J Med Syst (2014) 38:106
DOI 10.1007/s10916-014-0106-1

PATIENT FACING SYSTEMS

Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree

Cheng-Min Chao · Ya-Wen Yu · Bor-Wen Cheng · Yao-Lung Kuo

Received: 31 December 2013 / Accepted: 7 July 2014 / Published online: 14 August 2014

Abstract The aim of the paper is to use data mining technology to establish a classification of breast cancer survival patterns, and offers a treatment decision-making reference for the survival ability of women diagnosed with breast cancer in Taiwan. We studied patients with breast cancer in a specific hospital in Central Taiwan to obtain 1,340 data sets. We employed a support vector machine, logistic regression, and a C5.0 decision tree to construct a classification model of breast cancer patients' survival rates, and used a 10-fold cross-validation approach to identify the model. The results show that the establishment of classification tools for the classification of the models yielded an average accuracy rate of more than 90 % for both; the SVM provided the best method for constructing the three categories of the classification system for the survival mode. The results of the experiment show that the three methods used to create the classification system, established a high accuracy rate, predicted a more accurate survival ability of women diagnosed with breast cancer, and could be used as a reference when creating a medical decision-making frame.

Keywords Breast cancer · Support vector machine · Logistic regression · C5.0 decision tree · 10-fold cross-validation

This article is part of the Topical Collection on *Patient Facing Systems*

C.-M. Chao
Department of Business Administration, National Taichung University of Science and Technology, Taichung, Taiwan

Y.-W. Yu (✉) · B.-W. Cheng · Y.-L. Kuo
National Yunlin University of Science and Technology, Douliu, Yunlin Country, Taiwan
e-mail: g9821802@yuntech.edu.tw

Introduction

Breast cancer is the most common cancer diagnosed in women and is currently a widely discussed issue. From a clinical point of view, detecting early stage breast cancer is vital, but also difficult. In the United States, breast cancer is the most frequently diagnosed malignancy, and is the greatest cause of cancer deaths in women [1]. In Taiwan, breast cancer is the fourth leading cause of cancer deaths in women (Department of Health, Executive Yuan Report in Taiwan, 2001–2011), and the death rate is increasing each year [2]. Approximately 1.9 % of all female breast cancer patients are under the age of 35, and this diagnosis is physically and mentally overwhelming for these patients [3]. Taiwanese women are generally diagnosed with breast cancer at younger ages than their counterparts in the United States. Although breast cancer is the major cause of cancer deaths in women, survivability is high. With early diagnosis, 97 % of women live for 5 years or longer [4, 5].

A recent study indicated that the chance of a woman being affected by invasive breast cancer during her lifetime is approximately 1 in 8, and the chance of death is 1 in 35 [6]. Although breast cancer is curable when detected early, approximately one third of female breast cancer patients die of the disease [7]. However, despite early detection and new treatment options, up to 50 % of women will develop distant metastases, which are currently incurable. As the causes of breast cancer have not been identified, precise early detection is crucial to reduce the high mortality rate [6].

In the medical field, numerous researchers have adopted different methods in an attempt to improve the precision of data classification. Methods with better classification precision would provide more robust data for the identification of prospective patients and would improve diagnostic accuracy. Data mining tools [e.g., support vector machines (SVM), decision trees (DT) and neural networks], statistical tools

[e.g., logistic regression (LR) and discriminant analysis], and meta-heuristic algorithms (e.g., genetic algorithms, tabu searches, particle swarm optimization, and simulated annealing) have been applied in the medical field with significant results [8, 9]. This study, through the analysis of various clustering and classification techniques and also by considering the imbalanced nature of data, attempts to determine the best suited algorithm for early breast cancer identification. We address a critical problem in this health informatics research field.

Despite several currently recognized leading factors of breast cancer, researchers are still studying the main causes. The lymph status, tumor size, stage at presentation, and histological grade are considered to be the most critical factors in breast cancer survivability [3, 10]. Certain researchers [4, 6] argued that obesity, hormonal factors, and family history might increase the risk of a breast cancer diagnosis. Liao and Tsai [11] examined the risk factors associated with DNA viruses, fibroadenomas, and normal mammary tissues among with respect to a Taiwanese sample. They adopted a case–control design using 62 women with nonfamilial invasive ductal breast cancer, 32 women with mammary fibroadenomas, and 12 with normal mammary tissue. The results showed that DNA viruses, fibroadenomas, and normal mammary tissues are related to certain risk factors in breast cancer.

Certain researchers have developed a number of statistical methods for breast cancer diagnosis (e.g., [8, 12]) because traditional data analysis techniques are inadequate for processing large volumes of data. Data mining is a systematic approach used to locate underlying trends, patterns, and relationships buried in data. Data mining is broadly classified into two categories: technology and methodology. Applications of data mining have provided benefits in several medical procedures, including diagnosis, prognosis, and treatment [13]. Pendharkar et al. [14] indicated that data mining can be used in breast cancer diagnosis.

Other researchers (e.g., [8, 15]) have applied a popular technique for mammogram classification: the SVM. This technique has improved the prediction performance in breast cancer diagnosis. Ismail Saritas [16] using artificial neural network (ANN) and breast imaging reporting and data system (BI-RADS) evaluation and based on the age of the patient, mass shape, mass border and mass density to predicting the patients have breast cancer or not. The result showed that disease prediction rate was 90.5 % and the health ratio was 80.9 %. It is seen that the ANN model can be used in breast cancer. Shoorehdeli [17] using fuzzy neural networks (FNN), hierarchical fuzzy neural network (HFNN) and fuzzy gaussian potential neural network (FGPNN) to classify breast cancer into two groups; benign and malignant lesions. The result showed that the performance of HFNN and FGPNN are evaluated and compared with FNN. The technique has improved the prediction performance in breast cancer diagnosis. Huang et al. [18] recently compared the use of particle swarm optimizer (PSO) based ANN, the adaptive neuro-fuzzy inference system (ANFIS), and a case-based reasoning (CBR) classifier with a logistic regression model and decision tree model. The result showed that the ANFIS is better than others. Chen et al. [19] is proposed to using PSO_SVM for breast cancer diagnosis. The result showed that the proposed approach giving high predictive accuracy. The technique can ensure that the physicians make very accurate diagnostic decision in clinical breast cancer diagnosis. Huang et al. [20] is to combine the feature selection (FS) and optimization algorithms (using the levenberg marquardt (LM) and the PSO algorithms to devise the appropriate NN training weighting parameters) to improve the Wisconsin Breast Cancer Dataset classification. This technique has improved the prediction performance in breast cancer' classification accuracy and efficiency. Chhatwal et al. [12] proposed that LR models can discriminate between benign and malignant conditions in the decision making of early breast cancer detection, and can recognize the most crucial factors linked with the disease. Delen et al. [21] recently compared the use of LR, ANN, DTs (C5) for predicting the survival rates of diagnosed cases of breast cancer. They reported an accuracy rate of 91.2 % for artificial neural networks and 93.6 % for DTs, by employing a large data set of more than 200,000 cases collected between 1973 and 2000. A few published studies have compared different classification techniques in several areas. Lee et al. [22] further improved the cluster selection method by proposing a classifier which could classify instances of breast cancer into the same three survival categories with 82.7 % accuracy.

This study mainly focused on identifying the best suited algorithms for early breast cancer detection based on different factors. Thus, the purpose of this research was to develop predictive models to discover or explain the relationships among certain independent variables and the survival rates in the context of breast cancer. We used the breast cancer incidence database from a hospital in Southern Taiwan, focusing on three types of classification models: SVM, LR, and the C5.0 DT. In addition, a ten-fold cross-validation technique was used to compare the accuracy of these models.

Materials and Methods

Support Vector Machines

Data classification and regression are important in various research fields. SVMs are a type of supervised machine learning (ML) algorithm used to classify data points by maximizing the margin between classes in a high-dimensional space [23]. As an effective ML technique, SVM is used for decision making, for data classification and regression in many fields [23]. Since Cortes and Vapnik proposed the SVM in 1995, the technique has gained considerable attention in the ML

community because of its exceptional performance in various learning problems, including pattern recognition and bioinformatics [19], and medical diagnosis [8]. An SVM can produce a regression model or a classification function based on a set of training data.

The SVM, which seeks the optimal boundary between two classes, is used as the classifier in this paper. The maximum-margin hyperplane generates the maximum separation between decision classes. The training data closest to the maximum-margin hyper plane are called support vectors. The SVM can solve problems with linear and non-linear segmentations ([8, 24]) by using a training set with input vectors and target labels:

$$(x_1,\ y_i),\quad i = 1\cdots l, x\in R^n\ , y\in\{+1,-1\} \tag{1}$$

provided the following conditions are satisfied:

$$\begin{aligned} x_i + b &\geq +1 \ \text{ for } \ y_i = +1 \\ x_i + b &\leq -1 \ \text{ for } \ y_i = -1 \end{aligned} \tag{2}$$

which is equivalent to

$$y_i(w_i \times w + b) - 1 \geq 0 \quad \forall_i \tag{3}$$

This technique is used to search for a hyperplane $w \times x_i + b = 0$ to separate the data into classes of +1 and −1; it has a maximal margin in the feature space with a margin width between the hyperplanes equal to

$$\frac{2}{\|w\|^2} \tag{4}$$

The maximization of the margin is equivalent to the minimization of the norm of w. In primal weight space, the classifier uses the decision function (Eq. 5).

$$f(x) = sign(w \times x) + b \tag{5}$$

Thus, as Cristianini and Taylor [25] indicated, the SVM was trained to solve the following optimization problem:

$$Minimize: \ \frac{1}{2} W^T + C\sum_{i=1}^{N} \xi_i \tag{6}$$

subject to $y_i(w_i \times w + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1, \ldots, n$,

where C is a regularization parameter that imposes a trade-off between the training error and the generalization, and represents slack variables. These restrictions are imposed to ensure that no training patterns are within the margins. However, the restrictions are relaxed by using the slack variables to eliminate noisy data. The classifier represented in Eq. (6) was restricted because it performs only a linear separation of the data. The restriction can be overcome by mapping the input examples to a high-dimensional space, where they can be efficiently separated by a linear SVM. The mapping was performed using kernel functions, which allowed access to spaces of higher dimensions without explicitly knowing the usually relatively complex mapping function.

The SVM can also manage linearly inseparable problems by employing current data for training, and displays all data by choosing several support vectors from training data. Several extreme values are pre-eliminated, and a model is formed from the selected support vectors. The SVM was originally proposed to solve binary classification problems. However, SVM has recently been adopted for treatment prognosis, transition prediction, and disease diagnosis, and uses both structural and functional neuroimaging data. Therefore, this study used SVM as its research method, as proposed by Luo and Cheng [8]; we evaluated the SVM model to compare various classification techniques, and to predict patient deaths and their need for dialysis.

Logistic Regression

LR is a widely used and accepted statistical decision support tool which predicts the probability of an occurrence of an event by fitting data to a logistic function [26]. It is used to predict instant clinical outcomes for individual patients by using clinical, demographical, and other factors [27]. LR is also a multivariable method. It attempts to build a functional relationship between two or more predictor (independent) variables and the one outcome (dependent) variable. In this study, binary LR was chosen to predict the membership of only two categorical outcomes. Although the primary output of the LR model is the estimated odds or probability of a binary event, additional information can be acquired from the model and should be applied in decision making. In two-class problems, odds that are greater than 50 % are assigned to a class designated as "1." Other cases are designated as "0." Although LR is an effective modeling tool, it assumes that the response variables are linearly related to the coefficients of the predictor variables. Stepwise selections of the independent variables were stepwise incremented and the corresponding coefficients were computed. The general form of the LR functional model for n independent variables can be written as

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} = \frac{1}{1 + e^{-(b^T \times X)}} \tag{7}$$

where $x_0, x_1, \cdots x_n$ are the predictor variables, $\beta_0, \beta_1, \cdots \beta_n$ are the regression coefficients, and P(Y) is the probability of the presence of breast cancer. A binary LR was performed to determine vector b, which for the data set under consideration, associates each record (a patient) with the probability of breast cancer. We used LR and a forward stepwise variable selection process to predict the probability of breast cancer.

Decision Tree

A DT is a hierarchical model composed of decision rules that recursively split independent variables into homogeneous zones [28]. DTs are robust classification algorithms, which are becoming popular with the growth of data mining in the information systems field. DT-based classification algorithms have tree structures consisting of nodes (or leaves), and branches. The tree structure is constructed based on a set of decision rules applied in a certain order. Branches (i.e., splits) at a particular level of the tree are composed using different decision rules [29]. The purpose of DT building is to search for a set of decision rules to predict an outcome from a set of input variables [28]. A DT can be interpreted as a rule-induction technique if different scenarios are forecasted from the tree structure. A DT consists of three types of nodes: decision nodes, chance nodes, and end nodes. Several algorithms are used to construct decision-tree algorithms, including classification and regression trees (CART), ID3s, chi-square automatic interaction detector DTs (CHAIDs), and C4.5 and C5.0 DTs [26, 28]. The goal of these algorithms is to minimize the size of the tree; thereby preserving the quality of the final decision and shortening the decision process. C5.0 also applies a boosting technique to generate and connect multiple classifiers to enhance predictive accuracy. The error rate of C5.0-boosted classifiers is approximately one third of the error rate of C4.5-single classifiers [30]. Biggs et al. [30] used a DT to categorize an entire sample according to whether patients were likely to receive hemodialysis.

Measures Foe Evaluating Performance

In this study, we used the confusion matrix presented in Table 1 and the accuracy obtained from Eq. 8 [21] to evaluate performance.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

K-Fold Cross-Validation

In k-fold cross-validation (also called rotation estimation), a complete data set (D) is randomly split into k mutually exclusive subsets of approximately equal sizes (folds: D1, D2…, Dk). Empirical studies have shown that stratified cross-validation tends to generate comparison results with lower bias and lower variance compared to those of regular k-fold cross-validation. Furthermore, to estimate the performance of classifiers, a stratified ten-fold cross-validation approach is used. Empirical studies have shown that ten appears to be the optimal number of folds (e.g., [31]).

To estimate the performance of classifiers, a stratified ten-fold cross-validation approach was used. In ten-fold cross-validation, the entire data set is divided into ten mutually exclusive subsets with approximately the same class distribution as the original data set. Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds; this leads to ten independent performance estimates. In addition, when the number of survived: 1,272 and deceased: 68 cannot be divided by 10, the total is divided by 10 to make each fold of equal size with 134 cases with an additional training set of size 1,206 to attain the total of 1,340.

Data Analysis

Data Source We used data from the Diseases Database of one medical center spanning the years 2002 to 2010. The raw dataset was composed of 1,721 patients, of which only 1,340 cases were considered. Eight variables were in each file, and each record in the file was related to a specific incidence of breast cancer. There were eight variables, including one dependent variable and seven predictor variables. The dependent variable is qualitative, and is divided into survival and deceased. The predictor variables include age, tumor size, the number of examined lymph nodes, the number of attacked lymph nodes, and variable types that are quantitative variables. Pathological staging, chemotherapy, and radiotherapy are qualitative variables.

Understanding the data and the data preparation stages are among the most important steps of data mining applications. The majority of time spent on developing data mining applications is on these earliest stages [32]. The raw data were uploaded into an Excel database. The SPSS statistical analysis tool, the statistical data miner, and the Clementine data mining toolkit were used to explore and manipulate the data.

This section describes the surface complexities and structure of the data. After initial screening, 1,340 useable records

Table 1 A confusion matrix

Actual state		Class is “true” (positive)	Class is “false” (negative)
Predicted state	Classified as “true” (positive)	TP	FP
	Classified as “false” (negative)	FN	TN

remained. Patient age ranged from 25 to 92 years, and the mean age was 55±11 years. There were 1,272 survivors and 68 deaths. The means for the variable tumor sizes, the number of lymph nodes examined, and the number of lymph nodes attacked were 2.63±1.77, 17.38±9.84, and 2.41±5.60, respectively.

Data Processing We used an SVM, Logistic regression, and C5.0 decision tree to train two real-world data sets. Our tools included SPSS for Windows 12.0 and Clementine 7.2. All of our experimental results were obtained using an AMD Turion64 MK38 2.2 GHz PC and a Windows XP professional operating system. We performed experiments using published template data sets, and compared the results of SVMs, Logistic regression, and C5.0 decision trees. Because the appropriate kernel and the kernel parameters were unknown, several different parameters and randomly selected kernels were used for the SVM, logistic regression, and C5.0 decision tree to locate the best performance for prediction.

Results The experiments performed under this study used published template data sets for a comparison of the results of SVM, logistic regression, and C5.0 decision tree. Comparisons were based on 10-fold cross-validation. We evaluated the models based on the discussed accuracy measures for classification. The results were derived using 10-fold cross-validation for each model, and were based on the average results obtained from the test data set (10^{th} fold) for each fold.

Table 2 shows the comparisons of accuracy for the SVM, logistic regression, and C5.0 decision tree for the deceased data set. The SVM achieved a classification average accuracy of 95.22 %. The logistic regression model achieved a classification average accuracy of 95.1 %. The C5.0 decision tree model achieved a classification average accuracy of 93.95 %. These results show that the SVM is the best of the three categories of classification methods for survival mode.

The logistic regression model selected six variables (including: age, radiotherapy, tumor size, number of examined lymph nodes, number of attacked lymph nodes, and pathological staging), and the C5.0 decision tree model selected six variables (including: age, radiotherapy, tumor size, number of examined lymph nodes, number of attacked lymph nodes, pathological staging, and chemotherapy).

Table 3 The probabilities of the different models experiencing Type I and Type II errors

	SVM	Logistic regression	C5.0 decision tree
Type I error	0.002	0.003	0.018
Type II error	0.806	0.939	0.892

For classification problems, the average accuracy of the classification model should also consider the misplaced cost in determining the best classification model. When handling binary classification problems, the expected misclassification costs are as follows:

Expected misclassification costs = $C(0|1)\times P(0|1)\times \pi_1 + C(1|0)\times P(1|0)\times \pi_0$, where π_1 and π_0 are the a priori probabilities of the survival and incapable of survival groups, respectively; $P(0|1)$ and $P(1|0)$ are the probabilities of survival misjudged as incapable of survival (Type I error), and incapable of survival misjudged as survival (Type II error), respectively; and $C(0|1)$ and $C(1|0)$ are the costs of survival being misjudged as incapable of survival (Type I error), and the costs of incapable of survival being misjudged as survival (Type 2 error), respectively. The ratio of survival and incapable of survival extracted from the database is used as the a priori probability in this study. These probabilities are $\pi_1 = 0.95$ and $\pi_0 = 0.05$, respectively. The probabilities of Type I and Type II errors are the 10-group classification models obtained through 10-fold cross-validation for each type of classification tool, calculations for the average values of Type I and Type II errors are shown in Table 3.

The cost of Type II errors occurring was generally higher, compared to Type I errors. Hofmann suggested that the delay costs caused by the occurrences of Type II errors was five times greater than those of Type I errors. Therefore, we changed the cost ratio of the occurrences of Type I and II errors from (1: 1) to (1: 10) to compare the expected misclassification costs among the various models, as shown in Table 4. When considering the expected misclassification costs of Type I and II errors, logistic regression has the highest cost of the three.

Table 2 Results predicted for patients' deaths in 10-fold cross-validation for all folds and all model types

Fold No.	SVM		Logistic regression		C5.0 decision tree	
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Best accuracy	98.34	98.51	95.5	98.5	97.01	97.76
Mean	96.52	95.22	95.02	95.1	96.3	93.95
St. Dev.	1.26	2.52	0.31	2.43	0.55	2.54

Table 4 Expected misclassification costs for the various models

Cost ratio	SVM	LR	C5.0DT	Cost ratio	SVM	LR	C5.0DT
1:1	0.043	0.050	0.061	1:6	0.244	0.285	0.284
1:2	0.083	0.097	0.106	1:7	0.284	0.332	0.329
1:3	0.123	0.144	0.151	1:8	0.325	0.379	0.373
1:4	0.164	0.191	0.195	1:9	0.365	0.426	0.418
1:5	0.204	0.238	0.240	1:10	0.405	0.473	0.463

Discussion and Conclusions

For our study, we developed several prediction models for breast cancer survival. In particular, we used three popular data-mining methods: one for statistics (logistic regression), and two for machine learning (SVM and C5.0 decision trees). We analyzed a large data set (1,340 cases and 7 prognostic factors), and after a long process of data selection and transformation, we developed the prediction models. Our results showed that the SVM (95.15 %) outperformed the logistic regression (95.1 %) and the C5.0 decision tree (93.95 %) regarding prediction accuracy. In addition to assessing the average accuracy of each classification model, errors of Type I and Type II were considered. Type I and Type II errors are limited to the unbalanced number of survivals and the death rate from the data set. The imbalance resulted from inadequate death samples, causing the high inaccuracy of Type II error. Classification approaches can help physicians diagnose breast cancer, and our experiments with the learning-association rules show that risk assessment expert systems can be developed.

The logistic regression model selected six variables (ie, age, radiotherapy, tumor size, number of examined lymph nodes, number of attacked lymph nodes, and pathological staging), and the C5.0 decision tree model selected seven variables (ie, age, chemotherapy, radiotherapy, tumor size, number of examined lymph nodes, number of attacked lymph nodes, pathological staging, and chemotherapy). These variables contain critical information in the cancer staging system TN, M, and Nottingham Prognostic Index (NPI) (an indicator used for predicting cancer patients' survival). From the Delen et al. [21] study, we selected three key variables: pathological staging, radiotherapy, and tumor size [21]. According to D'Eredita et al. [10], lymph status, tumor size, and histological grade are typically the most critical factors for breast cancer survivability [10]. These three features also emerged as key variables in our study.

With the advancement of medical treatment and knowledge, doctors may be familiar with their own profession, but with the advancement of medical treatment and knowledge, they can nevertheless benefit from clinical learning models. Educational training and multidisciplinary cooperation are two methods that can be used to fill this gap. In clinical diagnosis, pattern recognition can assist in identifying symptoms of particular diseases. In contrast, complex symptoms must be analyzed based on patients' presentation, physical examination, and examination results. All diagnoses should be processed using a hypothetico-deductive method using the physician's medical knowledge. However, with the help of information technology, statistical analysis, and artificial intelligence, physicians may more effectively assess and deduce possible causes of disease and treatment results. Our attempts to discover the risk factors of breast cancer using information technology and the principles of medical management have been outlined. We hope that it may provide a reference point for physicians treating new cases.

For future research, we recommend that data be published using a standard vocabulary and format so that the information can be easily understood and shared [33]. In addition, it is vital to compare the results of data mining across various studies, comparing the performance of different models. However, the performances of certain classifiers, for example, those based on ANN, are critically dependent on fine-tuning the parameters. We did not examine this particular issue in this study.

Acknowledgments This research was performed under the auspices of Taiwan's National Science Council (NSC 99-2221-E-224-033-MY2).

References

1. Fabregue, M., Bringay, S., Poncelet, P., Teisseire, M., and Orsetti, B., Mining microarray data to predict the histological grade of a breast cancer. *J. Biomed. Inform.* 44(1):S12–S16, 2011. doi:10.1016/j.jbi.2011.03.002.
2. Department of Health, Executive Yuan, R.O.C., 2013. Retrieved from http://www.mohw.gov.tw/cht/DOS/Statistic.aspx?f_list_no=312&fod_list_no=2747.
3. Hartmann, S., Reimer, T., and Gerber, B., Management of early invasive breast cancer in very young women (<35 years). *Clin. Breast Cancer* 11(4):196–203, 2011. doi:10.1016/j.clbc.2011.06.001.
4. Jerez-Aragonés, J. M., Gomez-Ruiz, J. A., Ramos-Jimenez, G., Munoz-Perez, J., and Alba-Conejo, E., A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artif. Intell. Med.* 27(1):45–63, 2003. doi:10.1016/S0933-3657(02)00086-6.
5. O'Malley, C. D., Le, G. M., Glaser, S. L., Shema, S. J., and West, D. W., Socioeconomic status and breast carcinoma survival in four racial/ethnic groups: A population-based study. *Am. Cancer Soc.* 97(5):1303–1311, 2003. doi:10.1002/cncr.11160.
6. Nahar, J., Imam, T., Tickle, K. S., Ali, A. B. M. S., and Chen, Y.-P. P., Computational intelligence for microarray data and biomedical

image analysis for the early diagnosis of breast cancer. *Expert Syst. Appl.* 39(16):12371–12377, 2012. doi:10.1016/j.eswa.2012.04.045.
7. Keles, A., Keles, A., and Yavuz, U., Expert system based on neuro-fuzzy rules for diagnosis breast cancer. *Expert Syst. Appl.* 38(5): 5719–5726, 2011. doi:10.1016/j.eswa.2010.10.061.
8. Luo, S. T., and Cheng, B. W., Diagnosing breast masses in digital mammography using feature selection and ensemble methods. *J. Med. Syst.* 36(2):569–577, 2012. doi:10.1007/s10916-010-9518-8.
9. Fan, C.-Y., Chang, P.-C., Lin, J.-J., and Hsieh, J. C., A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Appl. Soft Comput.* 11(1):632–644, 2011. doi:10.1016/j.asoc.2009.12.023.
10. D'Eredita, G., Giardina, C., Martellotta, M., Natale, T., and Ferrarese, F., Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution. *Eur. J. Cancer* 37(1):591–596, 2001. doi:10.1016/s0959-8049(00)00435-4.
11. Liao, H. C., and Tsai, J. H., Data mining for DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue. *Appl. Math. Comput.* 188(1):989–1000, 2007. doi:10.1016/j.amc.2006.10.069.
12. Chhatwal, J., Alagoz, O., Lindstrom, M. J., Kahn, C. E., Jr., Shaffer, K. A., and Burnside, E. S., A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *Am. J. Roentgenol.* 192(4):1117–1127, 2009. doi:10.2214/AJR.07.3345.
13. Richards, G., Rayward-Smith, V. J., Sonksen, P. H., Carey, S., and Weng, C., Data mining for indicators of early mortality in a database of clinical records. *Artif. Intell. Med.* 22(3):215–231, 2001. doi:10.1016/S0933-3657(00)00110-X.
14. Pendharkar, P. C., Rodger, J. A., Yaverbaum, G., Herman, N., and Benner, M., Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Syst. Appl.* 17(3):223–232, 1999. doi:10.1016/S0957-4174(99)00036-6.
15. Acharya, U. R., Ng, E. Y., Tan, J. H., and Sree, S. V., Thermography based breast cancer detection using texture features and Support Vector Machine. *J. Med. Syst.* 36(3):1503–1510, 2012. doi:10.1007/s10916-010-9611-z.
16. Saritas, I., Prediction of breast cancer using artificial neural networks. *J. Med. Syst.* 36(5):2901–2907, 2012. doi:10.1007/s10916-011-9768-0.
17. Shoorehdeli, M. A., Breast cancer classification based on advanced multi dimensional fuzzy neural network. *J. Med. Syst.* 36(5):2713–2720, 2012. doi:10.1007/s10916-011-9747-5.
18. Huang, M. L., Hung, Y. H., et al., Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. *J. Med. Syst.* 36(2):407–414, 2012.
19. Chen, et al., Support vector machine based diagnostic system for breast cancer using swarm intelligence. *J. Med. Syst.* 36(4):2505–2519, 2012. doi:10.1007/s10916-011-9723-0.
20. Huang, M. L., Hung, Y. H., and Chen, W. Y., Neural network classifier with entropy based feature selection on breast cancer diagnosis. *J. Med. Syst.* 34(5):865–873, 2010. doi:10.1007/s10916-009-9301-x.
21. Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* 34(2):113–127, 2005. doi:10.1016/j.artmed.2004.07.002.
22. Lee, Y. J., Mangasarian, O. L., and Wolberg, W. H., Survival-time classification of breast cancer patients. *Comput. Optim. Appl.* 25(1–3):151–166, 2003. doi:10.1023/A:1022953004360.
23. Vapnik, V., *The nature of statistical learning theory.* Springer, New York, 1995.
24. Stoean, R., Stoean, C., et al., Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C. *Artif. Intell. Med.* 51(1):53–65, 2011.
25. Cristianini, N., and Taylor, J., *An introduction to support vector machines*. Cambridge University Press, Cambridge, UK, 2000.
26. Quinlan, J. R., *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, 1993.
27. Mazzocco, T., and Hussain, A., Novel logistic regression models to aid the diagnosis of dementia. *Expert Syst. Appl.* 39(3):3356–3361, 2012. doi:10.1016/j.eswa.2011.09.023.
28. Pradhan, B., A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* 51(1): 350–365, 2013.
29. Petrović, J., Ibrić, S., Betzb, G., and Durić, Z., Optimization of matrix tablets controlled drug release using Elman dynamic neural networks and decision trees. *Int. J. Pharm.* 428(1–2):57–67, 2012. doi:10.1016/j.ijpharm.2012.02.031.
30. Biggs, D., et al., A method of choosing multiway partitions for classification and decision trees. *J. Appl. Stat.* 18(1):49–62, 1991.
31. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.
32. Cios, K., and Moore, G., Uniqueness of medical data mining. *Artif. Intell. Med.* 26(1):1–24, 2002. doi:10.1016/S0933-3657(02)00049-0.
33. Szalay, A., and Gray, J., Science in an exponential world. *Nature* 440(1):413–414, 2006.