 5329 | 32.75 |
| | Two or more | 38 | 3.61 | 1626 | 10.68 | 1664 | 10.23 |
| 8 | Previous breast procedure ( $X_8$) | | | | | | |
| | No | 716 | 68.00 | 8925 | 58.65 | 9641 | 59.25 |
| | Yes | 337 | 32.00 | 6293 | 41.35 | 6630 | 40.75 |
| 9 | Result of last mammogram before the index mammogram ( $X_9$) | | | | | | |
| | Negative | 1032 | 98.01 | 13244 | 87.03 | 14276 | 87.74 |
| | False positive | 21 | 1.99 | 1974 | 12.97 | 1995 | 12.26 |
| 10 | Surgical menopause ( $X_{10}$) | | | | | | |
| | Natural | 576 | 54.70 | 7000 | 46.00 | 7576 | 46.56 |
| | Surgical | 250 | 23.74 | 5336 | 35.06 | 5586 | 34.33 |
| | Unknown | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |
| 11 | Current hormone therapy( $X_{11}$) | | | | | | |
| | No | 400 | 37.99 | 6382 | 41.94 | 6782 | 41.68 |
| | Yes | 426 | 40.46 | 5954 | 39.12 | 6380 | 39.21 |
| | Unknown or not menopausal | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |

**Table 2.** Summary of the training and testing data sets

| Data set | Malignant | Non-Malignant | Total |
|---|---|---|---|
| Training set | 829 | 1658 | 2487 |
| Test set | 224 | 3030 | 3254 |
| Total | 1053 | 4688 | 5741 |

Under the Bayesian approach, four types of networks were trained with different weight regularization techniques. The first network is trained using 10 fold cross validation along with a weight regularization. The second and third types of the networks are trained using Bayesian evidence procedure, one without and the other with ARD prior. For both of the above types, 10 different networks were trained with 10 different random initializations to examine the effect of local minima on solutions, and they were taken to construct the network committees. The optimal MLP with the lowest average regularized cost function in the training data (or the highest average log evidence) is then selected and used to predict the posterior probability of malignancy by simply averaging 10 network predictions from each committee. Additionally, a same type of neural network with one hidden node was built for a comparison, which is functionally equivalent to a logistic regression model.

As the final network type, 10 different networks were trained on 10 different random samples with varying number of hidden nodes along with evidence process and ARD prior. The best MLP is selected using the minimum of the regularized cost function.

## Model Evaluation

The selected ANN models are evaluated based on their accuracy, sensitivity, specificity values and the area under the receiver operating characteristic curve (AUC) for the testing data (Bradley, 1997; Friedman & Wyatt, 2005). The proportions of correctly identified malignant and non-malignant women from the ANN models are known as the model accuracies. The proportions of actual malignant patients who are correctly identified from the models are known as the sensitivities and the proportions of non-malignant women who are correctly identified from the models are known as the specificities.

A perfect desirable predictor would be described as 100% sensitive (i.e. predicting all people from the malignant group as malignant) and 100% specific (i.e. predicting all non-malignant people as non-malignant). However, for any test,

there is usually a trade-off between these two measures, and this can be represented graphically by a receiver operating characteristic curve.

## Results

The summary of our six optimal network types is given in Table 3. Overall accuracy in the logistic network ($6^{th}$ MLP in the table) is lower than all other MLPs except for the MLP trained without ARD prior. Moreover it has the second lowest sensitivity and specificity values with the highest error. However, these models are not directly comparable in terms of their errors, as they have different settings and different training samples.

**Table 3.** Classification summary of the different MLP

| No | MLP Type | Error(Cross Entropy/Cost) | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|
| 1 | Standard MLP | 641.96(valid error 16.50) | 78.43% | 55.36% | 80.13% |
| 2 | MLP with weight regularization | 434.77(valid error 8.28) | 74.09% | 53.57% | 75.61% |
| 3 | MLP with evidence, but without ARD prior | 548.63 | 72.99% | 60.71% | 73.89% |
| 4 | MLP with both evidence and ARD prior | 582.28 | 74.15% | 59.82% | 75.21% |
| 5 | MLP trained on different samples with evidence and ARD prior | 908.78 | 81.35% | 59.38% | 82.97% |
| 6 | MLP with one hidden node (logistic) | 1123.10 | 73.11% | 55.35% | 74.42% |

Out of these MLP types, the best network in terms of the highest accuracy and specificity is found to be the MLP trained using different samples along with both evidence procedure and ARD prior ($5^{th}$ MLP). As can be seen, use of the evidence procedure and the ARD prior has always resulted in better sensitivities. However, use of weight regularization without any optimization (evidence process) does not provide any significant improvement over the standard network training process.

It can be concluded that use of weight regularization techniques along with evidence process gives better results in Bayesian classification for most of the time. Apart from that, use of ARD prior helps to identify the most contributing variables to the network. Overall, Bayesian methods are preferred over the standard method mainly because of the natural way of handling the weight

regularization. By forming committees, we were able to reduce the network training error.

The minimum and maximum prediction accuracies from these MLPs are 73% and 81%, respectively. Sensitivity values are varying from a minimum of 54% up to a maximum of 61% while specificity values are varying from 74% to 83%.

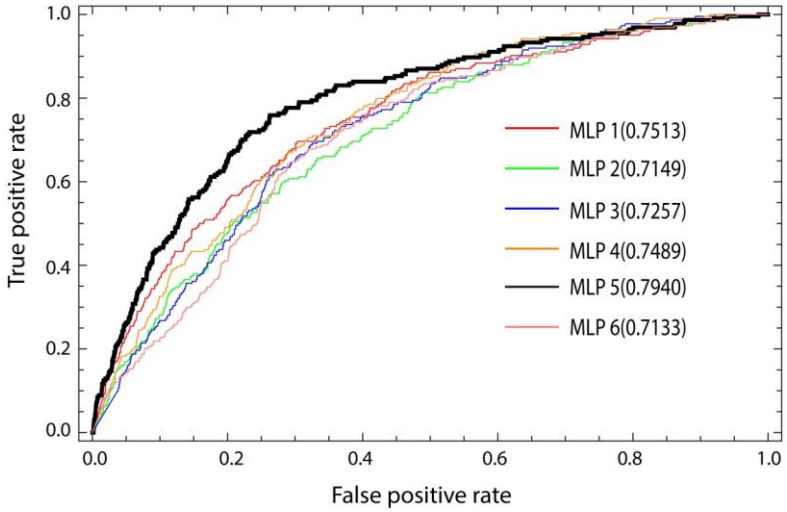


**Figure 2.** The receiver operating characteristic curves and the AUC values

The AUC values of all the above MLPs are greater than 70%, which implies a moderate classification model. Figure 2 represents the receiver operating characteristic curves with the corresponding AUC values. The posterior probabilities of malignancy were obtained from the best Bayesian MLP network selected.

ARD prior identifies the relevant importance of the inputs in the network. Table 4 includes the rankings of the variable based on these hyper-parameter values. Risk factors with smaller hyper-parameters are highly contributing to the model outcome. Being in the age group 75 to 79 is the most critical factor in diagnosing with malignant breast cancer. Having a prior false positive mammogram can be an indication of malignant breast cancer. In accordance with cancer literature, risk factors such as having heterogeneously or extremely dense breast densities, and having a BMI of 35 or more are significantly contributing to the model.

**Table 4.** Rankings of the attributable variables based on the ARD prior

| Rank | Alpha (hyper-parameter) | Variable | Risk Group |
|---|---|---|---|
| 1 | 0.3841 | agegrp9 | Age group 75-79 |
| 2 | 0.5550 | lastmamm | Result of last mammogram before the index mammogram - False positive |
| 3 | 0.6489 | density3 | Density - Heterogeneously dense |
| 4 | 0.6846 | density4 | Density - Extremely dense |
| 5 | 0.8251 | bmi4 | 35 or more |
| 6 | 1.3072 | agegrp2 | Age group 40-44 |
| 7 | 1.3872 | agegrp7 | Age group 65-69 |
| 8 | 1.6989 | hispanic | Hispanic or not - Yes |
| 9 | 1.7403 | nrelbc2 | Number of first degree relatives with breast cancer - 2 or more |
| 10 | 1.9510 | hrtYes | Current hormone therapy – Yes |
| 11 | 2.0528 | agegrp10 | Age group 80-84 |
| 12 | 2.0826 | bmi2 | 25-29.99 |
| 13 | 2.1980 | agegrp8 | Age group 70-74 |
| 14 | 2.2112 | hrtNo | Current hormone therapy - No |
| 15 | 2.8161 | agegrp6 | Age group 65-69 |
| 16 | 2.9341 | bmi3 | 30-34.99 |
| 17 | 3.2299 | agegrp5 | Age group 55-59 |
| 18 | 3.6520 | nrelbc1 | Number of first degree relatives with breast cancer - One |
| 19 | 3.7138 | surgnatural | Surgical menopause - Natural |
| 20 | 4.2249 | agegrp4 | Age group 50-54 |
| 21 | 5.0616 | surgsurgical | Surgical menopause - Surgical |
| 22 | 5.1547 | brstproc | Previous breast procedure - Yes |
| 23 | 5.7224 | density2 | Density - Scattered fibroglandular densities |
| 24 | 7.2989 | menopaus | Postmenopausal or age>=55 |
| 25 | 10.1388 | agenulli | Age at first birth - Nulliparous |
| 26 | 10.5538 | agegrp3 | Age group - 45-49 |
| 27 | 11.4664 | agegreater30 | Age at first birth - Age 30 or greater |

## Conclusion

A breast cancer detection model was introduced using artificial neural network theory. With the intention of having a better classification, different types of MLPs were developed. These models are trained using the standard and Bayesian techniques. The first two models were validated using 10-fold cross validation and we have constructed committees for the other models. Finally all MLPs were tested on a new set of test data.

The advantage of Bayesian MLP is that it gives the posterior probabilities for classification which can be used as a priori risk of diagnosing with breast cancer. The evidence procedure is used for the network regularization along with ARD prior. Use of ARD prior did not make any significant difference in the accuracy of our optimal MLP. Use of committees also did not show much difference in the overall results compared to the single network predictions alone. However, this has helped to give a low variance in the predictions.

The highest accuracy which was obtained from one of the Bayesian MLP is about 81% and this is a significant improvement over the other methods which used the same set of real data in terms of the discriminative accuracy. ROC curve provides information about a model's classification efficiency. A good classification model was obtained for the third and the fifth MLP with more than 75% area under the ROC curve. The model may be further improved by considering more relevant risk factors and more recent data, such as different races because ethnicity is one of the significant risk factors that contributes to the malignancy of breast cancer (Xu, Kepner, & Tsokos, 2011).

It was also confirmed that ANN may have an important role in improving the accuracy and consistency of medical diagnosis. The proposed approach in developing the ANN model is free of assumptions, as opposed to parametric regression and hence increases the validity of our findings.

## References

Amir, E., Evans, D. G., Shenton, A., Lalloo, F., Moran, A., Boggis, C., … Howell, A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *Journal of Medical Genetics, 40*(11), 807-814. doi: 10.1136/jmg.40.11.807

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E., Jr., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural

networks revisited: Discrimination and calibration. *Cancer, 116*(14), 3310-3321. doi: 10.1002/cncr.25081

Barlow, W. E., White, E., Ballard-Barbash, R., Vacek, P. M., Titus-Ernstoff, L., Carney, P. A., … Kerlikowske, K. (2006). Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute, 98*(17), 1204-1214. doi: 10.1093/jnci/djj331

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York, NY: Oxford University Press.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Claus, E. B., Risch, N., & Thompson, W. D. (1993). The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Research and Treatment, 28*(2), 115-120. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8173064

Domchek, S. M., Eisen, A., Calzone, K., Stopfer, J., Blackwood, A., & Weber, B. L. (2003). Application of breast cancer risk prediction models in clinical practice. *Journal of Clinical Oncology, 21*(4), 593-601. doi: 10.1200/jco.2003.07.007

Euhus, D. M., Leitch, A. M., Huth, J. F., & Peters, G. N. (2002). Limitations of the Gail Model in the specialized breast cancer risk assessment clinic. *The Breast Journal, 8*(1), 23-27. doi: 10.1046/j.1524-4741.2002.08005.x

Floyd, C. E., Jr., Lo, J. Y., Yun, A. J., Sullivan, D. C., & Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. *Cancer, 74*(11), 2944-2948. doi: 10.1002/1097-0142(19941201)74:11<2944::AID-CNCR2820741109>3.0.CO;2-F

Friedman, C. P., & Wyatt, J. (2005). *Evaluation methods in biomedical informatics*. Springer Science & Business Media. Retrieved from https://books.google.com/books?id=o2XyIpz_I1sC&pgis=1

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., & Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute, 81*(24), 1879-1886. doi: 10.1093/jnci/81.24.1879

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. USA: Prentice Hall.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359-366. doi: 10.1016/0893-6080(89)90020-8

Hung, M. S., Shanker, M., & Hu, M. Y. (2002). Estimating breast cancer risks using neural networks. *Journal of the Operational Research Society, 53*(2), 222-231. doi: 10.1057/sj/jors/2601276

Kline, D. M., & Berardi, V. L. (2005). Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing and Applications, 14*(4), 310-318. doi: 10.1007/s00521-005-0467-y

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation, 4*(5), 720-736. doi: 10.1162/neco.1992.4.5.720

MacKay, D. J. C. (1996). Hyperparameters: Optimize, or integrate out? In G. R. Heidbreder (Ed.), *Maximum entropy and bayesian methods* (pp. 43-59). Retrieved from http://link.springer.com/chapter/10.1007/978-94-015-8729-7_2\npapers3://publication/uuid/4C73D436-E70B-49D8-AA7E-031803718650

Nabney, I. T. (2002). NETLAB: Algorithms for pattern recognition. Gateshead: Springer.

Orr, R. K. (2001). Use of an artificial neural network to quantitate risk of malignancy for abnormal mammograms. *Surgery*, *129*(4), 459-466. doi: 10.1067/msy.2001.112069

Penny, W. D., & Roberts, S. J. (1999). Bayesian neural networks for classification: How useful is the evidence framework? *Neural Networks, 12*(6), 877-892. doi: 10.1016/S0893-6080(99)00040-4

Thodberg, H. H. (1996). A review of Bayesian neural networks with application to near infrared spectroscopy. *IEEE Computational Intelligence Society, 7*(1), 56-72. doi: 10.1109/72.478392

Van Asperen, C. J., Jonker, M. A., Jacobi, C. E., van Diemen-Homan, J. E. M., Bakker, E., Breuning, M. H., … de Bock, G. H. (2004). Risk estimation for healthy women from breast cancer families: new insights and new strategies. *Cancer Epidemiology, Biomarkers & Prevention, 13*(1), 87-93. doi: 10.1158/1055-9965.EPI-03-0090

Wu, Y., Giger, M. L., Doi, K., Vyborny, C. J., Schmidt, R. A., & Metz, C. E. (1993). Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. *Radiology*, *187*(1), 81-87. doi: 10.1148/radiology.187.1.8451441

Xu, Y., Kepner, J., & Tsokos, C. P. (2011). Identify attributable variables and interactions in breast cancer. *Journal of Applied Sciences, 11*(6), 1033-1038. doi: 10.3923/jas.2011.1033.1038