Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining

Md Faisal Kabir
Department of Computer Science
North Dakota State University, NDSU
Fargo, USA
mdfaisal.kabir@ndsu.edu

Simone A. Ludwig
Department of Computer Science
North Dakota State University, NDSU
Fargo, USA
simone.ludwig@ndsu.edu

Abu Saleh Abdullah
Boston Medical Center
Boston University
Boston, USA
abu.abdullah@bmc.org

***Abstract*—Breast cancer is the most common cancer in women worldwide. Prevention of breast cancer through risk factors reduction is a significant concern to decrease its impact on the population. Attaining or detecting significant information in the form of rules is the key to prevent breast cancer. Our objective is to find hidden but important knowledge of the form of rules from the risk factors data set of breast cancer. Mining rules is one of the vital tasks of data mining as rules provide concise statement of potentially important information that is easily understood by end users. In this paper, we use association rule mining, a data mining technique to attain information in the form of rules from breast cancer risk factors data that could be useful to initiate prevention strategies. We discovered rules of both breast cancer and non-breast cancer patients so that we can understand and compare the characteristics of both breast cancer and non-breast cancer individuals. The experimental results show that generated or mined rules hold the highest confidence level.**

***Index Terms*—data mining, association rule mining, breast cancer, risk factors, rules generation.**

I. Introduction

Cancer has become one of the most devastating disease worldwide, with more than 10 million new cases every year, according to World Health Organization (WHO) [1]. The causes and types of cancer vary in different geographical regions, however, nearly every family in the world is touched by cancer. The disease burden is enormous, not only for affected individuals but also for their family and society. Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 [1]. Breast cancer makes up 25 percent of all new cancer diagnoses in women globally, according to the American Cancer Society (ACS) [1]. Prevention of major types of cancer through a quantified assessment of risk is a major concern in order to decrease its impact on our society. Identifying risk factors of breast cancer is important whereby physicians can inform the patients about the potential cancer risks from the risk factors and suggest preventive measures. It is also more important to extract important knowledge from these available risk factors in the form of rules. By using these rules medical professionals or other health related organizations can develop strategies to identify and prevent its impact in the early stage.

Data mining is often referred to as knowledge discovery in databases and describes a process of nontrivial extraction of implicit, previously unknown and potentially useful information from a large volume of data [2]. The mined information is also referred to as knowledge in the form of rules, constraints and regularities. In data mining, rule mining is one of the vital tasks since rules provide concise statements of potentially important information that can be easily understood by end users [3].

The idea of association rules originated from the market basket analysis where a rule is sought to be like "When a customer buys a set of products what is the probability that he or she buys another product? Mathematically, an association rule is defined as $A \Rightarrow B$ where A (antecedent) and B (consequent) are logical predicates constructed by Boolean predicates. A logical predicate in an association rule consists of one or more Boolean conditions and they are connected by the logical AND ($\wedge$) operator. In a transactional data set (e.g., sale database of a supermarket), an association rule appears as $(item = milk) \wedge (item = bread) \Rightarrow (item = butter)$, which means when a customer buys milk and bread it is most likely that he or she also buys butter. The likelihood of an association rule is measured by many values, e.g., support, confidence, lift, and so on.

Association rule mining [4] has been introduced in 1993, and since then it has attracted considerable attention particularly for market basket analysis, where customers' buying patterns are discovered from retail sales transactions.

Discovery of association rules is an important component of data mining. Association Rule Mining (ARM) has been widely used by the retail industry under the name "market-basket analysis". However, the concept of association rules is general and has wide applicability also in the medical domain [5] [6] [7] [8]. In this paper, we demonstrate its applicability to a breast cancer risk factors data set. We investigated to discover hidden but significant rules that could be useful not only for medical professionals but also for health organizations.

This paper is comprised of three important sections following the introduction. The related work is discussed in Section II. The preliminaries including data description, data preprocessing, and the problem statement are described in Section III. The analytical workflow are discussed in Section IV. In this section, the binary logit model and association rule mining is discussed. In Section V, we show our experiments and

results. The outputs obtained from the logit model is discussed and shown. Also, the rule generation using the association rule mining technique is also shown in this section. Moreover, important rules along with their interpretation are shown in this section. Section VI is our discussion section. Section VII is the summary section of this paper where we conclude our paper and suggest possible future research directions.

II. Related Work

Researchers have developed different models for breast cancer risk prediction and association between risk factors [9] [10] [11] [12]. In [9], authors applied statistical methods to show a positive association between Hormone Replacement Therapy (HRT) and breast cancer risk, although this relationship varies according to race/ethnicity, BMI (Body Mass Index), and breast density. The Gali model is used to estimate the number of expected breast cancers for white females who are examined annually [10]. In [11], the authors used commonly identified risk factors such as race/ethnicity, breast density, BMI, and use of hormone therapy, type of menopause, and previous mammographic results to improve the model. In [12], the Breast cancer risk score is determined using a data mining approach called k-nearest-neighbor (KNN) to improve readability for physician and patients. In addition, authors [12] tried to get higher risk detection performances and impact levels of each risk factor.

Association rule mining has been used in the medical domain to find useful information from the data. In [5], authors used the ARM technique for generating the rules for heart disease patients. Based on the rules they discovered the factors which cause heart problems in men and women. In [6], the authors implemented the ARM based concept for finding co-occurrences of diseases carried by a patient using a healthcare repository. The authors extracted data from a patients' healthcare database and from that they generated association rules. Class association rule mining has also been used in the literature to discover the characteristics features [13]. A class association rule set is a subset of association rules with the specified classes as their consequents [14]. In traditional association rule mining, if the support value is kept too low, the class association rule mining will generate overfitting rules for frequent or majority classes; while keeping support value high will not generate sufficient rules for infrequent or minority classes. In class association rule mining this is not the case since mining is done according to the class, the algorithm is not influenced by the unequal distribution between the classes (imbalanced class).

In this paper, we used a risk factors data set from the Breast Cancer Surveillance Consortium (BCSC) [16] to examine significant rules of breast cancer and non-breast cancer patients. Rules of breast cancer patients can be useful for physicians to make informed decision as they have to inform patients about risk factors and alert patients about the potential risks of developing breast cancer (if any). This way, a prevention program or process can be initiated in the early stage of disease progression.

III. Preliminaries

A. Data Description

The data set includes information from $6,318,638$ mammography examinations obtained from the Breast Cancer Surveillance Consortium (BCSC) database collected from January 2000 to December 2009 [16]. Data for this study was obtained from the BCSC Data Resource and more information is available at http://www.bcsc-research.org.

B. Data Pre-processing

The data is aggregated such that the total number of instances or records is $1,144,565$, with 13 attributes or columns. The data set also contains missing or unknown values denoted by 9. To build a reliable model, we discarded the records containing at least one missing or unknown value. We also removed the attribute year that represents the calendar year of the observation. After discarding these records and one attribute, there are $219,524$ available records with 12 attributes. In the data set, there is an attribute named count, representing the number of records that have the combination of variable-values shown in the row. For instance, the value of the count column for the particular row is 12. It indicates that there were 12 similar records; the same as that particular row in the original data. For that reason, we created the number of rows or records the same as the count value in the original data set, and discarded the count column after that. Finally, there are a total of $1,015,583$ records with 11 attributes for building the model. Among $1,015,583$ records, $60,800$ individuals have prior breast cancer, and $954,783$ are non-breast cancer individuals. Among the 11 attributes, "prior breast cancer" values yes or no is considered as response or class variable and the remaining 10 attributes are considered as explanatory or predictors or independent variables. The distribution of all features are shown in Table I through Table X. Bar plots of the age group, age first birth, BMI group, and breast cancer history are shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4 respectively.

TABLE I
Distribution of Race/Ethnicity

Race/Ethnicity	**Count**
Non-Hispanic-White	902736
Asian_or_Pacific Islander	39139
Hispanic	35451
Other_or_Mixed	20972
Non-Hispanic-Black	14389
Native American	2896

TABLE II
Distribution of Hormone Replacement Therapy (HRT)

HRT	**Count**
No	849225
Yes	166358

TABLE III
DISTRIBUTION OF AGE GROUP

Age_group_range	Count
age_55_59	168659
age_50_54	168158
age_45_49	146665
age_60_64	127459
age_40_44	115237
age_65_69	93919
age_70_74	72315
age_75_79	53983
age_80_84	29750
age_35_39	21841
age_greater_equal__85	12557
age_30_34	4113
age_18_29	927

TABLE IV
DISTRIBUTION OF MENOPAUSAL STATUS

Menopaus	Count
Post_menopausal	687566
Pre_or_peri_menopausal	292699
Surgical_menopause	35318

TABLE V
DISTRIBUTION OF BODY MASS INDEX (BMI)

BMI_range	Count
10-to-lessThan_25	430102
25-to-lessThan_30	310555
30-to-lessThan_35	161785
35-or-above+	113141

TABLE VI
DISTRIBUTION OF BI-RADS BREAST DENSITY

BIRADS_breast_density	Count
Scattered_fibroglandular_densities	429488
Heterogeneously_dense	414732
Almost_entirely_fat	90005
Extremly_dense	81358

TABLE VII
DISTRIBUTION OF AGE FIRST BIRTH

Age_first_birth	Count
Age_20_24	331615
Age_25_29	216877
Nulliparous	166180
Age_less_20	157723
Age_greater_equal_30	143188

TABLE VIII
DISTRIBUTION OF FIRST DEGREE RELATIVE

First_degree_relative	Count
No	824472
Yes	191111

TABLE IX
DISTRIBUTION OF PREVIOUS BREAST BIOPSY

biopsy	Count
No	724364
Yes	291219

TABLE X
DISTRIBUTION OF PRIOR BREAST CANCER DIAGNOSIS

breast_cancer_history	Count
No	954783
Yes	60800

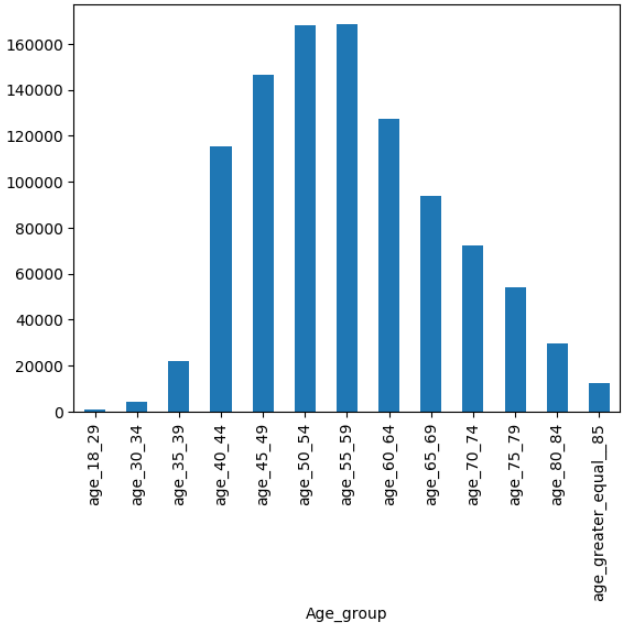

Fig. 1. Bar graph of age group for BCSC risk factors data.

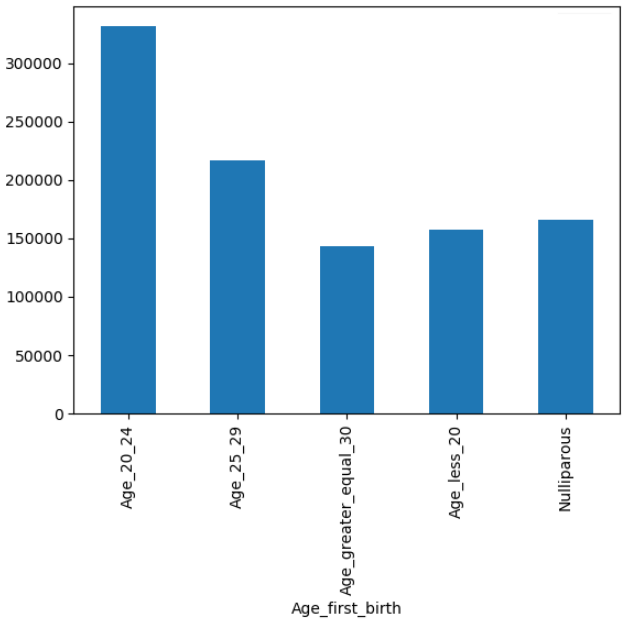

Fig. 2. Bar graph of age first birth for BCSC risk factors data.

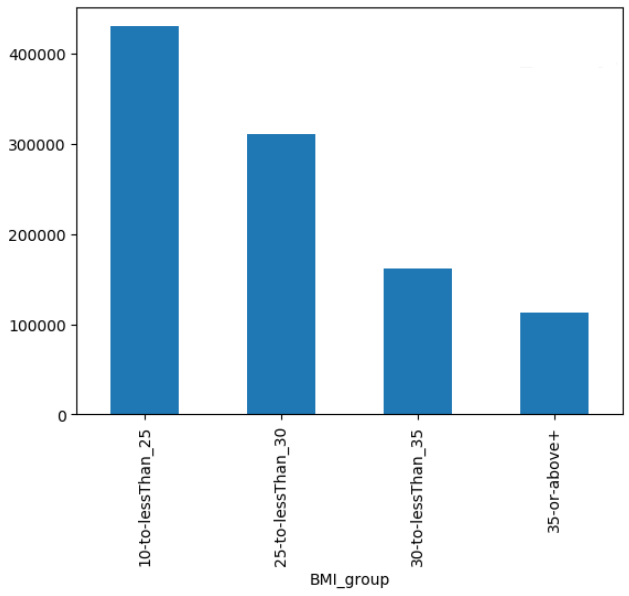

Fig. 3. Bar graph of BMI group for BCSC risk factors data.

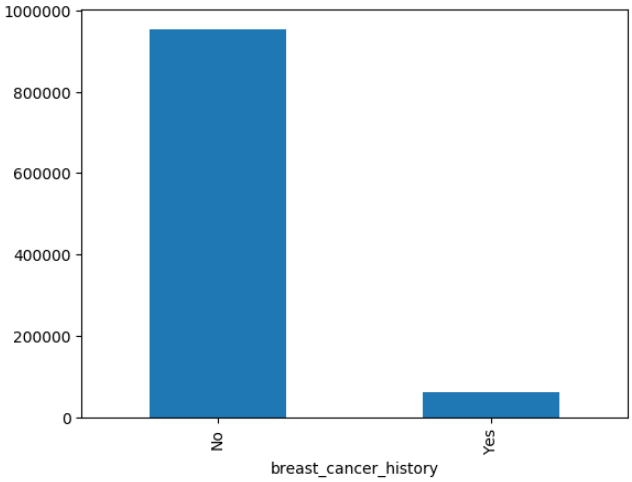

Fig. 4. Bar graph of prior breast cancer for BCSC risk factors data.

C. Conversion of Data Set into Transaction-like Database

For association and class rule mining, the data set has been converted into transactions. For instance, for feature such as race or ethnicity there were a total of six values namely non-Hispanic white, non-Hispanic black, Asian, native American, Hispanic, and mixed/other; for that six columns have been created accordingly with values Yes or No. For example, if an individual is a Native American, then Yes or 1 would be in the corresponding column and the remainder would be No or 0. This way, a total of 46 columns have been created. So, in total there were 1015583 records and 46 items or columns.

D. Problem Statement

Let, P = $\{p_1,p_2,p_3,...,p_n\}$ be the set of n patients and D = $\{d_1,d_2,d_3,...,d_m\}$ be the characteristics of patients, where m is the number of attributes of the patients. We define, C = $\{c_1,c_2\}$ be the class information or the breast cancer history (yes or no) of patients. In this paper, we are interested in finding the relationships among breast cancer risk factors. More specifically, we are interested to find the characteristics or rules in terms of risk factors of both the breast cancer and non-breast cancer individuals (i.e. $\{d_1, d_3, d_6\} \Rightarrow c_1$ and $\{d_2, d_5, d_7\} \Rightarrow c_2$).

IV. Analytical Workflow

In this section, we provide an overview of our framework. First, we used the logit model on the Breast Cancer Surveillance Consortium (BCSC) data set to identify appropriate factors that may affect the likelihood of breast cancer. After that we applied association rule mining and class association rule mining on these risk factors to find significant rules of both non-breast cancer and breast cancer patients.

A. Logit Model

In the current study, the dependent attribute of breast cancer (Yes or 1) or no breast cancer (No or 0) is dichotomous and thus represented as a binary variable. The binary logit model is extensively used in breast cancer investigations where the response variable is binary [15]. The model takes the natural logarithm of the likelihood ratio such that the dependent variable is 1 (breast cancer) as opposed to 0 (no breast cancer). Let, p_1 and p_0 represents the probabilities of the response to variable categories breast cancer and no breast cancer, respectively. The binary logit model is given as:

$$Y = log\left[\frac{p_0}{p_1}\right] = \alpha + \beta_i X_i \tag{1}$$

where Y is the Binary response or class variable; α is the intercept to be calculated; β_i is the estimated vector of parameters, and X_i is the vector of independent variables.

In Equation (1), the maximum likelihood estimation technique is used to estimate the parameters. The unit increase in the independent variables X_i, while keeping all the remaining factors constant, will result in the increase of the likelihood ratio by $exp(\beta_i)$. This states that the relative magnitude by which the response outcome (breast cancer) will increase or decrease, while considering a one-unit increase in the explanatory variable. The probability of breast cancer (p_1) is given by:

$$p_1 = \frac{exp(\alpha + \beta_i X_i)}{1 + exp(\alpha + \beta_i X_i))} \tag{2}$$

Similarly, the probability of no breast cancer (p_0) is given by:

$$p_0 = \frac{1}{1 + exp(\alpha + \beta_i X_i)} \tag{3}$$

We used the logit model to identify and select appropriate factors that may affect the likelihood of breast cancer.

B. Association Rule Mining

Association Rule Mining (ARM) is one of the key techniques to discover and extract useful information from a large data set. Mining association rules [2] can formally be defined as: Let $I = \{i_1, i_2, i_3, ..., i_n\}$, be a set of n binary attributes called items, and Let, $D = \{t_1, t_2, t_3, ..., t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$. The sets of items or item sets X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively. Often rules are restricted to only a single item in the consequent.

Association rules are rules which surpass a user-specified minimum support and minimum confidence threshold. The support $supp\,(X)$ of an item set X is defined as the proportion of transactions in the data set, which contain the item set and confidence of a rule as defined as:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{4}$$

Therefore, an association rule $X \rightarrow Y$ will satisfy $supp(X \cup Y) \geq \phi$ and $conf(X \rightarrow Y) \geq \delta$, which are the minimum support and minimum confidence, respectively. Minimum confidence can be interpreted as the threshold on the estimated conditional probability, the probability of finding the RHS of the rule in the transactions under the condition that these transactions also contain the LHS. Another popular measure for association rules used throughout this paper is lift [17]. The lift of a rule is defined as:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)} \tag{5}$$

It can be interpreted as the deviation of the support of the whole rule from the support expected under independence given the support of both sides of the rule. Greater lift values $(>> 1)$ indicate stronger associations. Measures like support, confidence, and lift are generally called interest measures because they help with focusing on potentially more interesting rules. For example, consider a rule such as $\{milk, sugar\} \Rightarrow \{bread\}$ with support of 0.1, confidence of 0.9, and lift of 2. Now, we know that 10% of all transactions contain all three items together, thus the estimated conditional probability of seeing bread in a transaction under the condition that the transaction also contains milk and sugar is 0.9; and we see the items together in transactions at double the rate we would expect under independence between the item sets milk, sugar and bread [18].

Rules can be generated from data sets having specified classes as their consequences under the name of class association rule mining. These rules have the form $\{A_1, A_2, A_3, ..., A_n \Rightarrow class\}$. The objective here is to focus on using exhaustive search techniques to find all rules with the specified classes as their consequences that satisfy support and confidence [19]. Appropriate values of support and confidence is the key for generating rules since keeping a very low support value will generate large rules and if the support value is too high, we may lose rare but important rules. In this paper, we generated rules from the data set having specified classes such as rules or characteristics of patients who have prior breast cancer. We also generated or mined rules for non-breast cancer individuals. Our goal is to find rules or characteristics rules for these two groups.

V. Experiments and Results

Results of the logit model and association rule mining are discussed in this section. Association rule mining and class association rule mining has been applied on the data set. By selecting the optimum value of support and confidence, we mined strong rules for both breast cancer, and non-breast cancer patients. In this section, we also interpret few strong rules for both groups.

A. Output of Logit Model

The binary logit regression model was used to estimate the coefficients of significant explanatory variables in the final model. The software package SAS was used for the model development. For the model, all attributes were used as input for the likelihood of breast cancer. Interestingly, all explanatory variables turned out to be statistically insignificant $(p < 0.0001)$. Table XI shows the predictor variables which are significant at the corresponding significance levels in the binary logit model, which can contribute to the likelihood of breast cancer.

TABLE XI
Predictor variables with corresponding p values.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	-9.1986	0.0544	28589	<.0001
Age_group	1	0.223	0.00228	9580	<.0001
Race_eth	1	0.0376	0.00463	66	<.0001
First_degree_relative	1	0.1068	0.0109	95	<.0001
Age_menarche	1	0.0259	0.00651	16	<.0001
Age_first_birth	1	0.0729	0.00375	377	<.0001
BIRADS_breast _density	1	-0.1035	0. 00682	230	<.0001
HRT	1	-1.9993	0.0238	7052	<.0001
Menopaus	1	0.4206	0.0132	1009	<.0001
BMI group	1	-0.0164	0.00512	10	0.0014
biopsy	1	5.511	0.0386	20417	<.0001

Positive values of coefficients express that the probability of breast cancer will increase by a certain amount for the specific

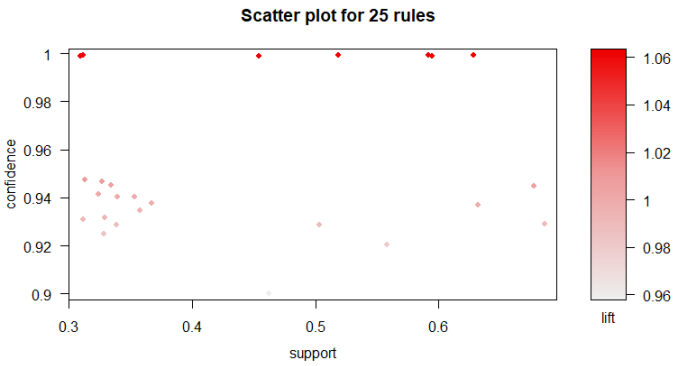

Fig. 5. Scatter plot of 25 rules with minimum support, and confidence of 30% and 80%, respectively.

predictor variables. Interestingly, all explanatory variables are significant at $p < .0001$ except the BMI group which is significant at .0014. From the table it can be referred that age group, race, first degree relatives, age menarche, age first birth, menopause, and biopsy has a positive relationship with previous breast cancer history. However, BIRADS breast density, HRT, and BMI group have negative relationship with breast cancer history.

B. Rules Generation from BCSC Risk Factors Data Set

Our goal is to extract characteristics of patients who have prior breast cancer and who do not have breast cancer. For that, we generated rules using the association rule technique with the specified support and confidence. We defined the consequent of a rule so that we can get our target rules that represent the characteristics of the patients who have breast cancer $(Breast\_cancer\_history = Yes)$ or who do not have breast cancer $(Breast\_cancer\_history = No)$. Support and confidence play an important role in rule generation. Initially, we set the minimum values of support and confidence to 30% and 80%, respectively. Also, we set the minimum length to 3, which means that the generated rules should have at least three items including the consequent. With these specified parameters the algorithm generated 37 rules and after pruning redundant rules we got 25 rules. The scatter plot of these 25 rules are shown in Fig. 5. From these 25 rules, 11 rules whose lift values are greater than or equal to one are shown in Table XII sorted by higher lift value with corresponding support, and confidence. The software R was used for the experiments.

It is worth to mention that we did not obtain any rules of patients who have prior breast cancer $(Breast\_cancer\_history = Yes)$ for the specified support and confidence. This is due to the given values of support, and confidence; also a very small number of instances in which patients have breast cancer compared to their counterpart (ratio is about 1:16).

To obtain the rules of patients having breast cancer we set support to 10% and keep the confidence the same as before (80%). After pruning the redundant rules, we have 165 rules. The scatter plot of these rules is shown in Fig. 6. We still did not obtain any rules having the consequent equals to Yes, which means rules of breast cancer patients.

TABLE XII
RULES GENERATED USING THE ASSOCIATION RULE TECHNIQUE WITH MINIMUM SUPPORT, AND CONFIDENCE VALUE 30% AND 80% RESPECTIVELY.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Race=Non-Hispanic-White, First_degree_relative=No,biopsy=No} =>{breast_cancer_history= No}	52	99	1.06
2	{Age_menarche=Age_12_13, biopsy=No} =>{breast_cancer_history=No}	31	99	1.06
3	{First_degree_relative=No, biopsy=No} =>{breast_cancer_history=No}	59	99	1.06
4	{Race=Non-Hispanic-White, biopsy=No} =>{breast_cancer_history=No}	63	99	1.06
5	{HRT=No, biopsy=No} =>{breast_cancer_history=No}	60	99	1.06
6	{BIRADS_breast_density= scattered_fibroglandular_densities, biopsy=No} =>{breast_cancer_history=No}	31	99	1.06
7	{Menopaus=post_menopausal, biopsy=No} =>{breast_cancer_history=No}	45	99	1.06
8	{First_degree_relative=No, BIRADS_breast_density= Heterogeneously_dense} =>{breast_cancer_history=No}	31	95	1.01
9	{First_degree_relative=No, BMI_group=10-to-lessThan_25} =>{breast_cancer_history=No}	33	95	1.01
10	{First_degree_relative=No, Age_menarche=Age_12_13} =>{breast_cancer_history=No}	33	95	1.01
11	{Race=Non-Hispanic-White, First_degree_relative=No} =>{breast_cancer_history=No}	68	95	1.01

After several experiments, we assigned the value of support to 0.001% but a high confidence value of 90%, and obtained 67 rules. Here, we set the consequent or class value to Yes $(breast\_cancer\_history = Yes)$ so that we can get the rules of breast cancer patients only. The scatter plot of these 67 rules is shown in Fig. 7. And from these 67 rules, the top 10 rules sorted by lift are shown in Table XIII.

C. Generating Strong Rules

We obtained many rules using our methods described earlier. Here, we show a few rules for both breast cancer and non-breast cancer patients that are strong or important as they

TABLE XIII
RULES GENERATED USING ASSOCIATION RULE TECHNIQUE WITH MINIMUM SUPPORT AND CONFIDENCE OF 0.001% AND 90%, RESPECTIVELY AND CONSEQUENT FIXED FOR BREAST CANCER PATIENTS ONLY.

Rules	Supp. (%)	Conf. (%)	Lift
{Age_group=age_greater_equal__85, Race=Hispanic, Age_first_birth=Age_less_20, BI-RADS_breast_density=Almost_entirely_fat, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.7
{Age_group=age_75_79, Race=Non-Hispanic-Black, Age_first_birth=Age_20_24, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_greater_equal_85, Race=Non-Hispanic-White, First_degree_relative=No, Age_first_birth=Nulliparous, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_75_79, First_degree_relative=Yes, Age_first_birth=Nulliparous, BIRADS_breast_density= Almost_entirely_fat,BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_75_79, Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, BI-RADS_breast_density=Heterogeneously_dense, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth=Age_less_20, BIRADS_breast_density= scattered_fibroglandular_densities, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_first_birth=Age_less_20, BIRADS_breast_density= scattered_fibroglandular_densities, HRT=No, BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Hispanic, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth=Nulliparous, BIRADS_breast_density= Heterogeneously_dense, HRT=No, Menopaus=post_menopausal, BMI_group=10-to-lessThan_25, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	95	16.7
{Age_group=age_80_84, First_degree_relative=Yes, Age_first_birth=Age_25_29, BIRADS_breast_density= scattered_fibroglandular_densities, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.002	95	15.66
{Age_group=age_80_84, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth=Age_less_20, BI-RADS_breast_density=scattered_fibroglandular_densities, BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes	0.002	95	15.66

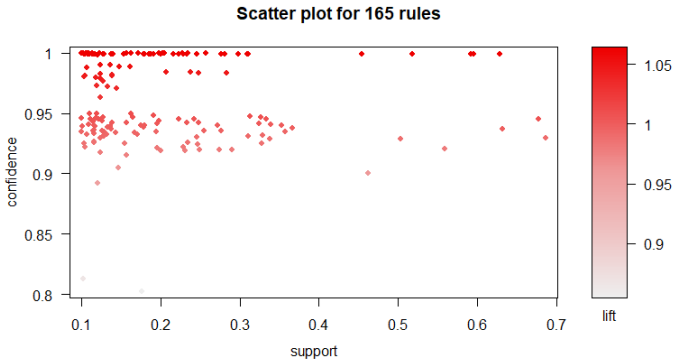

Fig. 6. Scatter plot of 165 rules with minimum support and confidence of 10% and 80%, respectively.

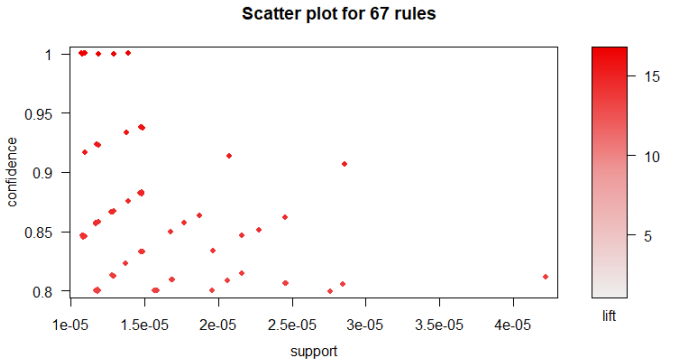

Fig. 7. Scatter plot of 67 rules with minimum support and confidence of 0.001% and 90%, respectively and the consequent is fixed for breast cancer patients only ($breast_cancer_history = Yes$).

have higher confidence and lift values. Strong rules of both non-breast cancer patients and breast cancer patients are shown in Table XIV and Table XV, respectively.

D. Interpreting Strong Rules

Rule 1 of Table XIV can be interpreted as "If a person is a non-Hispanic white with no breast cancer of first degree relatives, and has not had a previous breast biopsy then the individual is a non-breast cancer patient". Rule 4 can be interpreted as "If a person's first-degree relatives do not have breast cancer, and a person's BMI range is between 10 and 25 then the individual is a non-breast cancer patient".

We can interpret Rule 1 of Table XV as "If a patient's race is a Hispanic with age greater or equal to 85 and having had the first birth less than 20 years ago, with a BIRADS breast density being almost entirely fat, and had a previous breast biopsy then the person is a breast cancer patient". Likewise, Rule 2 can be interpreted as "If a person is a non-Hispanic black with an age between 75 and 79 years, the first birth age range between 20 to 24 years, BMI value 35 or above, and had a previous breast biopsy then the individual is a breast cancer patient".

TABLE XIV
STRONG RULES FOR NON-BREAST CANCER PATIENTS WITH CORRESPONDING SUPPORT, CONFIDENCE, AND LIFT VALUES.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Race=Non-Hispanic-White, First_degree_relative=No, biopsy=No} =>{breast_cancer_history=No}	52	99	1.062
2	{Race=Non-Hispanic-White, First_degree_relative=No} =>{breast_cancer_history=No}	68	95	1.005
3	{Age_menarche=Age_12_13, biopsy=No} =>{breast_cancer_history=No}	31	99	1.063
4	{First_degree_relative=No, BMI_group=10-to-lessThan_25} =>{breast_cancer_history=No}	33	95	1.007

TABLE XV
STRONG RULES FOR BREAST CANCER PATIENTS WITH CORRESPONDING SUPPORT, CONFIDENCE, AND LIFT VALUES.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Age_group=age_greater_equal__85, Race=Hispanic, Age_first_birth=Age_less_20, BI-RADS_breast_density=Almost_entirely_fat, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
2	{Age_group=age_75_79, Race=Non-Hispanic-Black, Age_first_birth=Age_20_24, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
3	{Age_group=age_greater_equal__85, Race=Non-Hispanic-White, First_degree_relative=No, Age_first_birth=Nulliparous, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
4	{Age_group=age_75_79, First_degree_relative=Yes, Age_first_birth=Nulliparous, BI-RADS_breast_density=Almost_entirely_fat, BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes}	0.002	100	16.70
5	{Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth=Age_less_20, BIRADS_breast_densit= scattered_fibrogland ular_densities, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.002	100	16.70

E. Interpreting Rules based on Support, Confidence, and Lift

If we consider the rules of both breast cancer and non-breast cancer individuals we can see the significant differences. For both non-breast cancer and breast cancer individuals, its observed confidence, which indicates how often the rule has been found to be true in the data set, is very high (close to 100 %). In case of support, which demonstrates how frequently the item set or factors appear in the data set, it is high (more than 30%) for non-breast cancer patients. However, for breast cancer patients support value is very low (about 0.001%).

For both groups, if we look at the lift value that measures the degree of dependence between the antecedent and the consequent value, we can see the differences. For non-breast cancer individual, lift value is just above 1.0 that means the relationship between factors of these rules (antecedent part) and consequent (non-breast cancer patients) are very low. On the other hand, for the breast cancer patients' lift value is very high (more than 16.0) that indicates a greater association between factors in the antecedent and the consequent (breast cancer patients).

VI. DISCUSSION

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice and clinical service. A reliable prediction will help oncologists and other clinicians in their decision-making process and allow clinicians in choosing the most reliable and evidence-based treatment and prevention strategies for their patients. Although, recent research has looked into various data mining techniques to aid clinicians in the diagnosis of breast cancer, however, there still remain gaps in suggesting an accurate prediction model. Our paper explores association rules for breast cancer and non-breast cancer patients by data mining of the BCSC risk factors data set. Our findings suggest association rules that could be used to predict breast cancer risks among the target population. The data-driven approach that we used in this paper can guide the efficient process of clinical data set to discover behavioral risk factor patterns and reveal hidden information for early detection and initiate prevention efforts as well as treatment strategies of at risk breast cancer patients. However, any prediction should be combined with clinical judgment and individual patient circumstances.

There are several limitations of the current paper. First, we used the BCSC data set which is robust, however, we did not have any control of the overall quality of the data collected. Second, in our data set there are a small number of instances in which patients have breast cancer compared to non-breast cancer patients. In our approach, we specified different support values for both target populations; for breast cancer patients we set a very low support value. In literature [20], we found that researchers used multiple support value for rare item problems and by using a low support value we attained rules of breast cancer patients that are rare in our cases. Although we used a low support value for breast cancer patients, however we set

a high confidence value that represents the predictive strength of the rules.

VII. Conclusion

Extracting useful rules has been generated from a breast cancer risk factor data set using association rule mining. Before applying association rule mining, we used the logit model to check the statistical significance of all predictors. We mined rules for both breast cancer and non-breast cancer patients with specified support and confidence. The experimental results showed that the generated rules hold the highest confidence level for both groups. However, in case of breast cancer patients we have to set a very low support value due to the imbalance of the data (small number of instances of patients having breast cancer compared to non-breast cancer individuals). We also mined strong rules from a huge set of generated rules and interpreted those rules accordingly. This research is an important step in improving risk prediction for people with potential risks for breast cancer.

We intend to extend this research by considering more risk factors to extract more useful and significant rules not only for breast cancer but also other cancer types using the association rule mining algorithm. Furthermore, we plan to build a predictive model using machine learning techniques for the breast cancer data set.

References

[1] J. Ferlay, et al. "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." International journal of cancer 136.5:E359-E386, 2015.

[2] J. Han, M. Kamber. "Data mining concept and technology." Publishing House of Mechanism Industry: 70-72, 2001.

[3] S. M. Monzurur Rahman, Md. F. Kabir, and F. A. Siddiky. "Rules mining from multi-layered neural networks." International Journal of Computational Systems Engineering1.1: 13-24, 2012.

[4] R. Agrawal, T. Imieliski, and A. Swami. "Mining association rules between sets of items in large databases." Acm sigmod record. Vol. 22. No. 2. ACM, 1993.

[13] Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings of the 2001 international conference on data mining. San Jose, CA, US; 2001. p. 36976.

[5] J. Nahar et al. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40.4: 1086-1093, 2013.

[6] K. Majid, and S. T. Tabibi. "Breast mass association rules extraction to detect cancerous masses." Technology, Communication and Knowledge (ICTCK), 2015 International Congress on. IEEE, 2015.

[7] C. Ordonez, C. A. Santana, L. De Braal. "Discovering Interesting Association Rules in Medical Data." ACM SIGMOD workshop on research issues in data mining and knowledge discovery, 2000.

[8] S. Stilou et al. "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare." Studies in health technology and informatics 2: 1399-1403, 2001.

[9] N. Hou, et al. "Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density." Journal of the National Cancer Institute 105.18:1365-1372, 2013.

[10] M. H. Gail et al. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." JNCI: Journal of the National Cancer Institute 81.24: 1879-1886, 1989.

[11] W. E. Barlow, et al. "Prospective breast cancer risk prediction model for women undergoing screening mammography." Journal of the National Cancer Institute 98.17: 1204-1214, 2006.

[12] E. Gauthier et al. "Breast cancer risk score: a data mining approach to improve readability." The International Conference on Data Mining. CSREA Press, 2011.

[14] P. Razan et al. "Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain." Journal of biomedical informatics 48: 73-83, 2014.

[15] A. F. Seddik, D. M. Shawky. "Logistic regression model for breast cancer automatic diagnosis." SAI Intelligent Systems Conference (IntelliSys), 2015. IEEE, 2015.

[16] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C).A list of the BCSC investigators and procedures for requesting BCSC data for research purposes, last retrieved July 2018 from http://www.bcsc-research.org.

[17] S. Brin et al. "Dynamic itemset counting and implication rules for market basket data." Acm Sigmod Record 26.2: 255-264, 1997.

[18] M. Hahsler, B. Grn, and K. Hornik. "Introduction to arulesmining association rules and frequent item sets." SIGKDD Explor 2.4: 1-28, 2007.

[19] P. Razan et al. "Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain." Journal of biomedical informatics 48: 73-83, 2014.

[20] B. Liu, H. Wynne, and M. Yiming. "Mining association rules with multiple minimum supports." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.