BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION

H. Yusuff[1*], N. Mohamad[2], U.K. Ngah[3] & A.S. Yahaya[4]
[1,2,] School of Electrical and Electronic Engineering, Engineering Campus, University Sains Malaysia 14300 Nibong Tebal , Penang, Malaysia
[3] Imaging & Computational Intelligence Research Group (ICI)
School of Electrical and Electronic Engineering, Engineering Campus, University Sains Malaysia
[4] School of Civil Engineering, Engineering Campus, University Sains Malaysia 14300 Nibong Tebal, Penang, Malaysia

ABSTRACT
In this study, the diagnosis of breast cancer from mammograms is complemented by using logistic regression. The radiologists can use the results to make a proper judgment as to the presence of breast cancer. Data were obtained from survey questions completed by the radiologist during his observation of the patients. The results using logistic regression cross tabulation was to obtain the significant values between the breast cancer factors. The classification table from 130 samples shows the occurrence from prediction and observation samples, producing percentage of correct classification for mammogram results is 91.5%. The accuracy is compared with validated samples which are 46 samples and the percentage of correct classification is 67.4%. The analysis for mammograms screening using parameter estimation is to identify all the factors that were available in the survey. The presence of mass, architectural distortion, skin thickening, and calcification had high odds of getting breast cancer.

Keywords: *breast cancer, mammograms, prediction, logistic regression, factors*

1. INTRODUCTION

There are many different types of breast cancer, with different stages or spread, aggressiveness, and genetic makeup. Survival rates for breast cancer may be increased when the disease is detected in its earlier stage through mammograms. The implementation of mass screening would result in increased caseloads for radiologists. This will increase chances of improper diagnosis. The prediction using logistic regression would aid the radiologist to detect the breast cancer.

The patient's history is used to predict and detect whether the patient had breast cancer or not. The patient's history include information about their age, menopause condition, age of menopause, whether the patient had any first degree relative with history of breast cancer or family member having cancer other than breast cancer, and if the patient had breast trauma. These variables may be the cause of breast cancer. Patient who has first degree relative with breast cancer or family member having cancer other than breast cancer has a high probability of breast cancer (Colditz, 1993; Gui, 2001; Rawal, 2006; Wiseman, 2004). A patient who had only a sister with breast cancer or other types of cancer has less probability of getting breast cancer compared to a patient who has a mother or sister and mother who has breast cancer or other types of cancer. A patient's menopause history can also be used to predict breast cancer (Ganz, 2005; Wiseman, 2004).

The patient's history can help the doctor to decide on the next mode of detection procedure. The next cause of action is to conduct clinical examination. The clinical examination is physical examination of both breasts by the doctors using the hands (Goodson III, 2010). The clinical examination includes inspection and palpation of the entire breast area including the lymph node areas above and below the collarbone under each arm. The doctor will gently palpate each breast. Special attention will be given to the shape and texture of the breasts, location of any lumps, and whether such lumps are attached to the skin or to deeper tissues. The lumps detected by doctors need not be cancer lumps as the lumps might be constituted by a trauma. Breast trauma is an injury to the breast. Thus, a patient with breast trauma may not have breast cancer because the lump may result from injury (Gatta, 2006). The doctor also inspects the condition of the nipple for the presence of any discharge.

The next procedure in breast examination is to undergo mammographic screening which is to aid in the diagnosis of breast disease in women. Mammography is a common screening method since it is relatively fast and is widely available in developed countries. Diagnostic mammography is used to evaluate a patient with abnormal clinical examination results. The results detected on the mammogram are mass, architectural distortion, skin thickening, and calcification (Balleyguier, 2007; Eltoukhy, 2010; Moezzi, 1996). Radiologists can predict the condition of the patient from the results of the mammogram. The levels of breast cancer may be based on the presence and conditions of mass and calcification as well as the presence of architectural distortion and skin

thickening (Balleyguier, 2007; Eltoukhy, 2010). The conditions of mass are location, margin, shape, size, and density. The conditions of calcification are their types, shape and distribution.

Logistic regression is one of the variety of popular multivariate tools used in biomedical informatics. It is one of the most common models for prediction and has been applied to cancer prediction (Samatha, 2009; Zhou, 2004). From previous studies, logistic regression is widely used in medical literature especially for corelating the dichotomous outcomes with the predictor variables that include different physiological data. In logistic regression, the predicted odd ratio of positive outcome is expressed as a sum of product. Product is formed by multiplying the values of independent variable and its coefficients. The probability of positive outcome is obtained from the odd ratio through a simple transformation (Samatha, 2009). The problems are formulated first from the logistic regression. Then, the coefficient obtained from the logistic regression is used to calculate the predictor variables (Zhou, 2004).

Because of the fact that the detection of breast cancer and prediction of the breast cancer level is important, numerous researches have been conducted in this area. These include prediction using logistic regression. Logistic regression is used for prediction by fitting data to the logistic curve. It requires the fitted model to be compatible with the data. In logistic regression, the variables are binary or multinomial. Multinomial Logistic Regression analysis is capable of showing the best way to find conclusion and be made as parsimonious model to describe the relationship between dependent and independent variables. Binary Logistic Regression is one of the logistic regression analysis methods whereby the independent variables are dummy variables. Independent variables consist of different size levels whereas dependent variables must be linear and fulfills the response that is needed for this method. A logistic regression model is the result of non-linear transformation of the linear regression model. The difference between logistic regression and linear regression is that the outcome variable in logistic regression is dichotomous (Hosmer & Lemeshow, 2000).

This study uses data from the mammogram results to determine the patient condition; (i) positive of breast cancer, (ii) uncertain of breast cancer, or (iii) negative of breast cancer. The logistic regression model from the mammogram is used to predict the risk factors of patient's history.

Logistic regression analysis can verify the predictions made by doctors and/or radiologists and also correct the wrong predictions. In this analysis, the logistic regression also calculates the mammogram results that contribute to breast cancer. Thus, the results of the analysis are compared to the prediction made by the doctor or radiologists.

Logistic regression models are then created from the mammogram results. The predictions of the risk factors of the patients' history are based on the mammogram results and logistic regression model. By using logistic regression to predict the causes from the patients' history, the patient does not need follow-up checks and may skip the subsequent steps in the course of her patient management. The healthcare giver can also determine the patient's levels of breast cancer without subjecting the patient to doing the mammogram screening.

2. MATERIALS AND METHODS

2.1 Data

The data for this study was collected from July to August 2006 by a consultant radiologist. Patients who had performed mammogram screening were randomly selected and interviewed about their history. Information from 176 patients were recorded based on their history and mammogram screening results in the survey form. Specialist records from the patients are patient menopause, first degree relative with breast cancer, family member with other cancer, patient's previous history of breast trauma, the presence of mass, architectural distortion, skin thickening, and the presence of the calcification. These eight variables are considered for classifying the presence of breast cancer.

The independent variables are patient menopause (H1), first degree relative with breast cancer (H2), family member with other cancer (H3), patient's previous history of breast trauma (H4), the presence of mass (M1), architectural distortion (M2), skin thickening (M3), and the presence of the calcification (M4).

The data description of the eight independent variables and the dependent variable are provided in Table 1. Data analysis is performed using SPSS (2007), V16.0, SPSS Inc.

Table 1: Description of variables

Variable	Definition	Characteristic
BC	The presence of breast cancer	0 = negative, 1 = uncertain, 2 = positive
H1	Patient Menopause	0 = No,1 = Yes
H2	First degree relative with breast cancer	0 = No,1 = Yes
H3	Family member with other cancer	0 = No,1 = Yes
H4	Previous history of breast trauma	0 = No,1 = Yes
M1	The presence of mass	0 = No,1 = Yes
M2	Architectural distortion	0 = No,1 = Yes
M3	Skin thickening	0 = No,1 = Yes
M4	The presence of calcification	0= No,1 = Yes

2.2 Logistic Regression Analysis

2.2.1 Variables Selection

It is important that the model include all relevant variables, it is also important the model does not start with more variables than are justified for the given number of observations (Bangley, 2001; Concato, 1993; Peduzzi, 1995). For the set of data, more variables generally produce a better model fit to the data. However, excessive variables will influence the coefficient in the model and contribute to the over-fitting model. A complicated model including many insignificant variables may result in less predictive power and it may often be difficult to interpret the results. There are two methods for variables selection namely filter and statistical (Austin, 2004; Genuer, 2010).

In the case of filter method, the variables are reduced based on the importance of the independent variables. By identifying the independent variables that will be used in the model, the risk factors are reduced. There are several procedures what need to be conducted. First, the patients are differed between studies. Some studies are based on patient enrolled in the tests conducted by the doctors. Second, the studies are differed in term of variables collected. Incomplete or missing variables are taken out from the analysis. Third, the remainders of the independent variables are compared with the previous study (Austin, 2004). The survey data by the breast specialist had too many variables and contribute to the imprecise of the analysis. So, the variables are reduced based on the previous research.

For statistical methods, correlation analysis is conducted. Two predictor variables that are highly correlated with each other present a problem for any regression analysis (Bangley, 2001), Feinstein (1996)). The variables that are highly correlated with each other can contribute to inaccuracy in the logistic regression analysis. There are two procedures for the statistical method of variables selection. First procedure is interaction test. Interactions are represented as product terms which is the term in the regression model and is not a single predictor variable but the product of two predictor (Bangley, 2001; Hosmer & Lemeshow, 2000; Kleinbaum, 1994). The interactions tests were performed to find the significant values of each variable. The significance of the interaction is measured and reported. The test is the cross tabulation test and the values were taken from Pearson Chi – Square. The Pearson Chi – Square is expressed as

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

The second procedure is co-linearity analysis. The variance associated with these coefficients will be increased with a consequent loss of statistical significance (Bangley, 2001). Co-linearity analysis was based on the significant values from the interaction test. The significant values for each variable must be lower than 0.20 (Hosmer & Lemeshow, 2000). The variables that had significant values below 0.20 are chosen into the logistic regression model analysis.

2.2.2 Validation

The validation analysis was performed so as to check whether the logistic regression analysis is suitable or not (Bangley (2001), Feinstein (1996)). The prediction percentage of correct cases from the main samples must be greater than or equal to the validated samples. The validation is using other sample data but having the same coefficient values as the main data to calculate the percentage of correct cases. First, the data were divided into two. The first data containing 80% of the samples is used as the main data and used to find the coefficient values. The second data which contain 20% of the samples is used to validate the main data. Secondly, after obtaining the coefficient values from the main data, the probability of each sample from the validated data are calculated. The probability is defined as:

$$P(Y = m) = \frac{\exp(g(x))}{1 + \sum \exp(g(x))} \quad (2)$$

Reference probability is defined as

$$P(Y = 0) = \frac{1}{1 + \sum \exp(g(x))} \quad (3)$$

with $g(x) = \beta_0 + \sum_{p=1}^{n} \beta_p x_p$ (4)

β_0 is the intercept coefficient values, β_p is the coefficient value for each factor that contributes to the occurrence. Thirdly, the probability of each sample is cross-validation with the observed probability. With the cross-validation, the percentage of correct cases of classification is obtained. Then, the percentage of correct cases of classification of the validated data is compared with the percentage of correct cases of classification of the main data.

The data were divided into two. The first 130 samples were used to obtain the logistic regression model. The remainders of the samples were used to validate the model. The confirmed results are used from the percentage of correct cases of the classification.

2.2.3 Logistic Regression Model

Several tests had been performed in the logistic regression analysis. The tests are model fitting test, parameter estimation and classification. Model fitting test is to check whether all the variables are suitable to be used in the logistic regression. Model fitting test is done by using likelihood ratio statistic. Likelihood ratio is defined as

$$LR[i] = -2(LL(\boldsymbol{a}) - LL(\boldsymbol{a}, \boldsymbol{B})) \quad (5)$$

where $LL(\boldsymbol{a})$ is the log-likelihood of the beginning model and $LL(\boldsymbol{a}, \boldsymbol{B})$ is the log-likelihood of the ending model. Likelihood ratio is distributed chi-square with i degree of freedom.

Parameter estimation is to estimate each independent variables that contribute to the presence of breast cancer. Parameter estimation is done using log-odd ratio. The log-odd ratio is defined as

$$B = \ln\left[\frac{P(Y=m)}{P(Y=0)}\right] = = \beta_0 + \sum_{p=1}^{n} \beta_p x_p \quad (6)$$

where *m = 1,2,…,n*.

To find the log-odd ratio, the probability of each event is calculated. Odds ratio measure the incidence when the independent variable increases by one unit. The odds ratio is defined as

$$\frac{P(Y=m)}{P(Y=0)} = \exp\left(\beta_0 + \sum_{p=1}^{n} \beta_p x_p\right) \quad (7)$$

Classification is to predict the patients in the presence of breast cancer. From the calculated coefficients, the probability of each sample is calculated. The probability is defined as

$$P(Y = m) = \frac{\exp(g(x))}{1 + \sum \exp(g(x))} \quad (8)$$

while for the reference category;

$$P(Y = 0) = \frac{1}{1 + \sum \exp(g(x))} \quad (9)$$

3. RESULTS AND DISCUSSION:

3.1 Data

Among the patients, 45.5% of them have gone through the stage of menopause while 54.5% have not. Patients who have had first degree relative with breast cancer are 17.6% while 82.4% do not. 4.5% of the patients had family members with other cancers while 95.5% of the patients do not have family member with other cancers. The patients who have had history of breast trauma are 4.0% whereas the patients who do not are 96.0%. The bar chart of the patients' history statistics is in Figure 1 and the summary of the patient's history statistic is in Table 2.

The remainders of the independent variables are detected through mammogram. About 48% of patients had mass while 52% had not. The patients who had architectural distortion are 6.8% while 93.2% did not. 3.4% of the patients had skin thickening while 93.6% had not. 34.1% of the patients had calcification whereas 65.9% did not. The bar chart of the mammogram results is described in Figure 2 and the summary of the mammogram statistics is in Table 3.

The eight variables are binary variable, where 0 indicates non-occurrence and 1 indicates occurrence. Dependent variable is multinomial, where 0 indicates negative of breast cancer, 1 indicates uncertain of breast cancer, and 2 indicates positive breast cancer.

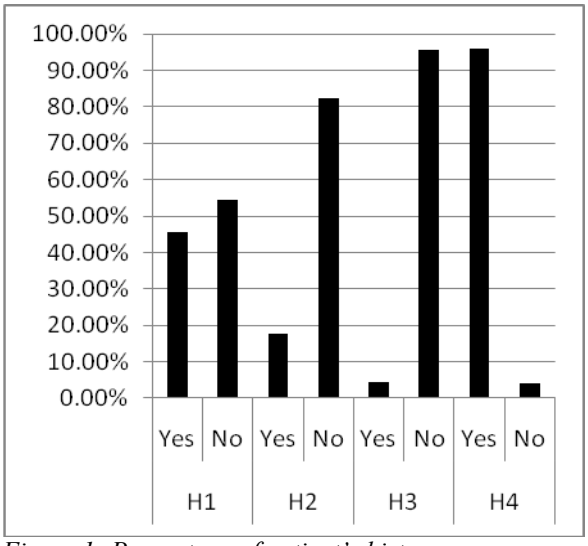

Figure 1: Percentage of patient's history

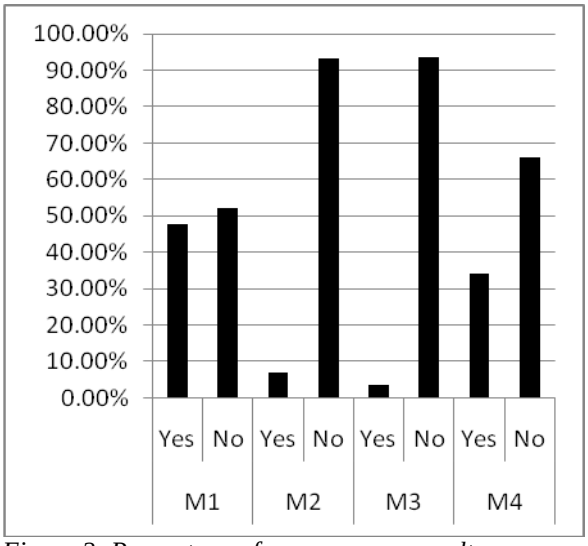

Figure 2: Percentage of mammogram results

Table 2: Patient's history statistic

	Yes	No
Menopause	45.5%	54.5%
1st degree relative with breast cancer	17.6%	82.4%
Family member with other cancer	4.5%	95.5%
Breast trauma	4.0%	96.0%

Table 3: Mammogram results statistic

	Yes	No
Mass	47.7%	52.3%
Architectural distortion	6.8%	93.2%
Skin thickening	3.4%	93.6%
Calcification	34.1%	65.9%

3.2 Variables Selection

First, the survey forms showed that several tests were conducted by the radiologist. By filtering the conducted test, only two tests are taken. The tests are patient's history and mammogram results. Mammogram results are used to create the model, so the mammogram results need to be considered. Second, the variables in the mammograms results and patient's history are filtered by the missing or incomplete data. Third, the remainders of variable are compared with the previous study. Based on previous research, a patient who has a relative with first degree breast cancer has a higher risk of having breast cancer (Colbitz (1993), Gui (2001), Rawal (1998)). In addition, the patients who have family members with other cancers, such as ovary, cervix, and lymphoma cancer, have a higher risk of breast cancer (Colbitz (1993), Gui (2001), Rawal (1998)). The patients with breast trauma probably have breast cancer (Gatta (2006)). The risks of breast cancer for patients who have mass, architectural distortion, skin thickening, and calcification through mammogram screening is high (Balleyguier (2007), Eltoukhy (2010), Moezzi (1996)). Thus, variables that are considered to be analyzed are patient menopause, first degree relative with breast cancer, family member with other cancer, patient's previous history of breast trauma, the presence of mass, architectural distortion, skin thickening, and the presence of the calcification.

The remainders of the independent variables are selected using statistical method. First, using the cross tabulation test to find the Pearson chi-square significant values, the values for each variable is recorded. The significant values for patient menopause, first degree relative with breast cancer, family member with other cancer, and patient's previous history of breast trauma are 0.496, 0.194, 0.089, and 0.328 respectively. The significant values for the presence of mass, architectural distortion, skin thickening, and the presence of the calcification are lower than 0.001. The results are in Table 4.

Table 4 shows the p-values for each variable. Using the values obtained in the interaction test, the correlation analysis is performed. The variables that had significant values greater or equal to 0.20 are correlated with other variables and the variables are patient menopause and patient's previous history of breast trauma. The variables that had significant values below 0.20 are considered to be taken into the logistic regression model. The variables that had values below than 0.20 are first degree relative with breast cancer, family member with other cancer, presence

of mass, architectural distortion, skin thickening, and the presence of the calcification. Thus, these variables are considered in the logistic regression analysis.

Table 4: Significant Values for Each Independent Variable

Variables	Significant Values
Patient menopause	0.496
First degree relative with breast cancer	0.194
Family member with other cancer	0.089
Patient's previous history of breast trauma	0.328
The presence of mass	< 0.001
Architectural distortion	< 0.001
Skin thickening	< 0.001
The presence of the calcification	< 0.001

3.3 Validation

The data were divided into two. Using the 130 samples of data, logistic regression model is created. The coefficient for logistic regression model is shown in Table 8. Using this coefficient, the logistic regression of each category for 46 samples is calculated. The logistic regression is as following

For positive category;

$$P(Y=2)=\frac{\exp g(2)}{1+\sum \exp(g(x))} \quad (10)$$

where; $$g(2)=-4.576+1.707*\mathrm{M1}+10.100*\mathrm{M2}+7.823*\mathrm{M3}+2.900*\mathrm{M4} \quad (11)$$

For uncertain category; $$P(Y=1)=\frac{\exp g(1)}{1+\sum \exp(g(x))} \quad (12)$$

where; $$g(1)=-1.389+3.076*\mathrm{M1}+33.105*\mathrm{M2}+2.172*\mathrm{M3}+2.472*\mathrm{M4} \quad (13)$$

For negative category as reference category $$P(Y=0)=\frac{1}{1+\sum \exp(g(x))} \quad (14)$$

where; $$g(x)=g(1)+g(2) \quad (15)$$

After logistic regression for each category is calculated, the predicted category is cross-validation with the observed category to obtain the percentage of correct classification. The percentage of correct classification for 46 samples is 67.4%. The percentages of correct classifications of two groups of samples are compared. For 130 samples, the percentage of correct classification is 91.5%. The percentage of correct classification for 130 samples is greater than percentage of correct classification for 46 samples. The validation criteria are satisfied. These total numbers of data is suitable to be used for logistic regression model. The percentages of correct classifications are in Table 5.

Table 5: Classification Summary of Percentage Correct

Groups	Percentage Correct
130 samples	91.7%
46 samples	67.4%

3.4 The Logistic regression model

3.4.1 Patient's History Analysis

The reference category for the respond variable is negative of breast cancer. Uncertain relative to negative of breast cancer, and positive relative to negative of breast cancer are used to estimate the parameter of the model. For patients with first degree relative with breast cancer, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is decreased by 0.648. The odd ratio for the uncertain of breast cancer relatively to negative is 0.523. The probability of uncertain of breast cancer is 0.523 times higher than negative of breast cancer. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 0.308. The odd ratio value is 1.360. The odd of getting breast cancer is 1.36 times if a patient has first degree relative with breast cancer.

For a patient with family member with other cancer, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is decreased by 1.426. The odd ratio for the uncertain of breast cancer relatively to negative is 0.240. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 0.487. The odd ratio value is 1.627. The odd of getting breast cancer are 1.63 times if patient has family member with other cancer.

Thus, evaluating the patients with these factors predicted have higher probability of getting breast cancer. Table 6 shows the parameter estimations for first degree relative with breast cancer and family member with other cancer.

Table 6: Parameter Estimations for Patient's History Analysis

The Presence of Breast Cancer		B	exp (B)
UNCERTAIN	Intercept	0.981	
	First degree relative with breast cancer	-0.648	0.523
	Family member with other cancer	-1.426	0.240
POSITIVE	Intercept	-2.027	
	First degree relative with breast cancer	0.308	1.360
	Family member with other cancer	0.487	1.627

3.4.2 Mammogram Analysis

For this analysis, the reference of the dependent variable is negative of breast cancer. Uncertain of breast relative to negative of breast cancer and positive of breast cancer relative to negative of breast cancer are used to estimate the parameter of the model. Since the parameter estimates are relative to the reference group, the standard interpretation of the multinomial logistic regression is that for a unit change in the predictor variable, the logistic regression of outcome relative to the reference group is expected to change by its respective parameter estimate which is in log-odds units, given that the other variables in the model are held constant.

For patients with mass, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is increased by 3.076. The odd ratio for the uncertain of breast cancer relatively to negative is 21.663. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 1.707. The odd ratio value is 5.512. The odd of patient afflicted with breast cancer is 5 times higher when the mass is detected in mammogram.

For patients with architectural distortion, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is increased by 33.105. The odd ratio for the uncertain of breast cancer relatively to negative is 2.38×10^{14}. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 10.100. The odd ratio value is 2.434×10^{4}. The odd patient getting breast cancer is 2.4×10^{4} times when architectural distortion is detected in mammogram.

For patients with skin thickening, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is increased by 2.172. The odd ratio for the uncertain of breast cancer relatively to negative is 8.777. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 7.823. The odd ratio value is 2.496×10^{3}. The odd patient afflicted with breast cancer is 2.5×10^{3} times when skin thickening is detected in mammogram.

For patients with calcification, the log-odd value, B, for uncertain of breast cancer relatively to negative of breast cancer is increased by 2.427. The odd ratio for the uncertain of breast cancer relatively to negative is 11.326. The log-odd value, B, for positive of breast cancer relatively to negative of breast cancer is increased by 2.9. The odd ratio value is 18.167. The odd for patient who has calcification detected is 18 times of getting breast cancer.

Table 7 shows the parameter estimations for mass, architectural distortion, skin thickening, and calcification. This mammogram analysis is used to create the model for patient's history analysis.

Using the coefficient from 130 samples, the logistic regression model is created. The model is used to predict the data of 46 samples. One of the observed uncertain of breast cancer is predicted as positive of breast cancer. The entire observed negatives of breast cancer are predicted as uncertain of breast cancer. Observed positive of breast cancer is predicted as uncertain of breast cancer. Thus, the percentage of correct classifications is 67.4%. The percentage is low because samples are not equally distributed. For observed positive of breast cancer, there is only one sample. This sample is predicted into the category of uncertain of breast cancer and this contributes toward the decrement of 100% of correct classifications. The classification of 46 samples is shown in Table 8.

Table 7: Parameter Estimations for Mammogram Analysis

The Presence of Breast Cancer		B	exp (B)
UNCERTAIN	Intercept	-4.576	
	Mass	3.076	21.663
	Architectural distortion	33.105	2.000×10^{14}
	Skin thickening	2.172	8.777
	Calcification	2.427	11.326
POSITIVE	Intercept	-1.389	
	Mass	1.707	5.512
	Architectural distortion	10.100	2.434×10^{4}
	Skin thickening	7.823	2.496×10^{3}
	Calcification	2.900	18.167

Table 8: Classification of 46 samples

	Predicted			Percentage Correct
Observed	Negative	Uncertain	Positive	
Negative	0	13	0	0.0%
Uncertain	0	31	1	96.7%
Positive	0	1	0	0.0%
Overall Percentage	0%	97.8%	2.2%	67.4%

4. CONCLUSIONS

Logistic regression analysis was performed using the variables from the mammogram results which are mass, architectural distortion, skin thickening, and calcification. A patient with mass detected on mammogram screening has probability of five times higher in getting breast cancer. Patients with architectural distortion or skin thickening has high probability of being afflicted with breast cancer. Also for patient with calcification detected, the probability of getting breast cancer is 18 times higher. Thus, a patient having any of the symptoms or a combination of these symptoms has greater probability of getting breast cancer. The study can assist radiologists to correctly diagnose breast cancer from using mammograms and referring to the patients' history.

5. ACKNOWLEDGEMENTS

The research was supported by an incentive grant from Universiti Sains Malaysia.

6. REFERENCES

[1]. Al-Ghamdi, A. S. Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention* **34**(6) (2002): 729-741.

[2]. Archer, K. J., S. Lemeshow, and Hosmer, D. W., Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis* **51**(9) (2007): 4450-4464.

[3]. Austin, P. C. and J. V. Tu, Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology* **57**(11) (2004): 1138-1146.

[4]. Bagley, S. C., H. White, and Golomb, B. A.. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology* **54**(10) (2001): 979-985.

[5]. Balleyguier, C., S. Ayadi, K. V. Nguyen, D. Vanel, C. Dromain, and R. Sigal ,BIRADS(TM) classification in mammography. *European Journal of Radiology* **61**(2) (2007): 192-194.

[6]. Colditz, G. A., W. C. Willett, D. J. Hunter,M. J. Stampfer, J. E. Manson, C. H. Hennekens, B. A. Rosner, and F. E. Speizer, Family History, Age, and Risk of Breast Cancer: Prospective Data From the Nurses' Health Study. *Journal of Clinical Medicine* **270**(3) (1993): 338-343.

[7]. Concato, J., A. R. Feinstein, and T. R. Holford, The Risk of Determining Risk with Multivariable Models. *Annals of Internal Medicine* **118**(3) (1993): 201-210.

[8]. Eltoukhy, M. M., I. Faye, and B. B. Samir, Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. *Computerized Medical Imaging and Graphics* **34**(4) (2010): 269-276.
[9]. Feinstein, A. R. *Multivariable analysis: an introduction*. New Haven, CT: Yale University Press,1996.
[10]. Ganz, P. A. Breast cancer, menopause, and long-term survivorship: critical issues for the 21st century. *The American Journal of Medicine* 118(12, Supplement 2) (2005): 136-141.
[11]. Gatta, G., A. Pinto, S. Romano, A. Ancona, M. Scaglione, and L. Volterrani, Clinical, mammographic and ultrasonographic features of blunt breast trauma. *European Journal of Radiology* **59**(3) (2006): 327-330.
[12]. Genuer, R., J. M. Poggi, and C. Tuleau-Malot Variable selection using random forests. *Pattern Recognition Letters* In Press, Corrected Proof, 2010.
[13]. Goodson III, W. H., T. K. Hunt, J. N. Plotnik, and D. H. Moore II Optimization of Clinical Breast Examination. *The American Journal of Medicine* 123(4) (2010): 329-334.
[14]. Gui, G. P. H., R. K. F. Hogben, G. Walsh, R. A'Hern, and R. Eeles, The incidence of breast cancer from screening women according to predicted family history risk: does annual clinical examination add to mammography? *European Journal of Cancer* **37**(13) (2001): 1668-1673.
[15]. Hosmer, D. W. and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley-Interscience Publication, 2000.
[16]. Johnson, J. L., J. L. Bottorff, Balneaves, S. Grewal, R. Bhagat, B. A. Hilton, and H. Clarke ,South Asian womens' views on the causes of breast cancer: images and explanations. *Patient Education and Counseling* **37**(3) (1999): 243-254.
[17]. Kleinbaum, D. G. *Logistic regression: a self-learning text*. New York: Springer-Verlag Telos, 1994.
[18]. McFadden, K. L. Predicting pilot-error incidents of US airline pilots using logistic regression. *Applied Ergonomics* **28**(3) (1997): 209-212.
[19]. Moezzi, M., J. Melamed, E. Vamvakas, G. Inghirami, J. Mitnick, A. Quish, S. Bose, G. Zelman, D. Roses, M. Harris, and H. Feiner, Morphological and biological characteristics of mammogram-detected invasive breast cancer. *Human Pathology* **27**(9) (1996): 944-948.
[20]. Morgan, S. P. and J. D. Teachman, Logistic Regression: Description, Examples, and Comparisons. *Journal of Marriage and the Family* **50**(4) (1988): 929-36.
[21]. Ngah, U. K. A Mammogram and Breast Ultrasound – Based Expert System with Image Processing Features for Breast Cancer, Ph.D Thesis, Universiti Sains Malaysia 2007.
[22]. Paul, P. S. Predictors of work injury in underground mines -- an application of a logistic regression model. *Mining Science and Technology (China)* **19**(3) (2009): 282-289.
[23]. Peduzzi, P., J. Concato, A. R. Feinstein, and T. R. Holford Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology* **48**(12) (1995): 1503-1510.
[24]. Rawal, R., L. Bertelsen, and J. H Olsen, Cancer incidence in first-degree relatives of a population-based set of cases of early-onset breast cancer. *European Journal of Cancer* **42**(17) (2006): 3034-3040.
[25]. Riegler, G., L. Caserta, F. Castiglione, I. Esposito, D. Valpiani, V. Annese, G. Zoli, P. Gionchetti, A. Viscido, G. C. Sturniolo, A. Rispo, F. R. De Filippo, A. de Leone, and R. Carratu, Increased risk of breast cancer in first-degree relatives of Crohn's disease patients: An IG-IBD study. *Digestive and Liver Disease* **38**(1) (2006): 18-23.
[26]. Rodeghie, M. *A practical guide to survey research using SPSS : Survey With Confidents*. Chicago: SPSS Inc 1997
[27]. Roy, S. S. and S. Guria, Diagnostics in logistic regression models. *Journal of the Korean Statistical Society* **37**(2) (2008): 89-94.
[28]. Samanta, B., G. L. Bird, M. Kuijpers, R. A. Zimmerman, G. P. Jarvik, G. Wernovsky, R. R. Clancy, D. J. Licht, J. W. Gaynor, and C. Nataraj, Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. *Artificial Intelligence in Medicine* **46**(3) (2009): 201-215.
[29]. Song, Y., Q. Cai, F. Nie, and C. Zhang, Semi-Supervised Additive Logistic Regression: A Gradient Descent Solution. *Tsinghua Science & Technology* **12**(6) (2007): 638-646.
[30]. Tahir, Z. and M. S. Abu *Analisis Data Berkomputer SPSS 11.5 for Windows*. Kuala Lumpur : Venton Publishing 2003.
[31]. Wiseman, R. A. Breast cancer: critical data analysis concludes that estrogens are not the cause, however lifestyle changes can alter risk rapidly. *Journal of Clinical Epidemiology* **57**(8) (2004): 766-772.
[32]. Zhou, X., K. Y. Liu, and S. T. C. Wong, Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics* **37**(4) (2004): 249-259.