Engineering Applications of Artificial Intelligence 26 (2013) 2194–2205

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Robust predictive model for evaluating breast cancer survivability

Kanghee Park [a], Amna Ali [b], Dokyoon Kim [c], Yeolwoo An [a], Minkoo Kim [b], Hyunjung Shin [a,*]

[a] Department of Industrial Engineering, Ajou University, Wonchun-dong, Yeongtong-gu, Suwon 443-749, South Korea
[b] Department of Computer Engineering, Ajou University, Wonchun-dong, Yeongtong-gu, Suwon 443-749, South Korea
[c] Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, South Korea

ARTICLE INFO

Article history:
Received 11 September 2012
Received in revised form
14 March 2013
Accepted 26 June 2013
Available online 25 July 2013

Keywords:
Machine learning
Semi-supervised learning
Breast cancer survivability

ABSTRACT

Objective: Many machine learning models have aided medical specialists in diagnosis and prognosis for breast cancer. Accuracy has been regarded as a primary measurement for the performance evaluation of the models, but stability which indicates the robustness of the performance to model parameter variation also becomes essential. A stable model is in practice of benefit to the medical specialists who may have little expertise in model tuning. The main purpose of this work is to address the importance of the stability of a model and to suggest one of such models.
Methods: A comparative study of three prominent machine learning models was carried out for the prognosis of breast-cancer survivability: support vector machines, artificial neural networks, and semi-supervised learning models.
Material: The surveillance, epidemiology, and end results database for breast cancer was used, which is known as the most comprehensive source of information on cancer incidence in the United States.
Results: The best performance was obtained from the semi-supervised learning model. It showed good overall accuracy and stability under model parameter variation. The sharpening procedure enhanced the stability of the model via the noise-reduction.
Conclusion: We suggest that semi-supervised learning model is a good candidate that medical professionals readily employ without consuming the time and effort for parameter searching for a specific model. The ease of use and faster time to results of the predictive model will eventually lead to the accurate and less-invasive prognosis for breast cancer patients.

1. Introduction

Breast cancer is one of the major diseases of the world. It is the most common type of cancer and the second leading cause of cancer deaths (after lung cancer) in women (Cancer Facts & Figures, 2010). In the United States, breast cancer is the most frequently diagnosed malignancy in women. Moreover, it was estimated that around 232,340 new cases of invasive breast cancer would be diagnosed in women in 2013; and around 40,678 women were expected to die of this disease (Siegel et al., 2013). Men can also suffer from this cancer (NC Institute. Breast Cancer Statistics, USA, 2010). According to the American Cancer Society, 1970 new cases of breast cancer would be found in men in the United States in 2010, and the expected number of deaths was 390 (Cancer Facts & Figures, 2010). Researchers are devoting considerable effort to search for enhanced and innovative techniques for the early detection and treatment of breast cancer. Therefore, the death rate for breast cancer has gradually decreased in women since 1990. A larger decrease has occurred for women younger than 50 years (3.2% annual decrease) than for those who are 50 years or older (2.0% annual decrease) (Cancer Facts & Figures, 2010).

The major clinical problem associated with breast cancer is to predict the outcome (survival or death) after the onset of this therapeutically resistant disseminated disease. In many cases, by the time the primary tumor is diagnosed, clinically evident metastases have already occurred. In general, treatments such as chemotherapy, hormone therapy, or a combination are considered to reduce the spread of breast cancer by decreasing the distant metastases, by one-third. However, studies have shown that 70% of patients receiving these therapies would have survived without them (Sun et al., 2007). Therefore, the ability to predict disease outcomes more accurately would allow physicians to make informed decisions on the potential necessity of adjuvant treatment. This could also lead to the development of individually tailored treatments to maximize treatment efficiency (Khan et al., 2008).

Prognosis helps to establish a treatment plan by predicting the outcome of a disease. There are three predictive foci of cancer prognosis: (1) prediction of cancer susceptibility (risk assessment),

* Corresponding author. Tel.: +82 312 192 417; fax: +82 312 191 610.
E-mail address: shin@ajou.ac.kr (H. Shin).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.06.013

(2) prediction of cancer recurrence (redevelopment of cancer after resolution), and (3) prediction of cancer survivability. In the third case, research focuses on predicting the outcome in terms of life expectancy, survivability, progression, or tumor-drug sensitivity after the diagnosis of the disease. We focus on the survivability prediction for a particular patient suffering from breast cancer. According to (Delen et al., 2005), survival analysis is a part of medical prognosis and involves the use of methods and techniques for predicting the survival of a particular patient on the basis of historical data of patients. In general, "survival" can be defined as the patient remaining alive for a specified period after the diagnosis of the disease. As recommended by (Delen et al., 2005) and, (Brenner et al., 2002), if the patient is still living for 1825 days (5 years) after the date of diagnosis, then the patient is considered to have survived.

Informed decision making for breast cancer patients is the basic motivation behind the growing emphasis on accurate and less-invasive personalized predictive models based on machine learning techniques. This approach can allow many breast cancer patients to avoid complex surgical biopsies, unnecessary adjuvant treatments, and high medical costs. Moreover, in situations where experienced oncologists are not available, predictive models created via machine learning techniques can support physicians' decision-making with acceptable accuracy (Amir et al., 2003). However, oncologists must determine the best parameters for the predictive models, and they have little or no expertise in such parameter selection. It would therefore be convenient to use a model that is robust to parameter variation. No matter how accurate a predictive model is, it will not be useful unless it is robust. To build a robust predictive model, information domain researchers need a large quantity of breast cancer survival data for analysis. There are two types of data: labeled (feature/label pairs) and unlabeled (features without labels). Accumulating a substantial quantity of labeled data is time-consuming, costly, and requires confidentiality agreements. In general, the collection of labeled survival data requires at least 5 years (Delen et al., 2005; Brenner et al., 2002). Moreover, oncologist consultation fees must be paid to confirm survivability. Furthermore, doctors and patients seldom reveal their information. Now, the subject of inquiry is that in order to acquire the survival data whether is it worthy to wait for 5 years, pay significant amount of fee and exert a great deal efforts to convince patients to disclose their personal medical data? Unlabeled data can be collected with much less efforts. Therefore, an economical solution is to use a large quantity of unlabeled data and a small quantity of labeled data, together with the semi-supervised learning (SSL) that has recently emerged in the machine learning domain.

SSL is an attractive method for improving classification models by using unlabeled data to support supervised learning (Zhong, 2006). In research areas such as speech recognition, text categorization, parsing, video surveillance, and protein structure prediction, it is difficult and time-consuming to collect labeled data whereas unlabeled data can easily be gathered (Zhu, 2005). For speech recognition, speech can easily be recorded from radio broadcasts; for text categorization, text documents can be collected from the Internet; for parsing, sentences are everywhere; for video surveillance, surveillance cameras run continuously; and for protein structure prediction, protein sequences are readily available from gene databases (Zhu, 2005).

SSL is an appealing method in areas where labeled data is hard to collect. It has been used in areas such as text classification (Subramanya and Bilmes, 2008), text chunking (Andoy and Zhangz, 2005), document clustering (Zhong, 2006), time-series classification (Wei and Keogh, 2006), gene expression data classification (Bair and Tibshirani, 2004; Gong and Chen, 2008), visual classification (Morsillo et al., 2009), question-answering task for ranking candidate sentences (Celikyilmaz et al., 2009), and web-page classification (Liu et al., 2006). However, it has not yet been employed for the prognosis of breast cancer survivability. In survival analysis, censored data are abundant because there are many cases that patient data have not been updated along time, and hence unlabeled. Therefore SSL is a good idea since it is able to use the censored data to either modify or reprioritize the predictions on survivability obtained from labeled patient data alone. To the best of our knowledge, our paper is the first to explore the use of graph-based SSL for the survival analysis of breast cancer. The successful implementation of SSL in this domain could offer predictability of survival outcomes with reasonable accuracy and stability, relieving oncologists of the burden of data collection.

We compare up-to-date machine learning models that predict the survivability of patients diagnosed with breast cancer. Note that the prediction on survivability is predominately used for the analysis where the interest is in observing time to death of a patient, but in our study, it is dealt with as a classification problem that predicts whether the patient belongs to the group of those who survived after a specified period. We aim to find an accurate and stable classification model. Such a model would allow medical oncologists to make efficient decisions for treating breast cancer patients. Contributions of this research include its authenticity and in-depth empirical study. We used three different classification models: support vector machine (SVM), artificial neural network (ANN), and SSL. We use the surveillance, epidemiology, and end results (SEER) cancer incidence database, which is the most comprehensive source of information on cancer incidence and survival in the United States (SEER et al., 2010).

The remainder of the paper is organized as follows. Section 2 provides background information on breast cancer research into survivability analysis. Section 3 introduces SVM, ANN, and SSL which are used for the comparative analysis. Section 4 gives the details of the SEER dataset, various performance measures, and the experimental results. Finally, Section 5 presents the conclusions.

2. Background

Research on breast cancer has led to enhanced methods, and improved treatments in the form of less-invasive predictive medicine. Thus, the death rate for this cancer has decreased in recent years (Foundation et al., 2010). We now discuss the related work in the area of breast cancer survivability.

An early work can be found in (Prentice and Gloeckler, 1978). In the study, the authors applied a statistical model, called the proportional hazards regression model, to breast cancer patient data in order to discern if the patient is survived.

The authors of (Delen et al., 2005) used two popular data mining algorithms, ANNs and decision trees, together with a common statistical method, logistic regression, to develop prediction models for breast cancer survivability. This research used SEER dataset from 1973 to 2000 which consists of 433,272 records and 72 variables. The decision tree turned out to be the best predictor among them by achieving the best performance of 0.9362 in terms of classification accuracy.

An improvement in the results of decision trees for the prognosis of breast cancer survivability is described in (Khan et al., 2008). The authors propose a hybrid prognostic scheme based on weighted fuzzy decision trees (FDT). This hybrid scheme is an effective alternative to crisp classifiers that are applied independently. It analyzes the hybridization of accuracy and interpretability in terms of fuzzy logic and decision trees. They used the SEER dataset from 1973 to 2003, which consists of 162,500 records with 17 variables after preprocessing. The resulting AUC values were 0.69 for FDT, and 0.77 for weighted FDT.

In (Thongkam et al., 2008), the authors carried out data preprocessing using the RELIEF attribute selection and then used the Modest AdaBoost algorithm based on classification and regression tree to predict breast cancer survivability (Dietterich, 1997; Breiman, 1996; Rätsch et al., 2001). They used the Srinagarind hospital database. The results showed that Modest AdaBoost performs better than Real and Gentle AdaBoost.

In (Harry et al., 1997), ANN was suggested as a predictive model for cancer survivability. Two datasets were used to evaluate the performance: the Commission on Cancer's breast and colorectal carcinoma Patient Care Evaluation (PCE) dataset and SEER dataset from 1977 to 1982. ANN achieved 0.770 of AUC for the prediction of 5-year survival of patients on the PCE dataset, and 0.730 for the prediction of 10-year survival of patients on the SEER dataset. ANN was also employed by the authors in (Lundin et al., 1999), in predicting 5-, 10-, and 15- year breast cancer survivability based on 951 patients' data consisting of eight cancer incidence variables including tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age. The reported AUC values were 0.909, 0.886, and 0.883, respectively.

On the other hand, other machine learning models including SVM also have been applied to the prediction problem of breast cancer survivability. In (Thongkam et al., 2009), the authors proposed a hybrid scheme to generate a high-quality dataset from Srinagarind hospital database in Thailand in order to develop improved breast cancer survival models. The scheme has two main steps: (a) use of an outlier-filtering approach based on C-support vector classification to remove outliers from the datasets; and (b) over-sampling. The results showed that this hybrid scheme improved the performance of SVM 29.83%, 29.83%, 47.34%, and 38.59% in terms of accuracy, sensitivity, specificity, and AUC score, respectively.

In (Cruz and Wishart, 2006) the authors conducted a wide-ranging investigation of different machine learning techniques, discussing issues related to the types of data incorporated and the performance of these techniques in breast cancer prediction and prognosis. This review provides detailed explanations leading to first-rate research guidelines for the application of machine learning methods to cancer prognosis.

Most recently, a literature survey of machine learning techniques for breast cancer prognosis prediction can be found in (Kim and Shin, 2013).

3. Predictive models

Medical domain experts are not familiar with parameter selection of a specific model. Therefore, they are unable to use machine learning models to predict the outcome of a disease. In contrast, information domain experts are good at choosing models and tuning the corresponding model parameters, but they may be unable to interpret the results. One solution is to develop a model that medical domain experts can use without the risk of modeling mistakes. Technically, this is related to the robustness or stability of a model under parameter variation. If the accuracies are comparable among the candidate models, then their stabilities become important. The remainder of this section presents short descriptions of three representative classification models: SVM, ANN, and SSL together with the explanation of their implementation in our research.

3.1. Artificial neural network (ANN)

An ANN is an analytical system inspired by the structure of biological neural networks and their way of encoding and solving problems (Delen et al., 2005; Peterson and Söderberg, 1993; Abraham, 2005). We employed a well-analyzed and frequently used ANN architecture known as multi-layer perceptron with back-propagation algorithm. The ANN comprises three types of layers: the input layer, hidden layers, and the output layer. The nodes in the input layer supply input signals (activation patterns from outside the system) to the nodes in the hidden layer via weighted connections. The overall result of the model is represented by the nodes in the output layer which send output signals (a weighted sum of the signals from the hidden nodes) on the basis of a transfer function. In ANN, the accuracy of the model often depends on the structure, i.e. the number hidden nodes, and the initial weights associated with the connections between the nodes. Generally, the number of hidden nodes is selected by trial-and-error fashion and the initial weights are randomly chosen.

3.2. Support vector machine (SVM)

SVM involves finding an optimal decision boundary i.e., maximizing the margin by finding the largest achievable distance among the separating hyperplane and the data points on either side (Kotsiantis et al., 2006). The classification can be represented by considering a set of input-output pairs $D=\{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\}$, where $i=1,\ldots, \ell$. Here $x\in X$ and $y\in Y$ where 'Y' represents the set of class labels, e.g., for binary classification $Y=\{-1, +1\}$. In a typical binary classification, training data points from two different classes are separated by a hyperplane. The separating hyperplane can be linear or non-linear. For linear classification, SVM computes the linear decision function in the central gap of the two classes by correctly classifying all the training data points and placing the decision function as far from the given data points as possible, to lessen the possibility of false prediction for the unseen data points (Cardoso and Cardoso, 2007). If classes are not linearly separable because of noisy data (measurement errors, uncertainty in class membership, etc.), we can still use the linear classifier with an error tolerance. In such a case, the aim is to find a balance between margin maximization and misclassification minimization. SVM solves the following quadratic programming problem to produce the maximum margin between the two classes (Shin and Cho, 2007):

$$\begin{aligned} &\min \Theta(\vec{w}, \xi) = \tfrac{1}{2}\|\vec{w}\|^2 + C\sum_i^M \xi_i, \\ &s.t. \quad y_i(\vec{w}\cdot\Phi(\vec{x}_i) + b) \geq 1-\xi_i, \\ &\xi_i \geq 0,\ i = 1, \ldots, M. \end{aligned} \tag{1}$$

The parameter C in Eq. (1) is the penalty for misclassifying a data point. The higher the value of C, the more the SVM training is compelled to avoid classification errors (Cardoso and Cardoso, 2007). The parameter ξ_i is the non negative slack variable, which allows a certain level of misclassification for an inseparable case. If the data points are separated by a non-linear hyperplane because of some intrinsic property of the problem, it is more appropriate to map the input feature space to a high-dimensional feature space where the data points are separated by a linear hyperplane. This mapping process φ is conducted by kernel functions. Among many types of kernel functions, the RBF kernel $k(\vec{u}, \vec{v}) = e^{-\gamma|\vec{u}\ \vec{v}|^2}$ is most widely used (Schölkopf and Smola, 2002). The parameter values of the tradeoff C and the kernel width γ are specified by users, and affect the performance of SVM.

3.3. Semi-supervised learning (SSL)

In many real world classification problems, the number of class-labeled data points is small because they are often difficult, expensive, or time-consuming to acquire, requiring qualified

human annotators (Lundin et al., 1999; Shin et al., 2010; Altman and Bland, 1994). On the other hand, unlabeled data can easily be gathered and can provide valuable information for learning (He et al., 2007). However, traditional classification algorithms such as supervised-learning algorithms use only labeled data; therefore, they encounter difficulties when only a few labeled data are given. On the other hand, unsupervised learning is usually employed to discover data structure from unlabeled data; however, the main use of unsupervised learning is limited for clustering, dimensionality reduction, outlier detection, not for classification.

In SSL, the most recent category of machine learning algorithms, it allows taking advantage of the strengths of both supervised learning and unsupervised learning; meaningful representation of input data is identified from unlabeled data, and then a classification function is achieved on both labeled and the unlabeled, which is smooth with respect to the underlying input geometry(Zhu, 2005; He et al., 2007; Chapelle et al., 2006).

In SSL, the classification function is trained with a small set of labeled data $(X^l, Y^l) = \{(X_i, Y_i)_{i=1}^{n_l}\}$ and a large set of unlabeled data $(X^u) = \{(X_j)_{j=n_l+1}^{n}\}$, where $Y = \pm 1$ indicates the labels. The total number of data points is $n = n_l + n_u$ (Wang, 2007). In our study, the graph-based SSL with sharpening is used, and the following sub-sections present the details.

3.3.1. Graph-based SSL

In graph-based SSL, a weighted graph is constructed in which the nodes represent the labeled and unlabeled data points and the edges reflect the similarity between data points. According to (Zhu, 2008), graph-based SSL methods are nonparametric, discriminative, and transductive in nature. They assume label smoothness over the graph. This assumption states that if two data points are coupled by a path of high density (e.g., it is more likely that both belong to same group or cluster), then their outputs are likely to be close, whereas if they are separated by a low-density region then their outputs need not be close (Chapelle et al., 2006).

There are many graph-based SSL algorithms, e.g., mincut, Gaussian random fields and harmonic functions, local and global consistency, Tikhonov regularization, manifold regularization, graph kernels from the Laplacian spectrum, and tree-based Bayes (Zhu, 2008). There are many differences in the technical details, but in all these methods the labeled nodes are set to the labels $Y^l \in \{-1, +1\}$, the unlabeled nodes are set to zero ($Y^U = 0$), and the pairwise relationships between nodes are represented via a similarity matrix (Shin et al., 2010). Fig. 1 depicts a graph with eight data points linked by similarity between them. In our study, the patients have survived in 5 years after diagnosis are labeled as '+1' (denoted as 'S' in the figure), and '−1' (denoted as 'D') otherwise. The patients whose survivability to be predicted are unlabeled and denoted as 'u'.

Given n data points, a graph is created based on the k-nearest neighbors algorithm (kNN). Two nodes X_i and X_j are connected by an edge if X_i is in X_j's k-nearest-neighborhood. The similarity between the two nodes X_i and X_j is represented via w_{ij} in a weight matrix W. Now, a label can propagate from (labeled) node X_i to node (unlabeled) node X_j only when the value of w_{ij} is large. The value of w_{ij} can be measured using the Gaussian function (Chapelle et al., 2006):

$$w_{ij} = \begin{cases} \exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{\alpha^2}\right) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

In Eq. (2), $i \sim j$ indicates that an edge (link) can be constructed between nodes X_i and X_j by kNN, where k is a user-specified hyperparameter that controls the density of the graph.

The algorithm will output an n-dimensional real-valued vector $f = [f_i^{n_l} f_{n_l+1}^T]^T = (f_1, \ldots, f_{n_l}, f_{n_{l+1}}, \ldots, f_n)^T$, which can be thresholded to make label predictions on $f_{i=1}, \ldots, f_n$ after learning. The graph-based SSL assumes label smoothness over the graph. This assumption states that if two data points are coupled by a path of high density (e.g., it is more likely that both belong to same group or cluster), then their outputs are likely to be close, whereas if they are separated by a low-density region then their outputs need not be close (Chapelle et al., 2003). This can be rephrased as f_i of a node should not be greatly different from f_j of its adjacent nodes (smoothness condition). The other assumption is about the loss or error, which states, in labeled nodes, the value of f_i should be similar to the value of the given label y_i (loss condition). Two conditions are included in the following quadratic objective function, and one can obtain the output vector f by minimizing (Choi et al., 2008; Shin et al., 2010; Belkin et al., 2004; Chapelle et al., 2003)

$$\min_f (f - y)^T (f - y) + \mu f^T L f, \quad (3)$$

where $y = (y_1, \ldots, y_l, 0, \ldots, 0)^T$ and the matrix L, called the graph Laplacian matrix, is defined as $L = D - W$ where $D = diag(d_i)$, $d_i = \sum_i w_{ij}$. The parameter μ trades off loss and smoothness. The solution of this problem is

$$f = (I + \mu L)^{-1} y \quad (4)$$

3.3.2. Graph sharpening

The graph-sharpening scheme is an elegant method to improve the performance of graph-based SSL algorithms by removing noisy or undesirable relationships from the graph of the raw data points (Shin et al., 2007, 2010). The algorithm operates on the similarity (weight) matrix W, adjusting the connections between data points. The graph-sharpening method is based on the following observation: in an un-directed graph, as shown in Fig. 1, all the relationships are reciprocated, so matrix of edge weights is symmetric (Shin et al., 2010). However, we know that W represents connections between labeled and unlabeled nodes. It is not desirable to consider all such edges (relationships) as symmetric because some edges may express more valuable information in one direction than in the other directions (Shin et al., 2010). Therefore, the algorithm to improve the performance of graph-based SSL is as follows:

- First, the algorithm penalizes the information flow from unlabeled to labeled points because this may affect the information in the system since such points hold uncertain information.
- Second, the algorithm disconnects the edges directly linked to oppositely labeled points because they may possibly transmit unconstructive information.
- Third, the spread of information between unlabeled points is different, so the algorithm allows the unlabeled nodes to synchronize with their neighbors.

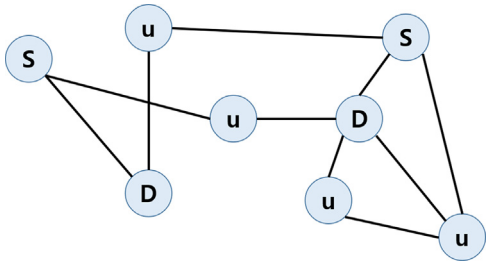

Fig. 1. Graph-based SSL: labeled nodes are represented by 'S' (Survival) and 'D' (Death), and unlabeled nodes are represented by 'u'.

The procedure is called *sharpening* and the resulting graph is called a *sharpened graph*. A sharpened graph is a directed graph. Fig. 2 depicts sharpening results on the original SSL graph in Fig. 1.

4. Experiments

4.1. Dataset

We used the SEER datasets. It is an initiative of the National Cancer Institute and is the premier source for cancer statistics in the United States. SEER claims to have one of the most comprehensive collections of cancer statistics. It includes, but is not limited to, incidence, mortality, prevalence, survival, lifetime risk, and statistics by race/ethnicity. The dataset can be requested online from the SEER website (http://www.seer.cancer.gov) (SEER et al., 2010). The SEER datasets have been used by many researchers. We used the breast cancer survivability dataset (1973–2003). It consists of 162,500 records with 16 predictor features and one target class variable. There are 16 features: tumor size, number of nodes, number of primaries, age at diagnosis, number of positive nodes, marital status, race, behavior code, grade, extension of tumor, node involvement, histologicalTypeICD, primary site, site specific surgery, radiation, and stage. The target variable "survivability" of SEER dataset is a binary categorical feature with values '−1' (if the patient had not survived longer than 5 years after diagnosis) or +1 (had survived).

Table 1 summarizes the features and their descriptions.

4.2. Measures of performance analysis

Sensitivity, specificity, accuracy, and the area under the ROC curve (AUC) are to measure the prediction accuracy on the basis of the entries of the confusion matrix that contains information about the actual and predicted classification (Thongkam et al., 2008). Sensitivity is defined as the proportion of true positives that are correctly identified by the classifier, which is $TP/(TP+FN)$, where TP and FN stand for the number of correct predictions and that of incorrect predictions, respectively, when a data point actually belongs to the positive class. In our study, the positive class indicates the group of the survived patients. Specificity is defined as the proportion of true negatives that are correctly identified by the classifier, which is $TN/(TN+FP)$, where TN and FP indicate the number of correct predictions and that of incorrect predictions, respectively, when a data point actually belongs to the negative class. Using sensitivity and specificity, we try to find what proportion of patients with abnormal test results is truly abnormal (Altman and Bland, 1994). To assess the overall value of a classifier, accuracy and AUC are used. Accuracy is a measure of the total number of correct predictions, which is defined as $(TP+TN)/(TP+FN+TN+FP)$ when the value of classification-threshold is set to 0. On the other hand, AUC is a threshold-independent measure off model performance based on the receiver operating characteristic (ROC) curve which plots the tradeoffs between sensitivity and 1-specificity for all possible values of threshold (Allouche et al., 2006).

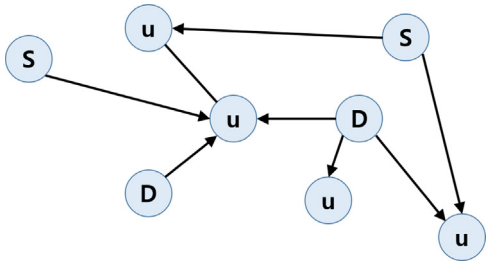

Fig. 2. Sharpened graph: information flows only from labeled nodes, S (Survival) or D (Death), to unlabeled nodes u, but influence between the nodes u is regarded as reciprocated.

4.3. Experimental setting

Three representative models, ANN, SVM, and SSL, are used to perform classification on breast cancer survivability. We evaluated these models on the basis of the aforementioned performance measures: accuracy, sensitivity, specificity, and AUC. The breast cancer survival dataset consisted of 162,500 data points: 128,469 positive cases and 34,031 negative cases. The dataset is large and class-imbalanced. The large-sized dataset imposes computational burden on most learning algorithms in training time and memory. Classification on imbalanced datasets, on the other hand, usually causes problems on biased accuracy to the overwhelmed class in size. To avoid the addressed difficulties, 40,000 data points for the training set and 10,000 for the test set are randomly drawn from the original dataset without replacement, and both sets are class-balanced to have the same proportion of positive and negative classes. The positive class and the negative class are composed of 25,000 data points randomly drawn from the 128,469 positive cases and 34,031 negative cases, respectively. Note that the test set is regarded as unlabeled data for SSL. The equipoise dataset of 50,000 data points is eventually divided into 10 groups. And for each set, five-fold cross validation is used and repeated five times. Fig. 3 shows the experimental setting.

Table 1
Prognostic elements of breast cancer survivability.

No.	Features	Description
1	Stage	Defined by size of cancer tumor and its spread
2	Grade	Appearance of tumor and its similarity to more- or less-aggressive tumors
3	Lymph node involvement	None, (1–3) minimal, (4–9) significant, etc
4	Race	Ethnicity: White, Black, Chinese, etc.
5	Age at diagnosis	Actual age of patient in years
6	Marital status	Married, single, divorced, widowed, separated
7	Primary site	Presence of tumor at particular location in body. Topographical classification of cancer
8	Tumor size	2–5 cm; at 5 cm prognosis worsens
9	Site-specific surgery	Information on surgery during first course of therapy, whether cancer-directed or not
10	Radiation	None, beam radiation, radioisotopes, refused, recommended, etc
11	Histological type	Form and structure of tumor
12	Behavior code	Normal or aggressive tumor behavior is defined using codes
13	Number of positive nodes examined	When lymph nodes are involved in cancer, they are called positive
14	Number of nodes examined	Total nodes (positive/negative) examined
15	Number of primaries	Number of primary tumors (1–6)
16	Clinical extension of tumor	Defines spread of tumor relative to breast
17	Survivability	Target binary variable defines class of survival of patient: '+1' if survived longer than five years, '−1' otherwise

The performance was measured under various combinations of model parameters. The ranges of the user-specified values are set as follows.

ANN

Random Seed={1, 3, 5, 7, 10}
Hidden Node={3, 6, 9, 12, 15}

SVM

C={0.2, 0.4, 0.6, 0.8, 1}
Gamma={0.0001, 0.001, 0.01, 0.1, 1}

SSL

Mu={0.0001, 0.01, 1, 100, 1000}
K={3, 7, 15, 20, 30}.

The number of 'hidden nodes' and the random seed number for 'initial weights' are user-specified parameters of ANN. There are some empirically derived rules-of-thumbs about how to set the number of hidden nodes, of these, the most commonly relied on is 'the optimal number of the hidden nodes is usually between the number of the input nodes and the number of the output nodes.' Since the dataset contains 16 input features and one output feature, the ranges of the number of hidden nodes are set within the bounds. For SVM, 'Gamma' and 'C' are determined by users where the former is the RBF kernel width and latter is the penalty for misclassifying a data-point. In order to find the relevant ranges of both parameters, a preliminary experiment with the broader ranges of C={0.1, 1, 10, 100, 1000} and Gamma={0.1, 1, 10, 100, 1000} was conducted *a priori*. Reasonable performances were obtained from the ranges of 0.1–1 for C and less than 1 for Gamma. Therefore, the ranges were narrowed and employed for comparison as above. Appendix A presents the details of the experimental results. Likewise, the user defined parameters for SSL are 'k' and 'Mu', where 'k' is the number of neighbors and 'Mu' is the tradeoff between smoothness condition and loss condition. It is difficult to

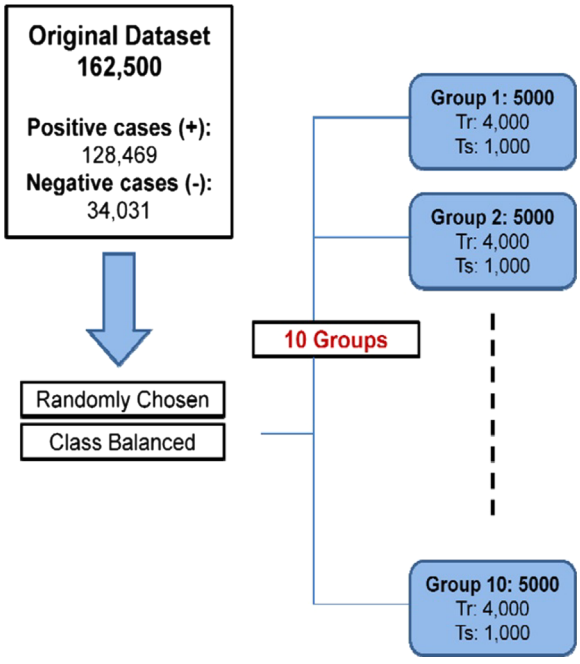

Fig. 3. Experimental setting.

Table 2
Best performance comparison: ANN, SVM, and SSL (the values of model parameters are presented in Appendix B Table B1).

Dataset	ANN				SVM				SSL			
	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
1	0.66	0.8	0.51	**0.68**	0.52	0.77	0.51	**0.79**	0.72	0.73	0.7	**0.78**
2	0.67	0.71	0.64	**0.72**	0.52	0.47	0.51	**0.79**	0.72	0.78	0.66	**0.79**
3	0.62	0.53	0.71	**0.68**	0.5	0.5	0.58	**0.8**	0.7	0.81	0.59	**0.78**
4	0.67	0.7	0.64	**0.72**	0.51	0.71	0.51	**0.79**	0.68	0.73	0.63	**0.76**
5	0.64	0.72	0.56	**0.66**	0.52	0.47	0.51	**0.82**	0.71	0.74	0.68	**0.78**
6	0.62	0.65	0.6	**0.68**	0.52	0.71	0.51	**0.78**	0.71	0.74	0.69	**0.77**
7	0.63	0.83	0.43	**0.67**	0.51	0.72	0.51	**0.79**	0.69	0.76	0.62	**0.77**
8	0.69	0.85	0.53	**0.73**	0.51	0.76	0.51	**0.82**	0.73	0.78	0.67	**0.8**
9	0.66	0.6	0.71	**0.71**	0.52	0.79	0.51	**0.81**	0.7	0.74	0.66	**0.79**
10	0.64	0.86	0.42	**0.73**	0.51	0.63	0.5	**0.81**	0.72	0.78	0.66	**0.8**
Avg. (±std)	**0.65 (±0.02)**	**0.73 (±0.11)**	**0.58 (±0.10)**	**0.70 (±0.03)**	**0.51 (±0.01)**	**0.65 (±0.13)**	**0.52 (±0.02)**	**0.80 (±0.01)**	**0.71 (±0.02)**	**0.76 (±0.03)**	**0.65 (±0.03)**	**0.78 (±0.01)**

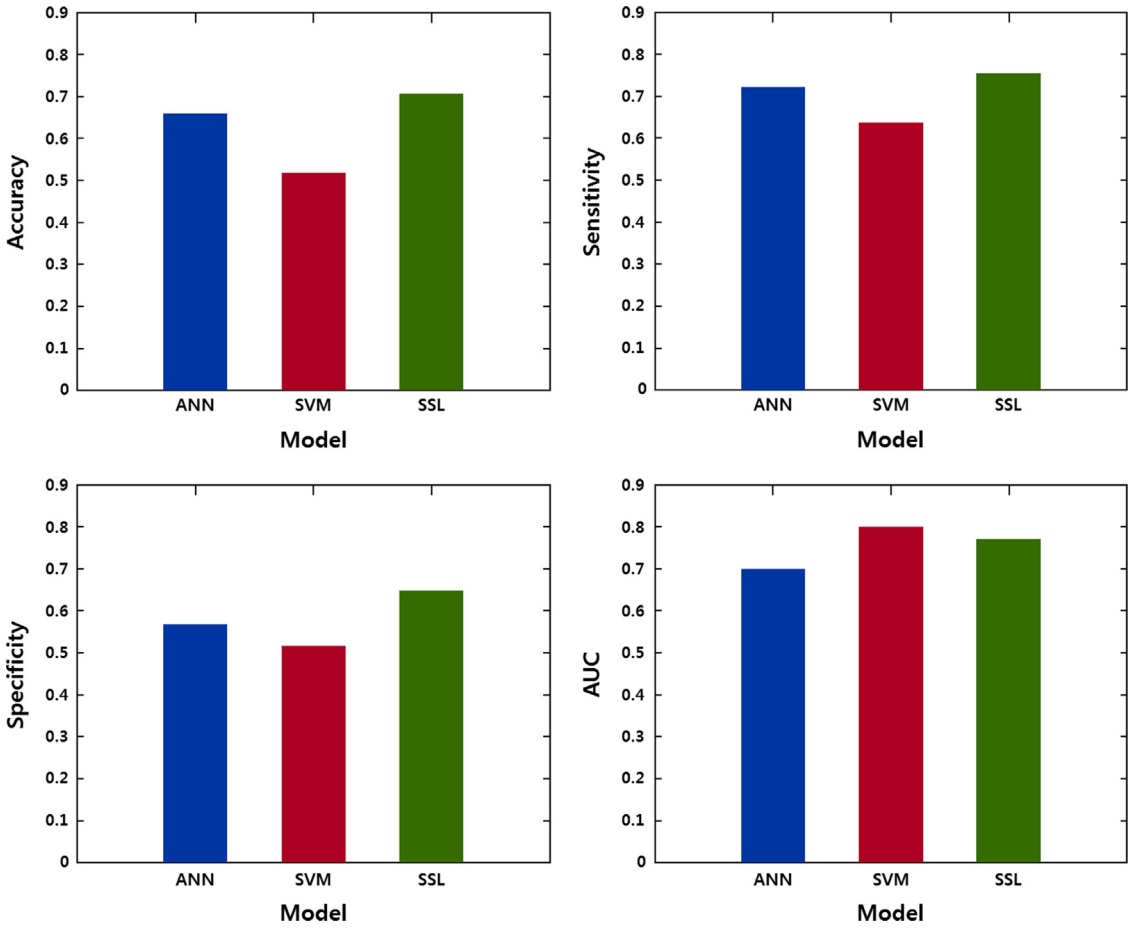

Fig. 4. Best performance comparison among ANN, SVM, and SSL in terms of accuracy, sensitivity, specificity, and AUC.

Table 3
Average AUC for different parameter combinations: ANN, SVM, and SSL.

ANN		**Random seeding**				
		1	**3**	**5**	**7**	**10**
Hidden node	**3**	0.48 ± 0.06	0.58 ± 0.06	0.57 ± 0.06	0.55 ± 0.07	0.57 ± 0.06
	6	0.52 ± 0.07	0.55 ± 0.06	0.59 ± 0.04	0.56 ± 0.06	0.58 ± 0.07
	9	0.52 ± 0.09	0.58 ± 0.09	0.60 ± 0.05	0.58 ± 0.05	0.60 ± 0.07
	12	0.57 ± 0.09	0.57 ± 0.08	0.58 ± 0.07	0.55 ± 0.06	0.60 ± 0.05
	15	0.54 ± 0.09	0.59 ± 0.09	0.57 ± 0.06	0.57 ± 0.05	0.61 ± 0.09
SVM		**C**				
		0.2	**0.4**	**0.6**	**0.8**	**1**
Gamma	**0.0001**	0.79 ± 0.01	0.80 ± 0.01	0.80 ± 0.01	0.80 ± 0.01	0.80 ± 0.01
	0.001	0.79 ± 0.01	0.79 ± 0.01	0.79 ± 0.01	0.79 ± 0.02	0.79 ± 0.02
	0.01	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.74 ± 0.01
	0.1	0.67 ± 0.02	0.67 ± 0.02	0.67 ± 0.02	0.67 ± 0.02	0.66 ± 0.02
	1	0.45 ± 0.04	0.45 ± 0.04	0.44 ± 0.04	0.44 ± 0.02	0.46 ± 0.03
SSL		**Mu**				
		0.0001	**0.01**	**1**	**100**	**1000**
K	**3**	0.71 ± 0.01	0.71 ± 0.01	0.71 ± 0.01	0.72 ± 0.01	0.72 ± 0.01
	7	0.75 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	0.76 ± 0.01	0.76 ± 0.01
	15	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
	20	0.77 ± 0.01	0.77 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01
	30	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01	0.78 ± 0.01

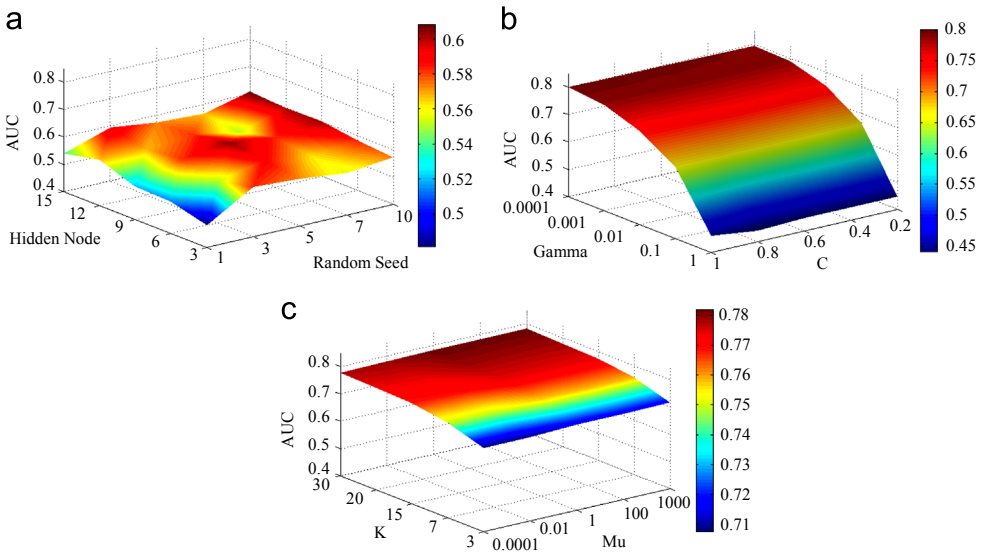

Fig. 5. Performance stability under model parameter variation: (a), (b), and (c) represent the AUC changes for parameter changes in ANN, SVM, and SSL, respectively.

Table 4
Expected AUC values without model tuning.

Dataset	ANN Avg_AUC	SVM Avg_AUC	SSL Avg_AUC	Friedman test (*p*-values)
1	0.59	0.68	0.76	5.14e−8
2	0.56	0.69	0.77	1.96e−7
3	0.55	0.68	0.75	1.82e−8
4	0.56	0.68	0.75	1.82e−7
5	0.56	0.70	0.77	7.99e−8
6	0.54	0.71	0.75	1.21e−8
7	0.57	0.67	0.75	1.23e−7
8	0.58	0.69	0.78	1.08e−7
9	0.56	0.70	0.76	1.79e−7
10	0.59	0.71	0.76	2.20e−7
Mean (±std)	**0.57 (±0.07)**	**0.69 (±0.13)**	**0.76 (±0.03)**	**1.38e−6**

find a known knowledge on the ranges where the optimal values of the parameters exist, the search space is ranged in a broad scale varying from 0.0001 to 1000 for 'Mu' and from 3 to 30 for 'K'. In the experiment, ANN and SVM were implemented using standard Matlab toolboxes (http://cogsys.imm.dtu.dk/toolbox/ann/index.html; http://sourceforge.net/projects/svm/), and for SSL, Matlab codes for (3) and (4) in Section 3.3.2 were implemented based on an optimization toolbox (http://www.alphaminers.net).

4.4. Results

Table 2 shows the results for the best parameter combination out of the 25 experiments for each of 10 datasets. In terms of accuracy, specificity, and sensitivity, the best performance was achieved by the SSL, with a mean accuracy of 0.71, a mean sensitivity of 0.76, and a mean specificity of 0.65. However, the SVM model had the best AUC performance with a mean of 0.80. ANN, which had the lowest AUC of the three models, had better performance than SVM in accuracy, specificity, and sensitivity. Thus, the comparison of the models depends on the chosen measure. Fig. 4 compares the best performance of the three models with respect to each of the four measures.

To find the right parameter values for a model, however, it requires many experimental repetitions and the technical

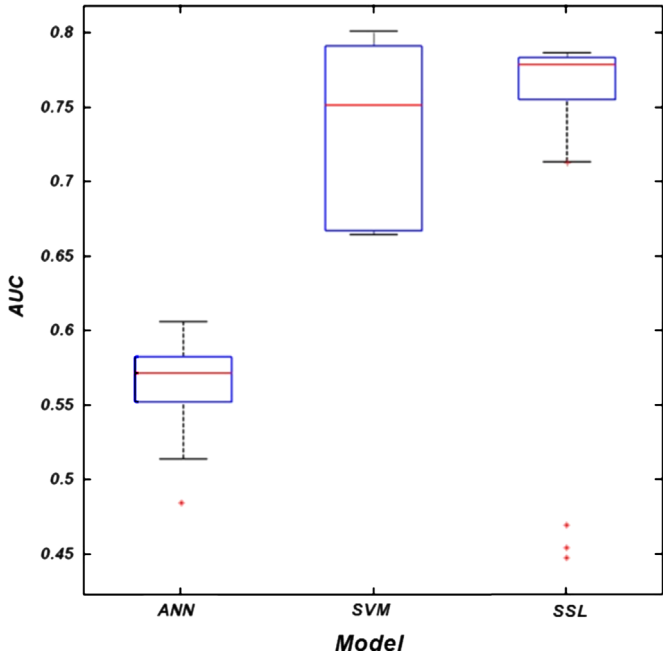

Fig. 6. Box-whisker-plot for performance variation across 25 combinations of model parameters presented in Table 3.

expertise for model-tuning. Thus, it is not easy to find the best parameter combination. Since the recent models have similar performance, we can consider the robustness of the models to parameter variation. If the performance of a model is not so sensitive to parameter variation, then the loss that may be incurred by layperson's choice for parameter values will be minimized, which may be a possible scenario that can apply to most interdisciplinary domain with informatics.

Table 3 shows the average AUC for different parameter combinations for ANN, SVM, and SSL, and Fig. 5 shows the AUC differences graphically. From the table and the figure, we see that SSL is stable under parameter variation at a fairly high AUC value, but SVM and ANN are highly sensitive to parameter variation. About the remark on stability over parameter variation, Appendix C presents additional experimental results conducted on other datasets.

Table 4 gives the expected AUC value of each model in the case where the user has little knowledge of model tuning and model parameter selection, so any values of the parameters might be chosen. The expected AUCs of ANN, SVM, and SSL are 0.57 ($\pm$0.07), 0.69 ($\pm$0.13), and 0.76 ($\pm$0.03), respectively. The Friedman test was used to validate the significance of differences in performance across the three models (Sheldon et al., 1996). The last column of the table shows the resulting p-values for the 10 datasets, which indicates the differences among the three models are statistically significant. The box-whisker-plot in Fig. 6 depicts the mean values and the standard deviations across 25 combinations of model parameters presented in Table 3. A smaller box area indicates more robustness or stability under parameter variation. The plot shows that SSL has smaller boxes with higher values of AUC, which makes them more accurate and stable (or robust) than the other models.

It is worth observing that some models with a higher accuracy than others may be heavily dependent on parameter selection. For instance, the performance of SVM in our experiment tends to depend on its two model parameters, the kernel parameter Gamma and the regularization coefficient C. This implies that the model selection should be carried out carefully by modeling experts. On the other hand, SSL has reasonably good accuracy and are stable under parameter variation, which makes the model selection easy and safe.

5. Conclusion

We have presented a comparison of various machine learning techniques for breast cancer prognosis analysis. We suggested using a model that bridges the domains of information and medicine to aid medical specialists and information experts in parameter selection when building a predictive model for medical domain. In practice, however, medical practitioners attempting IT work are expected to be either well versed with the modeling technique used or work closely with IT specialists. In such a viewpoint, our study may be regarded as overemphasizing the benefits of robustness in allowing laypersons to build good models which are not sensitive to modeling parameters. However, our study is not only be of benefit to laypersons but also IT specialists since in some learning problems, even simple parametric models are not sufficiently robust to provide accurate descriptions for input data or domain problems. There have been many such researches that guarantee more robust behavior of model under changes in parameter and input distribution (Bagnell, 2005; Hernandez-Lobato et al., 2008; Provost and Fawcett, 2001; Ramoni and Sebastiani, 2001).

In this paper, we attempted to discover the robust predictive model among well-known machine learning algorithms for breast cancer survivability. We compared three models: SVM, ANN, and SSL. And we suggested that the SSL model can be a candidate that would allow medical professionals or IT specialists to efficiently employ for survival analysis without needing to select parameters for a specific model. Appendix C presents additional experimental results on this suggestion.

The firsthand exploitation of SSL for the prognosis of breast cancer survivability is another attractive feature of our research. We used the SEER survival data for our experiments. The results showed that SVM performed reasonably well when the model parameters were carefully tuned, but performance can fluctuate

Table B1
Model parameters for best performance comparison in Table 2.

Dataset	ANN		SVM		SSL	
	Random seeding	Hidden node	C	Gamma	Mu	K
1	3	6	0.4	0.001	30	0.01
2	5	9	0.6	0.001	15	0.01
3	5	3	0.8	0.01	15	1
4	7	9	0.4	0.0001	20	100
5	10	9	0.4	0.001	30	0.01
6	5	12	1	0.01	15	1000
7	3	3	0.2	0.001	20	100
8	1	3	0.4	0.01	30	0.01
9	7	5	0.6	0.0001	30	1
10	10	9	0.8	0.0001	20	100

Table A1
Average AUC of SVM for different parameter combinations: broad ranges.

SVM	C				
	0.1	1	10	100	1000
Gamma					
0.1	**0.67 $\pm$ 0.02**	**0.66 $\pm$ 0.02**	0.66 $\pm$ 0.02	0.66 $\pm$ 0.01	0.66 $\pm$ 0.01
1	**0.45 $\pm$ 0.04**	**0.46 $\pm$ 0.03**	0.47 $\pm$ 0.03	0.47 $\pm$ 0.03	0.47 $\pm$ 0.03
10	0.44 $\pm$ 0.24	0.43 $\pm$ 0.01	0.44 $\pm$ 0.01	0.44 $\pm$ 0.01	0.44 $\pm$ 0.01
100	0.44 $\pm$ 0.04	0.43 $\pm$ 0.04	0.41 $\pm$ 0.02	0.40 $\pm$ 0.00	0.40 $\pm$ 0.00
1000	0.40 $\pm$ 0.00	0.40 $\pm$ 0.00	0.40 $\pm$ 0.00	0.40 $\pm$ 0.00	0.40 $\pm$ 0.00

Table A2
Average AUC of SVM for different parameter combinations: narrowed ranges.

SVM	C				
	0.2	0.4	0.6	0.8	1
Gamma					
0.0001	0.79 $\pm$ 0.01	0.80 $\pm$ 0.01	0.80 $\pm$ 0.01	0.80 $\pm$ 0.01	0.80 $\pm$ 0.01
0.001	0.79 $\pm$ 0.01	0.79 $\pm$ 0.01	0.79 $\pm$ 0.01	0.79 $\pm$ 0.02	0.79 $\pm$ 0.02
0.01	0.75 $\pm$ 0.01	0.75 $\pm$ 0.01	0.75 $\pm$ 0.01	0.75 $\pm$ 0.01	0.74 $\pm$ 0.01
0.1	0.67 $\pm$ 0.02	0.67 $\pm$ 0.02	0.67 $\pm$ 0.02	0.67 $\pm$ 0.02	0.66 $\pm$ 0.02
1	0.45 $\pm$ 0.04	0.45 $\pm$ 0.04	0.44 $\pm$ 0.04	0.44 $\pm$ 0.02	0.46 $\pm$ 0.03

depending on the choice of the parameters. ANN, on the other hand, did not perform well in terms of accuracy and stability. The best performance was obtained from SSL. They showed good overall accuracy and were stable under model parameter variation. The sharpening procedure seemed to enhance the stability of SSL by the noise-reduction function of the algorithm. The advantage of SSL is that the model extracts more information about the data by incorporating unlabeled data, usually discarded in other models. Finally, we expect that our findings will be helpful to medical clinicians and IT specialists when informed decision-making is required for more accurate and less invasive prognosis of breast cancer survivability.

Table C1
Result comparison from Wisconsin breast cancer data.

ANN		Random seeding					Avg(± std) [*p*-value]
		1	**3**	**5**	**7**	**10**	
Hidden node	**3**	0.760 ± 0.09	0.839 ± 0.06	0.740 ± 0.06	0.791 ± 0.06	0.755 ± 0.06	0.919 (± 0.074) [0.000]
	6	0.957 ± 0.06	0.957 ± 0.06	0.957 ± 0.08	0.957 ± 0.08	0.940 ± 0.08	
	9	**0.956** ± 0.05	0.954 ± 0.06	0.955 ± 0.07	0.956 ± 0.06	0.955 ± 0.09	
	12	0.954 ± 0.06	0.954 ± 0.05	0.954 ± 0.04	0.954 ± 0.05	0.955 ± 0.08	
	15	0.954 ± 0.09	0.954 ± 0.09	0.953 ± 0.06	0.954 ± 0.07	0.953 ± 0.09	
SVM		**C**					Avg(± std) [*p*-value]
		0.2	**0.4**	**0.6**	**0.8**	**1**	
Gamma	**0.0001**	0.935 ± 0.01	0.936 ± 0.01	0.938 ± 0.01	0.940 ± 0.01	0.939 ± 0.01	0.816 (± 0.188) [0.000]
	0.001	**0.942** ± 0.01	0.910 ± 0.01	0.906 ± 0.02	0.901 ± 0.02	0.900 ± 0.02	
	0.01	0.922 ± 0.01	0.922 ± 0.01	0.922 ± 0.02	0.922 ± 0.02	0.923 ± 0.02	
	0.1	0.848 ± 0.03	0.848 ± 0.01	0.851 ± 0.02	0.871 ± 0.02	0.871 ± 0.02	
	1	0.452 ± 0.04	0.452 ± 0.03	0.452 ± 0.04	0.452 ± 0.03	0.452 ± 0.03	
SSL		**Mu**					Avg(± std)
		0.0001	**0.01**	**1**	**100**	**1000**	
K	**3**	0.957 ± 0.01	0.958 ± 0.01	0.963 ± 0.01	0.964 ± 0.01	0.964 ± 0.01	0.969 (± 0.004)
	7	0.972 ± 0.01	0.972 ± 0.01	0.970 ± 0.01	0.968 ± 0.01	0.968 ± 0.01	
	15	0.971 ± 0.01	0.971 ± 0.01	0.970 ± 0.01	0.969 ± 0.01	0.969 ± 0.01	
	20	0.971 ± 0.01	0.970 ± 0.01	0.969 ± 0.01	0.969 ± 0.01	0.969 ± 0.01	
	30	**0.973** ± 0.01	0.973 ± 0.01	0.972 ± 0.01	0.972 ± 0.01	0.972 ± 0.01	

Table C2
Result comparison from Heart Disease data.

ANN		Random seeding					Avg(± std) [*p*-value]
		1	**3**	**5**	**7**	**10**	
Hidden node	**3**	0.522 ± 0.11	0.519 ± 0.09	0.513 ± 0.09	0.496 ± 0.12	0.557 ± 0.09	0.692(± 0.114) [0.000]
	6	0.625 ± 0.06	0.663 ± 0.04	0.644 ± 0.05	0.625 ± 0.06	0.590 ± 0.07	
	9	0.657 ± 0.07	0.704 ± 0.05	0.703 ± 0.07	0.752 ± 0.08	0.689 ± 0.09	
	12	0.722 ± 0.03	0.793 ± 0.04	0.806 ± 0.04	0.802 ± 0.06	0.756 ± 0.05	
	15	0.833 ± 0.06	0.809 ± 0.05	0.821 ± 0.06	**0.854** ± 0.07	**0.833** ± 0.07	
SVM		**C**					Avg(± std) [*p*-value]
		0.2	**0.4**	**0.6**	**0.8**	**1**	
Gamma	**0.0001**	0.763 ± 0.01	0.761 ± 0.01	0.763 ± 0.01	**0.766** ± 0.01	**0.769** ± 0.01	0.623(± 0.107) [0.000]
	0.001	0.722 ± 0.02	0.729 ± 0.01	0.727 ± 0.01	0.726 ± 0.01	0.730 ± 0.02	
	0.01	0.589 ± 0.03	0.589 ± 0.02	0.590 ± 0.02	0.591 ± 0.02	0.596 ± 0.02	
	0.1	0.525 ± 0.03	0.525 ± 0.02	0.525 ± 0.02	0.525 ± 0.02	0.525 ± 0.02	
	1	0.497 ± 0.04	0.497 ± 0.04	0.509 ± 0.03	0.509 ± 0.03	0.528 ± 0.03	
SSL		**Mu**					Avg(± std)
		0.0001	**0.01**	**1**	**100**	**1000**	
K	**3**	0.790 ± 0.01	0.791 ± 0.01	0.795 ± 0.01	0.802 ± 0.01	0.802 ± 0.01	00.846(± 0.033)
	7	0.822 ± 0.01	0.821 ± 0.01	0.831 ± 0.01	0.830 ± 0.01	0.830 ± 0.01	
	15	0.852 ± 0.01	0.850 ± 0.01	0.857 ± 0.01	0.857 ± 0.01	0.857 ± 0.01	
	20	0.865 ± 0.01	0.866 ± 0.01	0.866 ± 0.01	0.869 ± 0.01	0.869 ± 0.01	
	30	0.886 ± 0.01	0.886 ± 0.01	0.887 ± 0.01	**0.889** ± 0.01	0.889 ± 0.01	

Table C3
Result comparison from Spect data.

ANN		Random seeding					Avg (± std) [*p*-value]
		1	**3**	**5**	**7**	**10**	
Hidden node	**3**	0.588 ± 0.10	0.573 ± 0.06	0.604 ± 0.06	0.595 ± 0.06	0.595 ± 0.09	0.669 (± 0.043) [0.000]
	6	0.661 ± 0.09	0.672 ± 0.06	0.665 ± 0.06	0.657 ± 0.07	0.678 ± 0.08	
	9	0.685 ± 0.07	0.690 ± 0.06	0.689 ± 0.07	0.680 ± 0.07	0.689 ± 0.08	
	12	0.696 ± 0.07	0.702 ± 0.06	0.698 ± 0.08	0.692 ± 0.09	0.686 ± 0.07	
	15	**0.711** ± 0.06	**0.713** ± 0.06	0.701 ± 0.06	0.707 ± 0.09	0.708 ± 0.09	
SVM		**C**					Avg(± std) [*p*-value]
		0.2	**0.4**	**0.6**	**0.8**	**1**	
Gamma	**0.0001**	**0.745** ± 0.01	**0.739** ± 0.01	0.726 ± 0.01	0.726 ± 0.01	0.740 ± 0.01	0.729 (± 0.009) [0.000]
	0.001	0.726 ± 0.01	0.726 ± 0.01	0.731 ± 0.02	0.731 ± 0.01	0.732 ± 0.02	
	0.01	0.732 ± 0.01	0.732 ± 0.01	0.732 ± 0.03	0.732 ± 0.02	0.732 ± 0.02	
	0.1	0.735 ± 0.02	0.735 ± 0.02	0.735 ± 0.03	0.737 ± 0.04	0.741 ± 0.03	
	1	0.715 ± 0.03	0.713 ± 0.02	0.715 ± 0.03	0.715 ± 0.04	0.714 ± 0.04	
SSL		**Mu**					Avg (± std)
		0.0001	**0.01**	**1**	**100**	**1000**	
K	**3**	0.732 ± 0.02	0.731 ± 0.01	0.735 ± 0.01	0.744 ± 0.01	0.744 ± 0.01	0.783 (± 0.030)
	7	0.760 ± 0.02	0.760 ± 0.02	0.763 ± 0.02	0.784 ± 0.02	0.784 ± 0.02	
	15	0.821 ± 0.01	0.821 ± 0.01	0.815 ± 0.01	0.831 ± 0.01	**0.833** ± 0.01	
	20	0.791 ± 0.02	0.788 ± 0.01	0.791 ± 0.01	0.795 ± 0.01	0.794 ± 0.01	
	30	0.778 ± 0.01	0.777 ± 0.01	0.794 ± 0.01	0.801 ± 0.01	0.801 ± 0.01	

Acknowledgment

H.S. would like to gratefully acknowledge support from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013R1A1A3010440/2010-0028631)

Appendix A

For SVM, 'Gamma' and 'C' are determined by users where the former is the RBF kernel width and the latter is the penalty for misclassifying a data-point. In order to find the relevant ranges of both parameters, a preliminary experiment with the broader ranges of $C=\{0.1, 1, 10, 100, 1000\}$ and Gamma$=\{0.1, 1, 10, 100, 1000\}$ was conducted *a priori*. The following two tables, Table A1 and A2, present the details of the experimental results. From Appendix C Table C1, reasonable performances were obtained from the ranges of 0.1–1 for C and less than 1 for Gamma (the shaded cells in the upper-left corner). Therefore, the resulting ranges were narrowed to $C=\{0.2, 0.4, 0.6, 0.8, 1\}$ and Gamma$=\{0.0001,0.001, 0.01, 0.1, 1\}$, and employed for comparison.

Appendix B

The following table provides the optimal values of the model parameters presented in Tables B1.

Appendix C

Additional experiments were conducted to show that and SSL are stable under parameter variation at a reasonably high performance. The following Appendix C Tables C1–C3 show the average AUC for different parameter combinations for ANN, SVM, and SSL, for Wisconsin Breast Cancer, Heart Disease, and Spect datasets, respectively (available from UCI Machine Learning Repository, http://archive.ics.uci.edu/ml). Wisconsin Breast Cancer dataset is to discern malignant (cancerous) examples from benign (non-cancerous) examples, and Heart Disease dataset is to predict absence or presence of heart disease. Spect dataset is on diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each example of the patients is classified into two categories: normal and abnormal. The pairwise t-test was used to validate the significance of differences in performance between SSL and each of the remaining models. The numbers in the bracket in the last column of the table show the p-values. For the three datasets, SSL outperformed ANN and SVM, and the differences in performance were statistically significant.

References

Abraham A., 2005. Artificial neural networks. In: Sydenham, P., Thorn, R. (Eds.), Handbook for Measurement Systems Design, London.
Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223–1232.
Altman, D., Bland, M., 1994. Diagnostic tests; 1: sensitivity and specificity. Br. Med. J. 308, 1552. (1552).
Amir, E., Evans, D.G.R., Shenton, A, Lalloo, F, Moran, A, Boggis, C, et al., 2003. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J. Med. Genetics 40, 807–814.
Andoy R.K., Zhangz T., 2005. A high-performance semi-supervised learning method for text chunking. In: Knight, K., Ng, H.T., Oflazer, K., (Eds.), The Proceedings of the Forty-Third Annual Meeting on Association for Computational Linguistics Ann Arbor, Michigan, pp. 1–9.
Bagnell, J.A., 2005. Robust Supervised Learning. American Association for Artificial Inteligence.
Bair, E., Tibshirani, R., 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2, 0511–0522.
Belkin M., Matveeva I., Niyogi P., 2004. Regularization and Semi-supervised Learning on Large Graphs. In: Lecture Notes in Computer Science, vol. 3120, Springer, , pp. 624–638.
Breiman, L., 1996. Stacked regressions. Machine Learning 24 (1), 49–64.
Brenner, H., Gefeller, O., Hakulinen, T., 2002. A computer program for period analysis of cancer patient survival. Eur. J. Cancer 38, 690–695.
Cancer Facts & Figures, 2010. American Cancer Society. Atlanta, 2010.
Cardoso, J.S., Cardoso, M.J., 2007. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artif. Intell. Med. 40, 115–126.

Celikyilmaz A., Thint M., Huang Z., 2009. A graph-based semi-supervised learning for question-answering. In: The Proceedings of the Forty-Seventh Annual Meeting of Annual Meeting of the Association for Computational Linguistics Singapore, pp. 719–727.
Chapelle O., Weston J., Schölkopf B., 2003. Cluster kernels for semi-supervised learning. In: Advances in Neural Information Processing Systems, The MIT Press, Cambridge, England, pp. 585–592.
Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-supervised learning. The MIT Press, Cambridge, England, pp. 3–14.
Choi, I., Park, K., Shin, H., 2008. Sharpened graph ensemble for semi-supervised learning. Intell. Data Anal. 17, 387–398.
Cruz, J.A., Wishart, D.S., 2006. Applications of machine learning in cancer prediction and prognosis. Cancer Inf. 2, 59–78.
Delen, D., Walker, G., Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. 34, 113–127.
Dietterich, Thomas G., 1997. Machine-learning research. AI Magazine 18 (4), 97–136.
Foundation N.B.C., 2010. What is Breast Cancer?, National Breast Cancer Foundation, Inc ⟨http://www.nationalbreastcancer.org/?aspxerrorpath=/about-breast-cancer/what-is-breast%20cancer.aspx⟩ (Accessed: 11 July 2011).
Gong Y.C., Chen C.L., 2008. Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions. In: The Proceedings of Nineteenth International Conference on Pattern Recognition Tampa, ;FL, pp. 1–4.
Harry, B., Phillip, H., David, B., Donald, E., John, N., Frank, E., Jeffery, R., David, P., David, G., 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Am. Cancer Soc. 79 (4), 857–862.
He J., Carbonell J., Liu Y., 2007. Graph-based semi-supervised learning as a generative model. In: Veloso, M.M. (Ed.), The Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 2492–2497.
Hernandez-Lobato, J.M., Dijkstra, T., Heskes, T, 2008. Regulator discovery from gene expression time series of malaria parasites: a hierachical approach. Adv. Inf. Process. Syst. 20, 649–656.
Khan U., Shin H., Choi J.P., Kim M., 2008. wFDT – weighted fuzzy decision trees for prognosis of breast cancer survivability. In: Roddick, J.F., Li, J., Christen, P., Kennedy, P.J. (Eds.), The Proceedings of the Seventh Australasian Data Mining Conference Glenelg, South Australia, pp. 141–152.
Kim, J., Shin, H., 2013. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. J. Am. Med. Inf. Assoc. 20 (4), 613–618.
Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., 2006. Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26, 159–190.
Liu R., Zhou J., Liu M., 2006. Graph-based semi-supervised learning algorithm for page classification. In: The Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications China, IEEE Computer Society, pp. 856–860.
Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkänen, L., Joensuu, H., 1999. Artificial neural networks applied to survival prediction in breast cancer. Oncology 57, 281–286.
Morsillo N., Pal C., Nelson R., 2009. Semi-supervised learning of visual classifiers from web images and text. In: Boutilier, C. (Edr.), The Proceedings of the Twenty-First International Joint Conference on Artificial intelligence Pasadena, California, USA, pp. 1169–1174.
NC Institute. Breast Cancer Statistics, USA, 2010. National Cancer Institute, 2010, ⟨http://www.cancer.gov/cancertopics/types/breast⟩ (Accessed: 11 July 2011).
Peterson C., Söderberg B., 1993. Modern heuristic techniques for combinatorial problems. In: Sons, J.W. (Ed.), Artificial Neural Networks New York, USA, pp. 197–242.
Prentice, R.L., Gloeckler, L.A., 1978. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57–67.
Provost, F., Fawcett, T, 2001. Robust classification for imprecise environments. Mach. Learn. 42 (3), 203–231.
Ramoni, M., Sebastiani, P., 2001. Robust learning with missing data. Mach. Learn. 42 (2), 147–170.
Rätsch, G., Onoda, T., Müller, K.-R., 2001. Soft margins for AdaBoost. Mach. Learn. 42 (3), 287–320.
SEER, , 2010. Surveillance Epidemiology and End Results program National Cancer Institute ⟨http://www.seer.cancer.gov⟩ (Accessed: 11 July 2011).
Schölkopf, B., Smola, AJ., 2002. Learning with Kernels. The MIT Press, Cambridge, England.
Sheldon, M.R., Fillyaw, M.J., Thompson, W.D., 1996. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. Physiother. Res. Int. 1 (4), 221–228.
Shin, H., Cho, S., 2007. Neighborhood property-based pattern selection for support vector machines. Neural Comput. 19, 816–855.
Shin, H., Lisewski, A.M., Lichtarge, O., 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 23, 3217–3224.
Shin, H., Hill, N.J., Lisewski, A.M., Park, JS., 2010. Graph sharpening. Expert Syst. Appl. 37, 7870–7879.
Siegel, R., Naishadham, D., Jemal, A., 2013. Cancer Stat., CA: Cancer J. Clin., 63; 11–30
Subramanya A., Bilmes J., 2008. Soft-supervised learning for text classification. In: The Proceedings of the Conference on Empirical Methods in Natural Language Processing Honolulu, Hawaii, pp. 1090–1099.
Sun, Y., Goodison, S., Li, J., Liu, L., Farmerie, W., 2007. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 23, 30–37.
Thongkam J., Xu G., Zhang Y., Huang F., 2008. Breast cancer survivability via AdaBoost algorithms. In: Warren, J.R., Yu, P., Yearwood. J., Patrick, J.D., (Eds.), The Proceedings of the second Australasian workshop on Health Data and Knowledge Management Wollongong, NSW, Australia, pp. 55–64.
Thongkam, J., Xu, G., Zhang, Y., Huang, F., 2009. Towards breast cancer survivability prediction models through improving training space. Expert Syst. Appl. 36, 12200–12209.
⟨http://cogsys.imm.dtu.dk/toolbox/ann/index.html⟩.
⟨http://sourceforge.net/projects/svm/⟩.
⟨http://www.alphaminers.net⟩ (SSL Matlab codes will be available).
Wang, J., 2007. Efficient large margin semi-supervised learning. J. Mach. Learn. Res. 10, 719–742.
Wei L., Keogh E., 2006. Semi-supervised time series classification. In: The Proceedings of the Twelfth International Conference on Knowledge Discovery and Data Mining Philadelphia(KDD 2006), USA, pp. 748–753.
Zhong, S., 2006. Semi-supervised model-based document clustering: a comparative study. Mach. Learn. 65, 3–29.
Zhu X., 2005. Semi-Supervised Learning with Graphs, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, May.
Zhu, X., 2008. Semi-Supervised Learning Literature Survey, Computer Sciences TR 1530 Madison. University of Wisconsin.