Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data

Keerthana Rajendran[1], Manoj Jayabalan[2], Vinesh Thiruchelvam[3]
School of Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia[1, 2, 3]
Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool, UK[2]

***Abstract*—A widespread global health concern among women is the incidence of the second most leading cause of fatality which is breast cancer. Predicting the occurrence of breast cancer based on the risk factors will pave the way to an early diagnosis and an efficient treatment in a quicker time. Although there are many predictive models developed for breast cancer in the past, most of these models are generated from highly imbalanced data. The imbalanced data is usually biased towards the majority class but in cancer diagnosis, it is crucial to diagnose the patients with cancer correctly which are oftentimes the minority class. This study attempts to apply three different class balancing techniques namely oversampling (Synthetic Minority Oversampling Technique (SMOTE)), undersampling (SpreadSubsample) and a hybrid method (SMOTE and SpreadSubsample) on the Breast Cancer Surveillance Consortium (BCSC) dataset before constructing the supervised learning methods. The algorithms employed in this study include Naïve Bayes, Bayesian Network, Random Forest and Decision Tree (C4.5). The balancing method which yields the best performance across all the four classifiers were tested using the validation data to determine the final predictive model. The performances of the classifiers were evaluated using a Receiver Operating Characteristic (ROC) curve, sensitivity, and specificity.**

***Keywords*—Breast cancer; class imbalance; diagnosis; bayesian network**

I. Introduction

The World Health Organization reported in 2018 that there were 627,000 deaths worldwide due to breast cancer [1]. Breast cancer is the second most common cancer death among women, especially in developing countries [2]. This cancer type accounts for 25% of all cancers among women and affects 10% of women globally at some stage of their life [3]. This is a more common issue in developing countries where the mortality rate is greater due to the prohibitive cost incurred for extensive diagnostic tests and treatments required to treat breast cancer completely [4].

American Cancer Society statistics exhibited that there will be about 252,710 new patients with invasive breast cancer and 63,410 patients with in situ breast carcinoma that are expected to be diagnosed among US women in 2017 [5].

The clinicians must check the stage of breast cancer before conducting further assessments on the patients as this step is vital for starting the treatment process and to allow prognosis of the time of recurrence of cancer. However, the multitude of diagnoses carried out to assess the cancer stage require an extended period for the clinicians to obtain medical results. This period of waiting can cause deterioration of cancer where it will be too late for the patients to acquire any complete treatments. Researchers have suggested the involvement of an intelligent decision support system that can identify the cancer types, which will benefit both the patients and clinicians in aspects of treatment options and expenditures incurred [6].

The involvement of data mining techniques in predicting breast cancer based on the patterns and relationships found among the breast cancer risk factors reduce diagnosis time by physicians and cost [7]. Thus, the survival rate for breast cancer can be increased immensely with diagnosis and treatment at an early stage [8].

To predict the susceptibility to breast cancer depends on breast cancer risk parameters [9]. The risk factors of breast cancer include non-preventable factors such as gender, age and family history of cancer, and preventable factors such as body mass index (BMI) and hormone replacement therapy. Other risk factors include menopause, delayed pregnancy, race, radiation therapy before age 30 and high bone density. These abovementioned risk factors are included as part of this study. Genetic risk factors and lifestyle habits (smoking and alcohol consumption) which are also causatives of breast cancer are not included in this study.

As the chances of survival differ largely by breast cancer stages, the earliest diagnosis will improve the rate of survival greatly. Women who were diagnosed at the early, non-invasive stage will have better chances of survival than those diagnosed at the later invasive stages. It is crucial for clinicians to diagnose women who have breast cancer accurately and prevent false positive results [8]. Therefore, the purpose of this study is to develop a predictive model on the breast cancer occurrence for women by applying class balancing techniques on breast cancer risk factors data to cater a fair decision support system for medical practitioners to diagnose the incidence of breast cancer accurately and enhance the survivability rate of patients.

This study on breast cancer classification using BCSC dataset yielded two contributions. Class balancing methods on the imbalanced BCSC dataset were introduced in this study as the problem of class imbalance in BCSC dataset was not addressed in existing studies. Thus, several balancing methods were proposed in this study and the hybrid balancing method achieved greater performance across the proposed classifiers. The breast cancer predictive model developed using Bayesian Network was rarely explored in previous breast cancer studies and in this study, this classifier proved to achieve the highest

accuracy when compared to other works done using BCSC dataset. Thus far, the Bayesian Network model can be used as an effective model to predict the breast cancer occurrence based on the risk factors available in the BCSC data.

II. Related Works

The risk prediction model employing logistic regression on the BSCS dataset, which consisted of 2.4 million records of screening mammograms and breast cancer associated risk factors [10]. The risk models were built on two folds with the menopause status and four risk factors for premenopausal women and ten risk factors for postmenopausal women in second fold were found to be significant in the respective models. Compared to Gail's model that was developed in the late 1980s, this model enhanced the prediction of high-risk women through the addition of two more attributes, which were breast density and hormone replacement therapy [11]. The study reported a ROC of 0.631 for premenopausal women and 0.624 for postmenopausal women.

Another study by [12] was done to identify the factors for disparities in breast cancer outcomes between racial and ethnic groups. The prospective cohort study done showed that African-American women had a higher relative risk of advanced breast tumor compared to white women as African-American women had a less frequent mammographic screening. In [13] developed a breast cancer prediction model that includes breast density as an important risk factor. The researchers used BCSC mammography data where the proportional hazards model was employed to predict the hazard ratios for each BI-RADS breast density category in a 5-year follow-up cohort study. The model was validated using 5-fold cross-validation. The results showed that the average c-statistic was 0.6576 where there was slight discrimination between women who develop breast cancer and those who do not. This risk prediction model can assess 5-year risk for invasive breast cancer depending upon breast density and calibrate with common races and ethnic groups in the United States.

One study employed k-NN algorithm to develop a statistical risk score using four factors such as breast density, age, breast procedure and a number of first degree relatives which were based on the domain expert advice [14]. The area under the ROC was reported as 0.642, which suggested a better model compared to Barlow's logistic regression models [10].

Another study focused on comparative modeling to determine the threshold relative risks at which the harm-benefit ratio of screening women at two different age groups [15]. The authors used four microsimulation models on the film and digital mammography data obtained from BCSC. The results showed that the harm-benefit ratios for women aged 40-49 years with a two-fold elevated risk of breast cancer were similar to that for average-risked women aged 50-74 years for biennial screening mammography. The threshold relative risks were reported to be higher for annual screening using digital mammography, but the harm-benefit ratios were greater for film mammography as they have reduced the false-positive rate.

In [16] used 117,136 diagnostic mammograms pooled from six mammography registries under BCSC to construct logistic regression model to determine the adjusted effect on sensitivity, false positive rates, and cancer detection rates. Patient profile and mammography results were used as determinants in the model. The authors postulated that diagnostic interpretive volume was a crucial factor in considering the thresholds for abnormal diagnostic mammograms. Another study focused on generating an approximation to the logistic regression score function using four different algorithms, namely, ApproxMLE, W.ApproxMLE, WGD and WSGD [17]. These algorithms were applied to BCSC and record linkage datasets. The results showed that ApproxMLE method had excellent performance in aspects of accuracy, time scalability and parallel efficiency. This algorithm had an area under ROC of 0.92 and 3.24 minutes as execution time.

A study by [18] used the BCSC dataset in proposing a method to estimate the rate of missing values due to incomplete data in latent class regression. Two models, one without adjustment of mammography history and the other with the adjustment, were developed. Three approaches, namely maximum likelihood (ML) using the Expectation Maximization (EM) algorithm, multiple imputations (MI), and the proposed method of two-stage MI were compared in terms of the regression coefficient for both the models. The results highlighted that the proposed two-stage MI is better than the EM algorithm and standard MI as it allows for further separation of missing information rates into two parts.

In [19] introduced an adaptive online learning framework which integrated supervised learning (SL) and reinforcement learning (RL) models for clinical breast cancer diagnosis. Three machine learning algorithms such as linear regression, logistic regression, and neural network were employed on BCSC and WBC datasets. This framework had the leverage of gaining high diagnosis accuracy in real-time and reducing the amount of diagnosis required for efficient treatment. Logistic regression was found to achieve optimal performance rapidly. Overall, the SL model was reported to attain accurate risk assessment of breast cancer from incremental features and sequential data while the RL model catered better decision-making of clinical measurements. One study applied association rule mining with feature selection on three breast cancer datasets including BCSC dataset [20]. Syntax and dimension reduction constraints were applied to prune the association rules generated from the apriori algorithm. This resulted in a reduction of the feature subsets by more than 50%. The models were then validated using SVM and the results showed that this approach yielded a classification accuracy of approximately 98% for the BCSC dataset.

Most of the research papers that have published on the predictive model for breast cancer have shown relatively high prediction accuracies [7], [8], [21]. However, a widespread problem in medical data is a class imbalance, which was failed to be addressed by any of these previous papers. In case of breast cancer, most of the previous works done on breast cancer have employed datasets with extremely uneven distribution of the class labels, such as non-cancerous (97%) and cancerous (3%) in the BCSC dataset or survival (91%)

and non-survival (9%) of cancer in the SEER dataset [22] or 66% benign and 36% malignant in the WBC dataset [8]. Thus, the results are likely to be biased towards the majority class, which is non-cancerous or survival or benign group, even if the prediction accuracies were high. Most of the cancer diagnosis need pivotal information on the accuracy and false positive rates in the prediction of the cancerous cases. Development of a prediction model using a class-balanced data will cater to a more affirmative decision-making process during a breast cancer diagnosis.

III. Methodology and Techniques

The methodology deployed involves key processes such as the selection of target data, pre-processing the chosen data, transforming the data into a structured and comprehensible format, balancing the dataset, implementing supervised learning techniques and evaluating the machine learning performance using evaluation measures. These steps ultimately lead to knowledge extraction from the target dataset where new insights and ideas can be developed to assist in enhancing business operations or in this case, aid in early diagnosis and prediction of diseases such as breast cancer.

A. Dataset Selection

The data for this study was obtained from the BCSC Data Resource [23]. The dataset comprises of information on women with breast cancer in the age range of 35 years and above obtained from seven mammography registries of the BCSC-Carolina Mammography Registry, Colorado Mammography Project, Group Health Cooperative's Breast Cancer Surveillance Project, New Hampshire Mammography Network, New Mexico Mammography Project, San Francisco Mammography Registry, and the Vermont Breast Cancer Surveillance System in the United States from 1996 to 2002.

The dataset consists of 280,660 screening mammograms (known as "index mammogram") of the women. The data on the variables of interest were gathered via questionnaires given to women when they were present for their mammogram and through the radiologist who assessed the mammogram results at the screening facility. Besides this, cancer data and pathology registry were merged into the mammography data, thus adding on to the related variables of breast cancer [10]. The dataset was anonymized to preserve the confidentiality of the patients, the mammography registries and radiology facilities. Other identifiers such as the origin of the data, dates, patient identifiers and the index screening mammogram assessment results are not included in the data, refer Table I.

TABLE I. Dataset Summary

No.	Variable Name	Coded values
1.	menopaus	Indicates the stage of menopause of each patient. 0 = Premenopausal; 1 = Postmenopausal or age is more than 55; 9 = Missing or unknown
2.	agegrp	Indicates the age group (in years) that the patient belongs to. Kindly note that the code value 9 in this variable represents an age category, instead of a missing value. There is no missing value found in this variable. 1 = 35-39; 2 = 40-44; 3 = 45-49; 4 = 50-54; 5 = 55-59; 6 = 60-64; 7 = 65-69; 8 = 70-74; 9 = 75-79; 10 = 80-84
3.	density	Indicates the patient's breast density based on the findings from Breast Imaging Reporting and Data System (BI-RADS) mammogram screening. 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense; 9 = Unknown or different measurement system
4.	race	Indicates the ethnic background or race of the patient. 1 = White; 2 = Asian/Pacific Islander; 3 = Black; 4 = Native American; 5 = Other/mixed; 9 = Missing or unknown
5.	Hispanic	Indicates whether the patient has Hispanic background. 0 = No; 1 = Yes; 9 = Missing or unknown
6.	bmi	Indicates the Body Mass Index (BMI) of each patient under study. 1 = 10-24.99; 2 = 25-29.99; 3 = 30-34.99; 4 = 35 or more; 9 = Missing or unknown
7.	agefirst	Indicates the age of the patient when she had her first birth. 0 = Age less than 30; 1 = Age 30 or greater; 2 = Nulliparous (has not borne an offspring); 9 = Missing or unknown
8.	nrelbc	Indicates the number of first degree relatives of the patient with breast cancer. 0 = Zero; 1= One; 2 = Two or more; 9 = Missing or unknown
9.	brstproc	Indicates the presence of any previous breast procedure on the patient. 0 = No; 1 = Yes; 9 = Missing or unknown
10.	lastmamm	Indicates the outcome of the patient's last mammogram before the index mammogram. 0 = Negative; 1 = False positive; 9 = Missing or unknown
11.	surgmeno	Indicates whether the patient underwent surgical or natural menopause. 0 = Natural; 1 = Surgical; 9 = Missing or unknown or not menopausal (menopaus = 0 or menopaus = 9)
12.	hrt	Indicates whether the patient has undergone any current hormone replacement therapy. 0 = No; 1 = Yes; 9 = Missing or unknown or not menopausal (menopaus = 0 or menopaus = 9)
13.	invasive	Indicates the results for the diagnosis of invasive breast cancer of the patient within one year of the index screening mammogram. 0 = No; 1 = Yes
14.	cancer	Indicates the results for the diagnosis of invasive or ductal carcinoma in situ breast cancer of the patient within one year of the index screening mammogram. 0 = No; 1 = Yes

B. Data Pre-Processing and Transformation

In the stage of pre-processing, it is essential to eliminate any missing values, noise and other anomalies in the selected data. Any inconsistency in the chosen data, especially disease-related data may lead to unreliable results or misdiagnosis of test data, which could be fatal if the model is implemented in real-life situations [24]. One of the steps of pre-processing is the elimination of unrelated variables, as these variables are not required to meet the goal of the study. Besides that, missing values or anomalies occur due to lack of information and unprecise measurement values leading to inadequate accuracy and a greater percentage of error in the process of data evaluation. For handling missing values which is very common in cancer datasets, imputation have to perform before implementing the model [25]. The missing values for the nominal and numerical attributes in the dataset with the modes and means from the training data. Since all the variables were identified as the nominal (categorical) type, modes, which are values with the highest frequency, from the training data was used to impute the missing values.

For further processing, the data must be transformed into an appropriate format that is readable and compatible with the data mining techniques employed on the dataset [26], [27]. The transformation such as numerical to nominal conversion is done to cater to the requirements of distinct types of data mining techniques.

C. Class Balancing

The imbalance is a problem that is very commonly found in disease-related datasets, such as the breast cancer dataset used in this study, where the class with a greater number of instances is known as the majority class whereas the one with comparatively less number of instances is known as the minority class. In a scenario where the imbalanced dataset is used, the classifiers tend to favor the majority class, thus exhibiting very weak classification rates on the minority class. There is also a possibility that the classifiers predict all as the majority class and disregard the minority class. This is a very common scenario in medical datasets where the patient with the disease tends to be the minority class. Therefore, a good sampling technique is required for medical datasets. To solve the problem of class imbalance, various sampling techniques have been introduced which include undersampling, oversampling and a combination of both. Sampling strategies are introduced to overcome the class imbalance issue through the removal of some data from the majority class (undersampling) or the addition of some artificially synthesized or replicated data to the minority class (oversampling) [28]–[30].

To build a good prediction model from the training set, the data must be well-balanced. But, the class labels of the target variable, cancer in the breast cancer dataset used in this study are not balanced. This may result in a mediocre performance of the classifiers on the minority class label, which is the Yes label, especially when the data is extremely imbalanced with 97.2% of No and 2.8% of Yes. The key reason behind this is because the classifiers neglect the relative distribution of each class, but they tend to focus on optimizing the overall precision [28].

Oversampling methods multiply the number of members in the minority class in the training group. A benefit of oversampling is that there is no loss of information from the original training dataset as all the observations from the majority and minority classes are retained. The disadvantage of this technique is that it may take longer training time and result in over-fitting since there is a significant increase in the size of the training set. A well-known oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE), is used to oversample the minority class by creating synthetic instances to replicate the minority classes and increase their number of instances in the training set [31]. These synthetic instances are produced by considering two key parameters which are the number of instances (n) and the nearest neighbors (k). As the new minority instances are generated by interpolating between several minority instances that lie together, the problem of overfitting in prevented [32].

On the other hand, undersampling also overcomes the class imbalance problem wherein this technique, the number of samples in the majority class is decreased to balance the class distribution between the minority and majority classes. As the size of the training dataset is reduced significantly, the training time taken is lesser and more efficient which serves as an advantage of this technique. A disadvantage of this method is that important information in the training data will be lost. As for undersampling, SpreadSubsample can be used to decrease the number of samples in the majority class from the original dataset so that the class distribution can be balanced with the minority class. The distribution spread can be set as 0, 1 or 10 for different distribution spread of the class values. A value of 1 spreads the class into a uniform distribution, where the class labels are balanced equally.

A combination of oversampling and undersampling in some cases would be a better option as it generates better-defined areas in the data space and avoids over-generalization [29].

The training dataset which will be used in further analysis has the problem of class imbalance with regards to the target variable, cancer. In the imbalanced dataset, the values are more biased towards the value No than Yes. The distribution of No and Yes is largely uneven in the dataset. The dataset which has 180,465 observations consists of 175,339 No values (97.2%) and 5,126 Yes values (2.8%). This major difference between the values in the class variable could lead to the results to favor towards the majority value which is No. This impacts the efficiency of the results negatively which will, in turn, reflect uncertainty in the choice of an ideal prediction model developed from a machine learning algorithm. The training dataset without any class balancing is kept as it is to perform classification tasks on the data.

As one of the methods of class balancing, SMOTE is applied on the training dataset to oversample the minority class label and a separate dataset is created with this oversampling method. The dataset was resampled where the minority class value (which is Yes) is oversampled to increase its number of instances. The default parameter set in WEKA were used which include the creation of 100% SMOTE instances and the nearest neighbor of 5. Upon execution of the

SMOTE filter, the minority class value (which is Yes) has increased to 10,252 instances while the majority remains the same. Now, the target variable, cancer in the class-balanced training set distribution consisted of 10,252 Yes values (5.5%) and 175,339 No values (94.5%). The percentage of the minority class value has doubled via this method of class balancing. The total number of instances in the dataset was observed to be 185,591 instances.

As the second method of class balancing, the technique of undersampling was applied on the training dataset and a training dataset with this method was created. This was done using the SpreadSubsample method where the class distribution spread is set as 1.0 to allow uniform distribution between the two class values (Yes and No). As a result, the majority class value was undersampled to match the minority class value. Upon applying the SpreadSubsample filter, the number of instances in the majority class value (which is No) has reduced to 5,126 which is the same as in the minority class value (which is Yes). Both the class values, Yes and No, have an equal percentage of distribution, which is 5,126 instances (50%) respectively, for the target variable, cancer in the training dataset. The total number of instances in the dataset was observed to be 10,252 instances.

Following this, both the oversampling and undersampling techniques were combined to resample the imbalanced dataset and a training dataset with this method was created. The oversampling method (SMOTE) was applied first followed by the undersampling method (SpreadSubsample) to resample the distribution of the class values in the target variable, cancer. This resulted in the minority class which is the Yes value to be oversampled first, then the majority class which is the No value to be undersampled. The same parameters applied in the previous two methods of class balancing, which are SMOTE and SpreadSubsample, were applied here to obtain a uniform result when comparing these three techniques of balancing. The number of instances in both class values were equal with 10,252 instances (50%) in each of the class value, No and Yes. The total number of instances in the dataset was observed to be 20,504 instances.

D. Data Mining Techniques

Data mining plays as a powerful tool in acquiring valuable information from a large volume of transformed data to aid in quicker decision making and discovery of knowledge. Data mining techniques enable the identification of novel and hidden patterns from the data, facilitate the data experts in uncovering relationships among the data and make statistically-proven and informed decision. Employment of data mining techniques in medical diagnosis such as breast cancer prediction is of utmost importance as it allows the clinicians to make a quick decision on the effective treatment method, early detection, and prediction of cancer and other diseases which in turn improves the survival rate of patients and reduces the cost of treatment. There are various data mining techniques such as classification, clustering, association and regression which are commonly used in medical diagnosis and disease prediction [33].

Classification is a supervised learning technique which is employed to automatically generate a model that can classify or group a class of items, thus the unknown class values of future objects can be predicted. In this two-step process, a training step and validation step are involved to classify new objects. In the first step, the training dataset is used to construct a model to elucidate the characteristics of a group of data classes or notions. As the data classes or notions are predefined where the class which the training sample falls into is given, this process is known as supervised learning. In the second step, the model is implemented to predict the classes of future data or objects [34].

It is a popular technique in studies on cancer prediction and early diagnosis. Some of the classification algorithms which have been employed in previous studies on breast cancer prediction include Naïve Bayes, Bayesian Network, decision tree and association-based classification. In classification, the data is partitioned into two groups which are the training set and testing set. In the training phase, a model is constructed by employing the classifier on the training data and the performance of this model is validated by using the model to predict or assign a class label to the test data which is unlabeled.

1) Bayesian network: Bayesian Network is represented in a graphical model to portray the probabilistic relationships among the variables under study. The Bayesian model assumes conditional independence over the various random variables and this assumption gives information on the probability distribution that is illustrated within the network [35].

Overall, this Bayesian Network is made up of a qualitative element (structural model) that caters a visual depiction of the interactions among the variables, and a quantitative element (a group of local probability distributions) which allows probabilistic inference and mathematically measures the significance of a variable or a group of variables on others. These qualitative and quantitative components establish a singular joint probability distribution on the variables for a problem [36]. From a Bayesian point of view, the classification problem can be described as the challenge of identifying the class with maximum probability given that there is a set of observed variable values. Such probability is viewed as the posterior probability of a class by considering the given set of data and is computed based on the foundation of Bayesian theorem.

This classifier requires a very large training dataset to significantly analyze all the likely combinations and eventually estimate the probability distribution from the training set. This task could be arduous which serves as a disadvantage of this data mining technique [26]. One of the greatest advantages of Bayesian Network is that it permits the compact and economical representation of the joint probability distribution by using conditional independence extensively. The Bayesian Network is preferred as past academic works have shown that this classifier exhibits a strong correlation among the attributes in the patient disease diagnosis. Other than that, the classifier is robust to unrelated variables, noise and confounding factors that are not part of the classification [31]. Bayesian Network has been broadly employed in many medical diagnoses based on previous literature studies,

especially for cancer prediction and recently the use of Bayesian Network classifiers in breast cancer prediction is trending. It is a well-known classifier in medical diagnosis in case of the non-deterministic relationship between the class variable and the attribute set. One of the learning algorithms applied to the Bayesian Network known as K2 has been utilized in breast cancer classification due to its rapid convergence ability.

The structure of the Bayesian Network from the data is learned using search algorithms. Among the several types of learning algorithms such as AD (All Dimensions) Trees, TAN (Tree Augmented Naïve Bayes) and others, K2 is a very well-known algorithm used in cancer classifications which employs a heuristic greedy search method [35]. The K2 algorithm searches the space across all the potential acyclic digraphs by producing many distinct graphs in a heuristic way and based on this, their ability to interpret the data is compared. The apriori ordering of the variables limits the search space by only permitting parents that precede the variables in the ordering. The algorithm begins its search by assuming that each node has an empty set of parents and iteratively adds the parent based on a given ordering that the addition increases the probability (score) of the final structure the most. The algorithm stops the addition of the variables to a parental set when further addition of a parent does not increase the probability and carries on to the next variable present in the ordering. The model building process involves iterative permutations of the ordering and the network that gives the highest probability is selected. Once the structure has been learned, the conditional probabilities of the Bayesian Network are estimated directly from the data using a Simple Estimator method.

2) Random forest: Random Forest is a tree-based method where it creates multiple classifiers and aggregates the outcomes using ensemble learning method to make the predictions. The approach used in such Ensembles of Classifiers is that there is a level of randomness to generate their tree-based components. This technique creates a collection of hundreds to thousands of unpruned classification and regression trees (CART) based on the random selection of records in the original training data. Although Random Forest is derived from the CART technique, it differs from CART based on the non-deterministic growth via a two-level randomization process. Each tree is grown using the bootstrap sample of the training data and explores across a randomly chosen subset of features (input variables) to determine the split at the node level during the tree growth. The random selection of the features reduces the correlation between the trees which enhances the prediction power and gives higher efficiency. The low variance of the forest ensemble is known as the bagging phenomenon [37]. The splitting criterion in the Random Forest technique is based on the Gini measure of impurity where the lowest impurity value is computed at each node for a set of variables.

One of the key features of Random Forest in classification is that it provides the measure of variable importance where it shows the degree of association between a particular feature and the classification result. To test the trees developed from the bootstrap data, the out-of-bag samples can be used to provide the two by-products which are the unbiased test set error estimate and variable importance measure. Due to the many benefits offered by Random Forest, this method is very popular and preferred for classification tasks such as in breast cancer prediction studies. The advantages of this technique are given as the following:

- It can handle high dimensional data which contains missing values and variables which are continuous, binary and categorical.
- It is robust enough to overcome over-fitting of data, thus does not require pruning of the trees.
- It is a simple, efficient and comprehensible non-parametric method that can be employed on diverse types of datasets.
- It has greater prediction accuracy and better generalization.

3) Decision tree: A decision tree is a supervised technique which applies the reasoning approach to obtain solutions for a given problem. This data mining technique is very flexible and simple which makes it an attractive choice for applications in diverse fields, particularly because it exhibits advice-oriented visualization to make the prediction decision based on the observed outcomes. A decision tree is commonly applied in decision-making processes in the medical field for disease diagnosis such as cancer prediction. The tree-shaped structures in decision tree represent decision sets which are easy to interpret and understand for decision-makers to assess and choose the best course of action based on the risk and benefits for each possible outcome for distinct options [33]. The basic structure of a decision tree is composed of the following elements.

- A root node which does not have any incoming branch but consists of zero or more outgoing branches.
- Internal nodes where each node has one incoming branch and two or more outgoing branches.
- Leaf or terminal nodes, each of which comprises of one incoming branch but no outgoing branches.

Each node represents the attribute in the input attribute space, while each branch in the decision tree represents a condition value for the corresponding node. The non-terminal nodes have attribute test conditions to divide the records based on their distinguishing characteristics which are represented by the branches.

One of the popular classification types in the decision tree, especially in breast cancer prediction and diagnosis, is the C4.5 algorithm which is an extension of the ID3 algorithm [38]. C4.5 generates decision trees from a group of defined training data based on the information entropy concept. This approach utilizes the fact that each variable in the data is involved in decision-making by splitting the records into smaller subsets. Using the normalized information gain

(difference in entropy), C4.5 determines the selection process of an attribute to split the data. The attribute with the greatest normalized information gain is chosen as the decision node. The branch with zero entropy is taken as the leaf node in the decision tree. This algorithm runs recursively on smaller subsets which are non-leaf nodes with non-zero entropy. The splitting process stops when all the samples in a given subset or node fall under the same class. Then a lead node is generated so that the class can be chosen. But in case of lack of information gain from any of the attributes, the C4.5 generates a decision node using the class expected value from the nodes higher up in the tree.

The C4.5 algorithm has a few advantages such as it is easy and simple to construct in a comprehensible format, can be applied on data with discrete and continuous attributes, can handle attributes with missing values and differing costs in the training data and has greater precision due to pruning procedure. The disadvantages of C4.5 classifier include the expensive cost incurred and high computational time.

IV. Experimentation and Results

The experiment was carried out using the Waikato Environment for Knowledge Analysis (WEKA) [15]. Model validation using k-fold (10 folds) cross-validation was applied on the class-imbalanced training data as well as the three class-balanced training sets respectively. The generated classifier model in this study using the training dataset was validated using this 10-fold cross-validation method which is preferred in disease-related analysis, including breast cancer diagnosis and prediction.

The class balancing methods were applied to the training dataset which consists of 180,465 instances. These balancing methods were applied to overcome the issue of the class imbalance of the target variable, cancer. The balancing methods that were applied include SMOTE (for oversampling), SpreadSubsample (for undersampling) distribution spread was set as 1.0 and a combination of SMOTE and SpreadSubsample.

Each of the classifier performance was compared based on their sampling methods and the classifier with best overall performance was chosen as the best prediction model for the breast cancer dataset. The classifiers' performances were assessed based on several evaluation metrics which include the correctly classified instances percentage or the accuracy, ROC, PRC Area, FP Rate, specificity, precision, recall and F-measure. As the breast cancer dataset used in this study is medical data, there are certain evaluation measures which are of key importance in evaluating the prediction model developed from an algorithm. These measures are the accuracy, TP rate (or sensitivity or recall), FP rate, precision, ROC and PRC area.

For an ideal breast cancer prediction model, a greater TP rate indicates that cancer patients are predicted correctly to have cancer. A higher TN rate is also preferred but this measure does not carry as much importance as TP rate. A prediction model needs to detect the presence of a disease correctly and prevent any misdiagnosis. The misleading results due to FN rate and FP rate can be fatal to patients as cancer is a lethal disease and the earlier the diagnosis, the better are the chances of survival. The lesser the FP rate and FN rate, and the higher the TP rate and TN rate, the better is the performance of the classification model. This is some general criteria for a disease prediction model, but this may vary depending on the dataset and the type of classifiers.

Based on the analysis, it was found that across all the four classifiers there were two sampling methods which showed better measures to evaluate the performance of the classifiers. These methods are without balancing, where the original training set is used, and applying a combination of SMOTE and SpreadSubsample on the training set.

It was observed that within each classifier, without balancing and combination of SMOTE and SpreadSubsample methods show better evaluation for certain measures. In Bayesian Network, Random Forest and Decision Tree C4.5 models, the accuracy, sensitivity, and precision values are greater in without balancing method, but the ROC is higher in the combination method. As a lower FP rate is better for disease diagnosis, the combination (hybrid) method shows a lower FP rate for all the four abovementioned classifiers.

Overall, it can be concluded that although all these four classifiers do perform well without class balancing method, the results produced are likely to be biased as the distribution of the class value is imbalanced. The accuracy, sensitivity, and specificity will be biased towards the majority class value, which in this case is the No cancer class value. Thus, the combination of SMOTE and SpreadSubsample method is a better model that can be used to validate these four classifiers using a validation (test) dataset. This hybrid balancing method has an even spread of the Yes cancer and No cancer class value which will aid in the development of a fair predictive model.

Further performance measures used to evaluate the classifiers were analyzed to determine the optimum classification model for the prediction of breast cancer using the BCSC dataset. The validation dataset which was used to validate the classifiers generated on the training set resulted in the construction of classifiers with similar values of evaluation metrics and there is no enormous difference between the classifiers from the training and validation set. This shows that all the classifiers have performed well upon the evaluation using the test set. But, to determine the best classifier or the most robust one among the four proposed classifiers, some of the common evaluation metrics found in medical diagnosis were compared between these classifiers.

Table II shows Bayesian Network classifiers have the highest accuracy with 99.1%. Random Forest yielded an accuracy of 94.8% which is the lowest. For the Yes class label, Bayesian Network have the lowest FP rate where it was shown that there were 0% of FP that was predicted. It is important to obtain an FP rate as low as possible to avoid the mistake of diagnosing healthy patients as having breast cancer. Bayesian Network classifier portrayed the greatest precision values for both the class labels and weighted average. The average sensitivity for Bayesian Network is given as 99.1% and the ROC is given as 93.7%, where these two measures are the highest across all the classifiers. Overall,

the results show that Bayesian Network can be adopted as the predictive model for this breast cancer study using BCSC data.

The rationale behind selecting Bayesian Network as the best classification model for this study is because this data mining technique have been commonly employed in the diagnosis of cancer and thus, has an evident record of working well as a prediction model for cancer studies. Further, the Bayesian network model is more comprehensible for the human brain as the model can be easily visualized using a graph.

To further prove that the Bayesian Network model yields a better prediction compared to other models employed in previous literature on the same BCSC dataset, the Table III the comparison of the evaluation measures between the models in the previous literature and this study. Bayesian Network model has yielded the highest accuracy and ROC compared to the other models. Thus, the Bayesian Network acts as a better predictive model in the classification of breast cancer occurrence based on the related risk factors. Besides that one major difference with the previous studies were the use of class balancing techniques. None of these studies addressed the class imbalance issue on the BCSC and employed several balancing techniques, which were done in this study. The approach of hybrid balancing technique with Bayesian Network produced a better prediction model, as was shown in this study.

TABLE II. PERFORMANCE EVALUATION METRICS

Classifier	Class label	Performance evaluation metrics				
		Accuracy	FP Rate	Precision	Sensitivity or Recall	ROC
Naïve Bayes	No	0.991	0.219	0.991	1.000	0.937
	Yes		0.000	1.000	0.781	0.937
	Weighted average		0.210	0.991	0.991	0.937
Bayesian Network	No	0.991	0.219	0.991	1.000	0.937
	Yes		0.000	1.000	0.781	0.937
	Weighted average		0.210	0.991	0.991	0.937
Random Forest	No	0.948	0.197	0.991	0.955	0.913
	Yes		0.045	0.434	0.803	0.913
	Weighted average		0.191	0.968	0.948	0.913
Decision Tree C4.5	No	0.984	0.214	0.991	0.993	0.914
	Yes		0.007	0.832	0.786	0.914
	Weighted average		0.206	0.984	0.984	0.914

TABLE III. COMPARISON WITH EXISTING STUDIES

Previous literature	Predictive model	Evaluation measure	Scope of study
[10]	Logistic regression	ROC = 0.631 (premenopausal); 0.624 (postmenopausal)	Risk prediction model for premenopausal and postmenopausal women
[13]	Proportional hazard models	c-statistic = 0.6576	5-year risk prediction model for invasive breast cancer based on breast density
[14]	k-NN	ROC = 0.642	Statistical risk score using four risk factors
[17]	ApproxMLE algorithm	ROC = 0.92	Approximation to logistic regression score function
[20]	Association rule mining with SVM	Accuracy = 98%	An association rule model with feature selection on the dataset
This study	Bayesian Network	Accuracy = 99.1% ROC = 0.937	Predictive model for breast cancer occurrence depending on risk factors

V. CONCLUSIONS

This study was conducted using the BCSC dataset which consisted of 280,660 screening mammography results and demographic profiles of breast cancer patients who are women aged 35 years and above. The issue of class imbalance in the training dataset was solved using three-class balancing techniques, namely, SMOTE, SpreadSubsample and hybrid of SMOTE and SpreadSubsample. These methods were used to construct the Bayesian Network, Random Forest and Decision Tree C4.5 classification models. When the sampling techniques were compared across each classifier using the performance evaluation metrics, the results showed that the classifiers generated using the hybrid balancing method had

the best performance in terms of false positive rate and area under the ROC. Thus, the best-suited class balancing method for the BCSC dataset was determined to be a hybrid method which was statistically proven to perform well over other sampling techniques.

The results showed that the Bayesian Network generated from the class balanced BCSC data using the hybrid method had greater overall performance in terms of ROC (0.937), sensitivity (78.1%), and False Positive rate (0%) or specificity (100%). This study proves that the Bayesian Network model can serve as a better decision support system for physicians, and as means for early diagnosis and treatment for patients by predicting the occurrence of breast cancer based on the risk factors.

In conclusion, this study proved that the hybrid balancing method with Bayesian Network algorithm achieved the greatest efficiency in predicting the breast cancer occurrence based on the risk factors. With this approach, clinicians can make fair and statistically-proven decisions on diagnosis and treatment options, while breast cancer patients can gain a better understanding of the disease and its risk factors.

Further works can involve feature selection on the BCSC dataset and segmentation of the variables with similar characteristics. A predictive model constructed from feature selection and similar variables may produce a generalized model with a minimum number of risk factors to diagnose. As it is not guaranteed that conclusions from this study could be generalized to other mammography datasets with different properties, it would also be interesting to apply this methodology on other data with features such as shape, location, tumor size or radiation intensity.

ACKNOWLEDGMENT

We are grateful to the Breast Cancer Surveillance Consortium for making the dataset accessible for the research. 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 are provided at: http://breastscreening. cancer.gov/.

FUNDING

This work was supported by the Asia Pacific University of Technology & Innovation (APU) under Internal Grant FCET/10/2018.

REFERENCES

[1] World Health Organisation, “Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018,” Int. Agency Res. cancer, no. September, pp. 13–15, 2018.

[2] E. Kharazmi, A. Försti, K. Sundquist, and K. Hemminki, “Survival in familial and non-familial breast cancer by age and stage at diagnosis,” Eur. J. Cancer, vol. 52, pp. 10–18, Jan. 2016, doi: 10.1016/J.EJCA.2015.09.015.

[3] S. Kharya, “Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease,” Int. J. Comput. Sci. Eng. Inf. Technol., vol. 2, no. 2, pp. 55–66, 2012, doi: 10.5121/ijcseit.2012.2206.

[4] D. Carvalho, P. R. Pinheiro, and M. C. D. Pinheiro, “A Hybrid Model to Support the Early Diagnosis of Breast Cancer,” Procedia Comput. Sci., vol. 91, no. Itqm, pp. 927–934, 2016, doi: 10.1016/j.procs.2016.07.112.

[5] C. E. DeSantis, J. Ma, A. Goding Sauer, L. A. Newman, and A. Jemal, “Breast cancer statistics, 2017, racial disparity in mortality by state,” CA. Cancer J. Clin., vol. 67, no. 6, pp. 439–448, 2017, doi: 10.3322/caac.21412.

[6] R. J. Oskouei, N. M. Kor, and S. A. Maleki, “Data mining and medical world: Breast cancers’ diagnosis, treatment, prognosis and challenges,” Am. J. Cancer Res., vol. 7, no. 3, pp. 610–627, 2017.

[7] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, “Data Mining Techniques For Diagnosis And Prognosis Of Cancer,” Ijarcce, vol. 4, no. 3, pp. 613–615, 2015, doi: 10.17148/IJARCCE.2015.43147.

[8] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” Int. J. Comput. Appl., vol. 98, no. 10, pp. 16–24, Jul. 2014, doi: 10.5120/17219-7456.

[9] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” PLoS One, vol. 12, no. 1, pp. 1–14, 2017, doi: 10.1371/journal.pone.0161501.

[10] W. E. Barlow et al., “Prospective breast cancer risk prediction model for women undergoing screening mammography,” J. Natl. Cancer Inst., vol. 98, no. 17, pp. 1204–1214, 2006, doi: 10.1093/jnci/djj331.

[11] M. H. Gail et al., “Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually,” JNCI J. Natl. Cancer Inst., vol. 81, no. 24, pp. 1879–1886, Dec. 1989, doi: 10.1093/jnci/81.24.1879.

[12] R. Smith-Bindman et al., “Does Utilization of Screening Mammography Explain Racial and Ethnic Differences in Breast Cancer?Mammography Use and Breast Cancer Rates in Ethnic Minorities,” Ann. Intern. Med., vol. 144, no. 8, pp. 541–553, Apr. 2006, doi: 10.7326/0003-4819-144-8-200604180-00004.

[13] J. A. Tice, S. R. Cummings, R. Smith-Bindman, L. Ichikawa, W. E. Barlow, and K. Kerlikowske, “Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model,” Ann. Intern. Med., vol. 148, no. 5, pp. 337–347, 2008, doi: 10.7326/0003-4819-148-5-200803040-00004.

[14] E. Gauthier, L. Brisson, P. Lenka, and S. Ragusa, “Breast cancer risk score: a data mining approach to improve readability.,” Int. Conf. Data Min., pp. 15–21, 2011.

[15] N. T. van Ravesteyn et al., “Tipping the Balance of Benefits and Harms to Favor Screening Mammography Starting at Age 40 Years: A Comparative Modeling Study of Risk,” Ann. Intern. Med., vol. 156, no. 9, pp. 609–617, May 2012, doi: 10.7326/0003-4819-156-9-201205010-00002.

[16] S. Haneuse et al., “Mammographic Interpretive Volume and Diagnostic Mammogram Interpretation Performance in Community Practice,” Cancer, vol. 262, no. 1, pp. 69–79, 2012, doi: 10.1148/radiol.11111026/-/DC1.

[17] C. Ngufor and J. Wojtusiak, “Learning from large-scale distributed health data: an approximate logistic regression approach,” Proc. ICML 13 Role Mach. Learn. Transform. Healthc., vol. 28, 2013.

[18] L. K. Koegel, T. W. Vernon, R. L. Koegel, B. L. Koegel, and A. W. Paullin, “Improving Social Engagement and Initiations Between Children With Autism Spectrum Disorder and Their Peers in Inclusive Settings,” J. Posit. Behav. Interv., vol. 14, no. 4, pp. 220–227, Oct. 2012, doi: 10.1177/1098300712437042.

[19] T. Chu, J. Wang, and J. Chen, “An adaptive online learning framework for practical breast cancer diagnosis,” 2016, p. 978524, doi: 10.1117/12.2216550.

[20] M. U. Salma and others, “Reducing the Feature Space Using Constraint-Governed Association Rule Mining,” J. Intell. Syst., vol. 26, no. 1, pp. 139–152, 2017.

[21] A. Hazra, S. K. Mandal, and A. Gupta, “Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms,” Int. J. Comput. Appl., vol. 145, no. 2, pp. 975–8887, 2016.

[22] K.-M. Wang, B. Makond, W.-L. Wu, K.-J. Wang, and Y. S. Lin, “Optimal Data Mining Method for Predicting Breast Cancer Survivability,” Int. J. Innov. Manag. Inf. Prod., vol. 4, no. 2, pp. 28–33, 2013.

[23] “Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C),” http://www.bcsc-research.org/.

[24] N. Rathore, S. Agarwal, and others, "Predicting the survivability of breast cancer patients using ensemble approach," pp. 459–464, doi: 10.1109/ICICICT.2014.6781326.

[25] H. L. Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, "Prediction of breast cancer survival through knowledge discovery in databases," Glob. J. Health Sci., vol. 7, no. 4, p. 392, 2015.

[26] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 1, pp. 39–45, 2013.

[27] H. Sawhney and H. Kaur, "Implementation and Applications of Data Minin in Medical Decision Making Predictions," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 7, 2017.

[28] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," Int. J. Mach. Learn. Comput., vol. 3, no. 2, p. 224, 2013.

[29] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," Int. J. Comput. Sci. Netw., vol. 2, no. 1, pp. 83–87, 2013, doi: 10.1109/SIU.2013.6531574.

[30] R. Kothandan, "Handling class imbalance problem in miRNA dataset associated with cancer," Bioinformation, vol. 11, no. 1, p. 6, 2015.

[31] R. R. Rao and K. Makkithaya, "Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance," Int. J. Electr. Comput. Eng., vol. 7, no. 4, p. 2215, 2017.

[32] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, 2004.

[33] T. M. Fahrudin, I. Syarif, and A. R. Barakbah, "Data Mining Approach for Breast Cancer Patient Recovery," Emit. Int. J. Eng. Technol., vol. 5, no. 1, pp. 36–71, 2017.

[34] S. Kharya and S. Soni, "Weighted naive bayes classifier: A predictive model for breast cancer detection," Int. J. Comput. Appl., vol. 133, no. 9, pp. 32–37, 2016.

[35] H. You and G. Rumbe, "Comparative study of classification techniques on breast cancer FNA biopsy data," IJIMAI, vol. 1, no. 3, pp. 6–13, 2010.

[36] C.-R. Nicandro et al., "Evaluation of the diagnostic power of thermography in breast cancer using bayesian network classifiers," Comput. Math. Methods Med., vol. 2013, 2013.

[37] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

[38] D. Patel, B. Tanwala, and P. Patel, "Breast Cancer Using Data Mining Techniques," Int. J. Comput. Sci. Eng., vol. 6, no. 7, pp. 1531–1536, 2018, doi: 10.26438/ijcse/v6i7.15311536.