Computers in Biology and Medicine 63 (2015) 124–132

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques

Guanjin Wang [a,b,*], Kin-Man Lam [c], Zhaohong Deng [d], Kup-Sze Choi [a,b,e]

[a] School of Nursing, Hong Kong Polytechnic University, Hong Kong, China
[b] Centre for Smart Health, School of Nursing, Hong Kong Polytechnic University, Hong Kong, China
[c] Department of Surgery, Tseung Kwan O Hospital, Hong Kong, China
[d] School of Digital Media, Jiangnan University, Wuxi, Jiangsu, China
[e] Interdisciplinary Division of Biomedical Engineering, Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 2 February 2015
Accepted 17 May 2015

Keywords:
Bladder cancer
Radical cystectomy
Mortality
Prediction
Prognosis
Machine learning

ABSTRACT

Bladder cancer is a common cancer in genitourinary malignancy. For muscle invasive bladder cancer, surgical removal of the bladder, i.e. radical cystectomy, is in general the definitive treatment which, unfortunately, carries significant morbidities and mortalities. Accurate prediction of the mortality of radical cystectomy is therefore needed. Statistical methods have conventionally been used for this purpose, despite the complex interactions of high-dimensional medical data. Machine learning has emerged as a promising technique for handling high-dimensional data, with increasing application in clinical decision support, e.g. cancer prediction and prognosis. Its ability to reveal the hidden nonlinear interactions and interpretable rules between dependent and independent variables is favorable for constructing models of effective generalization performance. In this paper, seven machine learning methods are utilized to predict the 5-year mortality of radical cystectomy, including back-propagation neural network (BPN), radial basis function (RBFN), extreme learning machine (ELM), regularized ELM (RELM), support vector machine (SVM), naive Bayes (NB) classifier and k-nearest neighbour (KNN), on a clinicopathological dataset of 117 patients of the urology unit of a hospital in Hong Kong. The experimental results indicate that RELM achieved the highest average prediction accuracy of 0.8 at a fast learning speed. The research findings demonstrate the potential of applying machine learning techniques to support clinical decision making.

1. Introduction

Bladder cancer is a common cancer in genitourinary tract [1], affecting mainly the elderly population. In Hong Kong, a survey reported in 2011 that bladder cancer constituted 1.4% of all new cases of cancers [2]. Among various types of bladder cancer, muscle invasive bladder cancer has poor prognosis, with high tendency of metastasis and mortality that necessitate prompt treatment [3]. The most effective treatment approach is the surgical excision of bladder and the surrounding lymphatic tissue, which is known as radical cystectomy. Radical cystectomy is a major surgery that has significant morbidity and mortality [2]. The early postoperative complication rate is between 25% and 57% [4,5], and the early mortality rate is around 3%. The rates are higher for elderly patients [6–8]. Efforts have been made to identify the risk factors in order to maximize the operative outcomes, particularly the long-term survival after surgery. In a retrospective review, the association between clinicopathological factors and mortality for 117 patients treated with radical cystectomy for bladder cancer was investigated from statistical inference perspective [2]. In this study, instead of statistical inference, seven machine learning methods – back-propagation neural network (BPN), radial basis function network (RBFN), extreme learning machine (ELM), regularized ELM (RELM), support vector machine (SVM), naive Bayes (NB) classifier and k-nearest neighbour (KNN) algorithm-are exploited as alternative approaches to predict the overall 5-year survival based on the same clinicopathological dataset of 117 patients treated with radical cystectomy.

1.1. Bladder cancer

The indications for radical cystectomy include treatment failures in non-muscle invasive bladder cancer, and T2-T4aN0M0

* Corresponding author at: School of Nursing, Hong Kong Polytechnic University, Hong Kong, China. Tel.: +852 2766 3913; fax: +852 2364 9663.
E-mail address: guanjin.br.wang@connect.polyu.hk (G. Wang).

http://dx.doi.org/10.1016/j.compbiomed.2015.05.015
0010-4825/

muscle invasive bladder cancer [9,10]. Radical systectomy includes the excision of bladder, prostate (in case of men), urethra, part of distal ureter, and lymphatic tissue of pelvis. After the removal of urinary bladder, urinary diversion is performed to divert the urine produced from the kidneys outside of body. Urinary diversion can either be continence diversion by a neo-bladder or continence pouch, or urinary conduit and ureterocutaneostomy. Radical cystectomy and urinary diversion can be performed by conventional open surgery or minimally invasive surgery.

1.2. Prediction and decision support

In the fields of medicine, clinical data are essential for marking diagnosis, formulating treatment and predicting prognosis. Clinicians make use of the knowledge in different specialties to analyze the histological (cell-based), clinical (patient-based), and demographic (population-based) information [11], where statistical methods such as Cox regression, logistic regression and Kaplan–Meier estimator are conventional employed in the analyses. For example, Kaplan–Meier method and Cox proportional hazards model were used to evaluate the prognostic factors of recurrence, progression and disease mortality in patients with bladder cancer [12]. Logistic regression based on 12 variables was used to identify the predictors of overall 5-year survival of patients who had undergone radical cystectomy for bladder cancer [13]. However, with rapid development of health technologies and informatics, medical data of high dimensionality are made available in both volume and variety. The accuracy of outcome prediction depends heavily on effective information integration of data acquired from various sources, clinically or pathologically, making the conventional statistical analyses that rely on clinicians' knowledge and experience a difficult task. The weakness of statistical methods is more apparent when handling medical data with high variability, nonlinear interactions among the variables, and heterogeneous distributions. For example, regression model, a common statistical technique, often requires some explicit assumptions on the relationships among the data that may be practically invalid [13]. Hence, researchers have begun to investigate alternative techniques for clinical outcome prediction, where computational approaches are a main focus. In particular, machine learning has been introduced into the medical domain to overcome the problems with statistical methods and uncover the knowledge hidden in the complex clinical data.

1.3. Machine learning in medicine

Machine learning is a field in computer science leveraging knowledge from artificial intelligence, optimization and statistics to develop algorithms based on the available data. The approach is to build a model by learning from experience (i.e. the existing data, or the known samples acquired) and use the model to make predictions for the new samples [14]. While the quality and size of the samples can affect the prediction performance, machine learning methods are able to handle large, noisy and complex datasets, rendering it a promising technique for broad application in various areas. They have been explored as a more powerful alternative to statistical methods for classifying patterns and making predictions using techniques such as unconventional optimization strategies, conditional probabilities or absolute conditionality [15].

In medicine and healthcare, machine learning has been applied for personalized and predictive medicine [16], cancer diagnosis and detection [15], and for the study of prevention and treatment policy [11]. For bladder cancer, robust outcome predictions of patients undergoing radical cystectomy was achieved using artificial neural work (ANN) prediction model, with configuration optimized by genetic algorithm (GA) [17]. The system was user-friendly and had potential for widespread use for medical decision support. Histology type and bilharziasis datasets were employed to construct a model using ANN and radial basis function network to predict the survival of bladder cancer patients after diagnosis [18]. Besides, clinicopathological and molecular markers were also used to create an ANN model for predicting one-year survival of patients with muscle invasive bladder cancer [19].

As there is never a best algorithm for all problems, it is necessary to test the performance of different algorithms on a specific problem and identify the optimal one [20]. In this paper, seven machine learning methods are investigated to evaluate their performance in predicting bladder cancer mortality after radical cystectomy for the purpose of prognosis. The seven methods – BPN, RBFN, ELM, RELM, SVM, NB classifier and KNN – were selected because they are representatives among the algorithms in their respective domains. Details of these methods will be provided in Section 3. The implementation of these methods included two major processes. In the training process, the learning methods made use of the training dataset, e.g. the supervised input-output pairs, to identify the relationship directly from the clinicopathological data and built the corresponding model. In the testing process, the classification ability of the model is evaluated using the testing dataset. The predicted outputs were compared with the actual outputs of the testing dataset to measure the performance in terms of accuracy, sensitivity, specificity and precision.

The reminder of this paper is organized as follows. Section 2 describes the clinicopathological data adopted in the study. Section 3 briefly reviews the principles of the seven machine learning methods. In Section 4, the 10-fold cross validation strategy and performance indices used in the experiment are introduced. In Sections 6 and 7, the prediction results are presented and discussed. A conclusion is given in Section 7.

2. Clinical data

The dataset employed in this study originated from a retrospective review on the 5-year survival of patients treated with radical cystectomy for bladder cancer [2]. The data were retrieved from computerized clinical records of 117 patients who had undergone radical cystectomy within the period from 2003 to 2011 in a urology unit in Hong Kong. The purpose in the retrospective study was to examine whether age, tumor stage and preoperative serum albumin level are independent predictors of survival after radical cystectomy. Ninety-nine of the patients were male. The mean age was 68 years old (SD = 10 years). There was no loss of follow-up. The mean follow-up time was 31 months (SD = 29 months). The 30-day mortality, 5-year cancer-specific mortality, other-cause mortality, and the overall mortality rates were 3%, 33%, 22% and 55% respectively. Open radical cystectomy was performed in 71 cases and laparoscopic/robotic-assisted radical cystectomy was conducted for the rest. 96 patients had ileal conduit and 21 patients had continent diversion. Other data includes hospital stay duration, preoperative serum albumin level, Charlson comorbidity index, tumor grade, tumor stage and pathological lymph node status. Further details about the dataset can be found in [2].

For those attributes covering wide numerical range when compared with the others, i.e., age, preoperative serum albumin level and follow-up time, pre-processing was performed to normalize them into the range of [0, 1]. With the advice from physicians, irrelevant data were ignored for the study (e.g. date of operation, date of death) and the final dataset used in the study contained 10 attributes (input) and 1 class attribute (output), as shown in Table 1.

Table 1
Clinicopathological dataset used in the study.

Input	Attributes	Input values of machine learning algorithms
1	Gender	1 (female) 2 (male)
2	Age	Normalized to [0,1]
3	Age range	1 (age ≤ 75 years) 2 (age > 75 years)
4	Albumin level	1 (albumin level ≤ 39 g/L) 2 (albumin level > 39 g/L)
5	Surgical approach 1	1 (open surgery) 2 (laparoscopic surgery) 3 (robotic surgery)
6	Surgical approach 2	1 (open surgery) 2 (minimally invasive surgery)
7	Preoperative serum albumin level	Normalized to [0,1]
8	Tumor stage	1 (T1) 2 (T2) 3 (T3) 4 (T4)
9	Follow up period	Normalized to [0,1]
10	Type of diversion	1 (ileal conduit) 2 (neo bladder)
Output	Attribute	Output values
1	5-year mortality	0 (dead) 1 (alive)

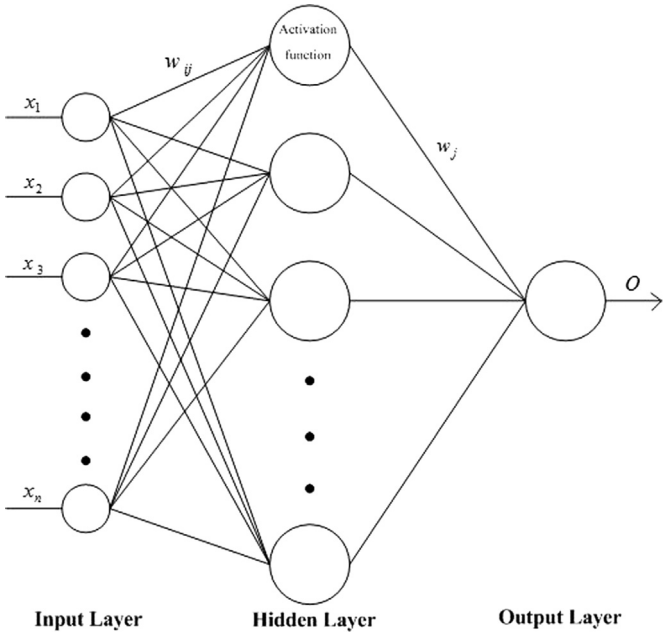

Fig. 1. The BPN structure.

3. Prediction models

In this section, the seven machine learning methods adopted in the study are briefly presented. The setting of the model parameters in the experiment is also discussed.

3.1. Back-propagation network algorithm

ANN emulates the learning ability of biological neural networks where appropriate interconnections (i.e., nodes) are made between the neurons for information transmission and parallel processing at the neurons. BPN is the most prevailing algorithm developed to train the ANN [21]. In BPN, the multilayer perception (MLP) architecture is usually used for data classification and prediction. As shown in Fig. 1, the architecture contains one input layer, one output layer, and one or more hidden layers. The nodes in the ANN are connected by links, each associated with a weight, to model the nonlinear relationship between the input and output layers.

The network is trained using supervised learning with known input–output pairs. In the training process, data fed forward into the input layer pass through the nodes in the hidden layer via the interconnections to produce the predicted results at the nodes in the output layer. The value at each node is computed using a mathematical function, known as activation function [22]. The differences between the prediction and the actual outcome are back-propagated to optimize the network structure (i.e., to learn the nonlinear relationship). In this study, the value H_j of a hidden node j is expressed as

$$H_j = f\left(\sum_{i=1}^{n} w_{ij}x_i - b_j\right) \quad j = 1, 2, \ldots, l, \tag{1}$$

where w_{ij} is the weight of the link connecting the input node i and the hidden node j, x_i is the value of the input node i, b_j is the bias, n is the number of input nodes, l is the number of hidden nodes, and f is the activation function. In this study, f is given by

$$f(x) = \frac{2}{(1+\exp(-2x))} - 1. \tag{2}$$

There was only one output node in the study, which is to predict 5-year survival after surgery. The value O of the node in the output layer is given by

$$O = g\left(\sum_{j=1}^{l} H_j w_j - b\right), \tag{3}$$

where w_j is the weight of the link connecting the hidden node j and the output node, b is the bias, and g is the activation function of the output node, which is given by

$$g(x) = x. \tag{4}$$

If the predicted result is unacceptable, the error e between the computed output O and the target output Y, i.e.,

$$e = Y - O \tag{5}$$

will be back-propagated through the network, so that all the weights and the biases will be re-adjusted to minimize the error. The weight update rules are given by

$$w_{ij} = w_{ij} + \eta H_j(1-H_j)x_i w_{ij} e, \; i = 1, 2, \ldots, n; j = 1, 2, \ldots, l, \text{ and} \tag{6}$$

$$w_j = w_j + \eta H_j e, \; j = 1, 2, \ldots, l \tag{7}$$

where η is the learning rate. The bias update rules are

$$a_j = a_j + \eta H_j(1-H_j)w_j e, \; j = 1, 2, \ldots, l, \text{ and} \tag{8}$$

$$b = b + e \tag{9}$$

The BPN algorithm executes iteratively until an acceptable error is reached or the maximum iteration limit is met.

In this study, the BPN had 10 input nodes and 1 output node, corresponding to attributes of the clinical data listed in Table 1. The number of hidden nodes was determined through repeated testing using the data to obtain the optimal number. While underfitting of the data by the network model can be avoided by using more hidden nodes, the computational time and the model generalization capability are adversely affected. In addition, the

learning rate was also varied in the experiment to identify the optimal value, and thus the corresponding weights and bias values of the resulting BPN. Otherwise, it would result in oscillations and instability [23] if the rate is too high, or leading to a slow training process if the rate is too low. The momentum of the learning process, a parameter to control the ability to get rid of suboptimal solutions during the learning process, is also set appropriately so that the network can converge to the optimal structure in a stable and swift manner.

In the experiment, the number of hidden neurons, the momentum and the learning rate were varied by taking a value from the sets $\{3,5,7,9,11,13,15,17,19,21,23,25,27,29\}$, $\{0.9,0.5,0.2,0\}$ and $\{0.01,0.05,0.09\}$ respectively in each trial. Different combinations were used to evaluate the performance of the BPN obtained, thereby identifying the optimal values and the best BPN structure.

3.2. Radial basis function network

In addition to BPN, another algorithm for training ANN is also investigated in the study, namely, the RBFN [24]. The structure of RBFN adopted in this study was similar to that of the BPN, as shown in Fig. 1. In RBFN, radial basis function is used as the activation function for the hidden nodes. Gaussian function is typically adopted as the radial basis function since it allows for factorization and linear optimization in the formulation. In RBFN, each hidden node j is parameterized with the center vector c_j, and the width σ_j [25]. When the Gaussian function is adopted, the output of the network O is given by

$$O=\sum_{j=1}^{l} w_j \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{c}_j\|^2\right) \tag{10}$$

where $\boldsymbol{x}=(x_1,x_2,...,x_n)^T$ is the input vector, $\boldsymbol{c}_j$ is the center vector for the hidden node j, $\|\boldsymbol{x}-\boldsymbol{c}_j\|$ is the Euclidean distance between the input vector and the center vector, w_j is the weight of the connection between node j and the output node, and l is the number of hidden neurons. The width σ_j of the jth hidden node, also called the spread, is defined as

$$\sigma_j=\frac{c_{\max}}{\sqrt{2l}} \quad j=1,2,...,l \tag{11}$$

where $c_{\max}$ is the maximum of the Euclidean distances between the centers [26].

It can be seen from Eq. (10) that the output of the RBF network is close to 1 if the distance between the input vector and the center vector is small. Geometrically, the hidden layer of the RBFN maps the input vector in the low dimensional space into the high dimensional space. In other words, the problem in the input space becomes linearly separable with a hyperplane in the high dimensional space.

In the experiment, the spread of the RBFN was varied to identify the optimal value. An RBFN with large spread is prone to misclassification while a small spread has poor generalization capability [27]. The spread of the RBF network in the experiment was selected from 25 different values ranging from $2e-12$ to $2e12$, at powers of 10, to compare the performance and find out the one yielding the highest accuracy.

3.3. Extreme learning machine

ELM is a fast learning method for single-hidden layer neural network [28,29]. A key feature of ELM is that the weights and the bias between the input and the hidden layers are randomly assigned, whereas the weights between the hidden and the output layers are analytically determined by utilizing Moore–Penrose (MP) generalized inverse operation of the hidden output matrices [30]. In this study, the algorithm used to implement the ELM is summarized as follows.

Step 1: Randomly assign the input weight w_{ij} between the input and the hidden layers, and the bias b_j. The nonlinear system is then transformed into a linear system and can be expressed as

$$H\boldsymbol{\beta}=\boldsymbol{T} \tag{12}$$

where $\boldsymbol{H}$ is the hidden-layer output matrix, $\boldsymbol{\beta}$ is the matrix of output weights and $\boldsymbol{T}$ is the matrix of the desired output. Further details about the formulation can be found in [30].

Step 2: Calculate the hidden layer output matrix $\boldsymbol{H}$.

Step 3: Calculate the output weights vector $\boldsymbol{\beta}$ by obtaining the least-square solution of the linear system in Eq. (13).

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\boldsymbol{T} \tag{13}$$

where $\boldsymbol{T}=[y_1,...,y_N]^T$ is the output vector, N is the number of input samples and $\boldsymbol{H}^{+}$ is the MP generalized inverse of the matrix $\boldsymbol{H}$.

3.4. Regularized ELM

A variation of ELM, RELM, was also investigated in this study. RELM inherits the fast learning feature of ELM while its generalization performance is enhanced by using the least squares regression methods to identify the degree of relevance of the weight linking a hidden node to the output layer, where penalties are applied to the coefficient vectors [31]. In RELM, the regularization parameter γ is introduced to improve the controllability [32]. In order to reduce the effect of noise, RELM introduces a weighting factor v_i to weigh the error between the output of RELM and the actual output of the ith input sample. Hence, the output weight $\boldsymbol{\beta}$ in Eq. (13) can be expressed as

$$\boldsymbol{\beta}=\left(\frac{I}{\gamma}+\boldsymbol{H}^T\boldsymbol{D}^2\boldsymbol{H}\right)^{+}\boldsymbol{H}^T\boldsymbol{D}^2\boldsymbol{T}, \tag{14}$$

where $\boldsymbol{D}=dialog\,(v_1,v_2,...v_N)$ and γ is the regularized factor. In the experiment, the number of hidden nodes in both the ELM and RELM was set experimentally by selecting a value from $\{20,22,24,26,28,30,32,34,36,38\}$.

3.5. Support vector machine

SVM is a typical kernel-based technique for supervised data classification. The basic principle is to create a hyperplane as the decision surface for classification, where the edge of the isolation between different categories of data is maximized. In the process, the input data vectors are first mapped to a high-dimensional space [33]. The SVM algorithm then searches for a hyperplane with the largest margin in order to achieve the best generalization ability. Fig. 2 gives an example with two linearly separable classes and the data points are denoted by crosses and triangles respectively. The points closest to the decision surface are the supporting vectors and the distance between support vectors and surface is the margin.

Consider a given training dataset $\boldsymbol{T}=\{(\boldsymbol{x_1},y_1),...,(\boldsymbol{x}_N,\ y_N)\}\in(X\times Y)$ with $\boldsymbol{x}_i\in X=\boldsymbol{R}^n, y_i\in Y=\{1,-1\}(i=1,2,...,N)$, where the training data matrix X has two separable classes with the class labels -1 and $+1$ stored in the vector Y. Applying the Lagrangian multiplier with the kernel function $K(\boldsymbol{x},\boldsymbol{x}')$ and the regularization parameter$C$, the dual formulation of SVM is written as follows,

$$\min \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{j=1}^{N}\alpha_j$$

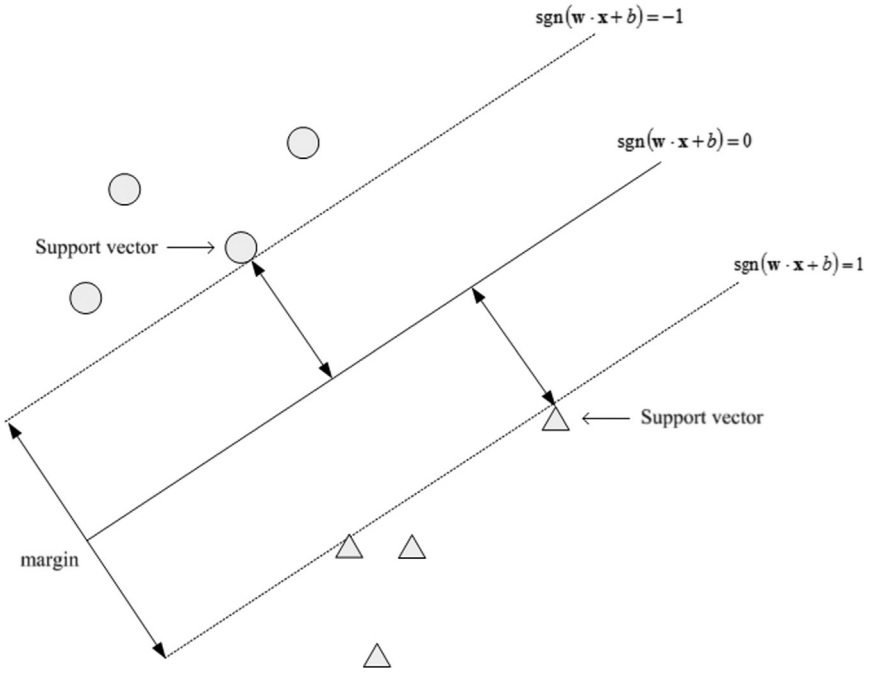

Fig. 2. SVM learns the hyperplane that best separates the two classes with data points denoted by circles and triangles, with the label $f(\boldsymbol{x}) = -1$ and $f(\boldsymbol{x}) = +1$ respectively.

$$\text{s.t.} \sum_{i=1}^{N} y_i \alpha_i = 0,\ 0 \le \alpha_i \le C, i = 1, ..., N \tag{15}$$

where the optimal solution $\boldsymbol{\alpha}^* = (\alpha_1^*, ..., \alpha_N^*)^T$ and threshold value b^* can be obtained. Thus the decision function for the classification is given by

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b^*\right). \tag{16}$$

The summation in Eq. (16) is performed only on a small group of support vectors with non-zero α_i.

In this study, a polynomial kernel function was used in the experiment, where the optimal value of the regularization parameter C is determined using the 10-fold cross-validation strategy (to be discussed in Section 4.1) to achieve a balance between classification accuracy (with a larger C) and generalizability (with a smaller C, i.e. larger margin). The value of C in the experiment was chosen from $\{150, 200, 250\}$.

3.6. K-nearest neighbour algorithm

The KNN algorithm was also investigated in this study for its simplicity in implementation [34]. In KNN, the training dataset is reserved and used to classify a new unclassified testing dataset. Classification is achieved by comparing the testing dataset with the groups in the training set to identify the most similar one based on a distance function [35]. In this study, the similarity between two neighboring data points **x** and **y** is measured by the Euclidean distance

$$\sqrt{\sum_i (x_i - y_i)^2}, \tag{17}$$

where $\mathrm{x} = (x_1, x_2, ..., x_n)$ and $\mathrm{y} = (y_1, y_2, ..., y_n)$ are vectors in n-dimensional space.

The number of neighbors k has to be set appropriately to reduce the effect of noise and outliers on the classification (k too small) and to avoid the dominance of local behavior (k too large) [35]. In the study, the value of k was chosen experimentally with values in $\{10, 12, 15, 18, 20\}$ to identify the one with the best performance.

3.7. Naive Bayes classifier

Lastly, the NB classifier was employed in the study by assuming that all the attributes were conditionally independent of the class labels [36,37]. In the training stage, the NB classifier estimated the parameters of the class priors and the probability distribution of the attributes by analyzing the training samples. In the testing stage, the method calculated the posterior probability of every sample belonging to each class. The NB classifier then selected the class with the largest posterior probability as the output of the testing samples.

4. Evaluation

4.1. Performance indices

In supervised learning, for a given sample with known input and output, an ideal prediction model taking the same input is expected to produce a result consistent with the known output of the sample. In other words, the 5-year mortality suggested by the prediction model is “alive” for a sample indicating that the bladder cancer patient was alive after radial cystectomy (i.e., true positive). The output of the prediction model is “dead” for a sample recording that the bladder cancer patient had been dead after the surgery (i.e., true negative). In the experiment, the reliability of the prediction models was quantified by the four standard performance measures, namely, accuracy, sensitivity, specificity and precision, which are given by

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN), \tag{20}$$

$$\text{Sensitivity} = TP/(TP + FN), \tag{21}$$

$$\text{Specificity} = TN/(TN + FP) \text{ and} \tag{22}$$

$$\text{Precison} = TP/(TP + FP), \tag{23}$$

where *TP*, *TN*, *FP*, *FN*denotes the number of true positives, true negatives, false positives and false negatives respectively.

4.2. Ten-fold cross validation

In this study, the training and testing datasets required for evaluating the machine learning methods were generated randomly using the clinicopathological dataset of the bladder cancer patients. To reduce the biasing effect that may be introduced in the generation of the training and testing datasets, which can lead to over-fit or under-fit models, the 10-fold cross-validation strategy was used in the experiment. Here, the original clinicopathological dataset was randomly divided into 10 subsets. The subsets had a class distribution similar to that of the original dataset. Cross validation was performed by alternately using one subset as the testing dataset to evaluate the model built using the remaining 9 subsets as the training dataset. This process was repeated 10 times so that every subset took turn to serve as the testing dataset, thereby resulting in 10 models. The mean and standard deviation of the performance indices of the 10 models created by the cross-validation strategy were calculated [38].

5. Experiments

The models for predicting the 5-year mortality after radical cystectomy were experimentally studied by using the clinicopathological data described in Section 2 and the seven machine learning methods discussed in Section 3. Three experiments were conducted.

5.1. Prediction performance

First, experiments were conducted to identify the optimal parameters for these models, except the NB classifier where the model was built based on prior knowledge deduced from the clinicopathological data. The parameters are shown in Table 2. The seven prediction models were evaluated and the experimental results are given in Table 3.

5.2. Predictor identification

Second, as discussed in Section 2, the dataset used here was originally collected to examine the association between patient age, tumor stage, preoperative serum albumin level and mortality following radical cystectomy [2]. In this study, experiments were designed to serve the same purpose using machine learning techniques. This was achieved by using the random permutation strategy [39] in which an attribute can be evaluated by randomly permuting the values of that attribute with the data in the same dataset, and then observing any reduction in accuracy of the machine learning methods before and after random permutation. The larger the decrease is in accuracy, the stronger the association of that parameter is with the outcome. In experiment, we first randomly permuted the values of the attribute "age" and then run the machine learning models on the randomly permuted dataset. The experiment was repeated ten times to calculate the average accuracy. The experiments were conducted in the same way for the attributes "tumor stage" and "preoperative serum albumin level" respectively. As it is evident from Table 3 that the performance of ELM and RELM is outstanding among the seven machine learning methods, emphasis was put on these two methods and the experimental results presented in Tables 4, 5 and 6.

5.3. Comparison with the common-used nomogram

Third, after the seven machine learning models were built by using the clinicopathological data described in Section 2, experiments were conducted to do comparison with the common-used nomogram prediction method designed by Lughezzani [40]. The method, which is also available on-line [41], is developed using the smoothed Poisson regression model according to disease stage and patient age [40,41]. A dataset with attributes required in both the nomogram prediction method and the machine learning models here was acquired in order to carry out the comparisons. Records of 30 patients were contained in the dataset. For fair comparison, we set the prediction output of this common nomogram to 1 if its probabilistic output is bigger than 0.5, 0 otherwise. Similarly, emphasis was put on the two outperforming machine learning methods ELM and RELM. The experimental results are shown in Table 7.

6. Discussion

The paper investigated the use of seven machine learning methods to predict the mortality of after the mortality after radical cystectomy. The performance of the seven prediction models could be broadly classified into three tiers, refer to experimental results in Table 3. The prediction models developed using BPN, NB classifier and KNN exhibited relatively low performance, with a mean classification accuracy of 0.7222, 0.7333 and 0.7222 respectively. The middle tier includes the prediction models developed using SVM and RBFN. Their mean classification accuracy was 0.7556 and 0.7667 respectively. For ELM and RELM, the resulting prediction models developed showed the best classification performance, with RELM achieving the highest mean accuracy of 0.8000. While their mean classification accuracy obtained using ELM and RBFN was the same, the maximum value achieved by the former was higher. In this practical medical application, ELM and RELM, as theoretically well-founded machine learning techniques, were found to outperform the other five methods.

Regarding the low performers, while the principles of the NB classifier is relatively more intuitive to clinicians, the assumption of conditional independence of the attributes is practically difficult to be fully satisfied in many medical applications. The probabilities of the events are usually obtained individually and their associations and interactions are unclear, which affect the validity of the assumption of the NB classifier and thus the classification accuracy. If the assumption is hold, the NB classifier can converge quickly and require less training data. It can be seen from the results in Table 3 that the NB classifier slightly outperformed the KNN method, in the terms of mean and maximum accuracy. As the size of the dataset in this study was relative small, the NB classifier,

Table 2
Parameter settings of the prediction models.

Models	BPN	ELM	RELM	RBFN	SVM	KNN
Parameter settings	Hidden neurons=15 Learning rate=0.05 Momentum=0.9	Hidden neurons=24	Hidden neurons=28	Spread=2*e*4	C=200	k=18

Table 3
Performance of seven models on the adopted dataset.

Performance		Machine learning models						
		BPN	ELM	RELM	RBFN	SVM	NB	KNN
Accuracy	Mean	0.7222	0.7667	**0.8000**	0.7667	0.7556	0.7333	0.7222
	SD	0.0586	0.0820	0.0625	0.0268	0.0486	0.0799	0.0642
	Max.	0.8056	0.8611	0.9167	0.8056	0.8333	0.8611	0.8333
	Min.	0.5833	0.6111	0.7222	0.7222	0.6944	0.6111	0.6111
Sensitivity	Mean	0.7764	0.7349	**0.8559**	0.7900	0.7542	0.7375	0.7507
	SD	0.0944	0.1064	0.0876	0.0928	0.1016	0.1215	0.0798
Specificity	Mean	0.6811	**0.8151**	0.7236	0.7533	0.7704	0.7342	0.7014
	SD	0.0862	0.1101	0.0967	0.0667	0.1257	0.1144	0.1226
Precision	Mean	0.7083	**0.8493**	0.8160	0.7441	0.7870	0.7264	0.7630
	SD	0.0982	0.0917	0.0536	0.0668	0.1177	0.1141	0.0680
Running time (ms)		4651	45	20	479	223	30	36

Table 4
Performance of ELM and RELM on dataset with "age" randomly permuted.

Performance		ELM	Accuracy decrease (%)	RELM	Accuracy decrease (%)
Accuracy	Mean	0.6750	11.96	0.7167	10.42
	SD	0.0572	–	0.0464	–
	Max.	0.7500	–	0.7778	–
	Min.	0.5833	–	0.6389	–

Table 5
Performance of ELM and RELM on dataset with "tumor stage" randomly permuted.

Performance		ELM	Accuracy decrease (%)	RELM	Accuracy decrease
Accuracy	Mean	0.6389	16.67	0.6917	13.54
	SD	0.0540	–	0.0592	–
	Max.	0.7222	–	0.7778	–
	Min.	0.5556	–	0.5833	–

Table 6
Performance of ELM and RELM on dataset with "preoperative serum albumin level" randomly permuted.

Performance		ELM	Accuracy decrease (%)	RELM	Accuracy decrease (%)
Accuracy	Mean	0.6556	14.49	0.7517	6.040
	SD	0.0480	–	0.0613	–
	Max.	0.7500	–	0.8889	–
	Min.	0.5556	–	0.6111	–

Table 7
Performance of ELM, RELM and the adopted nomogram prediction method on the test data.

Performance	ELM	RELM	The nomogram prediction method
Accuracy	0.7333	0.7750	0.6330

which is a high-bias-low-variance approach, was advantageous over low-bias-high-variance KNN method. This is mainly because the NB classifier is a linear method that constructs only a hyperplane, whereas KNN is a nonlinear method that builds variable boundaries between classes. The latter can adversely increase the probability of misclassification, particularly for noisy data. Moreover, the value of k in KNN is case dependent and the method is prone to over-fitting [42].

Among the seven prediction models, the models developed using RELM and ELM exhibited the highest mean sensitivity and specificity respectively (over 0.8), while the ones built with ELM and BPN had the lowest mean sensitivity and specificity respectively. For all the prediction model, the mean values of sensitivity and specificity values were all above 0.7, except the mean specificity achieved using BPN (i.e. 0.6811). For the models built using ELM and SVM, the mean sensitivity was lower than the mean specificity, the reverse was true for the other machine learning methods, except the NB classifier, where the mean sensitivity and specificity were about the same. The difference between the mean sensitivity and mean specificity was relatively large for the model developed using regularized ELM (0.8599 vs. 0.7236). Similar situations were also observed from the models built using the ELM and BPN. Ideally, prediction models with high sensitivity and high specificity are desired. The small sample size in this study may be an obstacle towards an ideal model. Taking precision into consideration, the models developed using ELM and RELM exhibited both high mean accuracy and high mean precision, which is desirable for outcome prediction.

On the other hand, the neural networks constructed using ELM were found to contain more hidden nodes when compared to that using BPN [30], which complicated the network architecture and incurred more computations. Nevertheless, the timing performance of the machine learning methods ELM and RELM was indeed outstanding whereas the learning speed of the models based on BPN and RBFN was relatively slow. The slow learning was attributed to the fact that the conventional gradient-descent based methods were used in BPN and RBFN [43], where multiple iterations were required to adjust and optimize the weights and bias of the neural network to avoid sub-optimal solutions resulting from the problem of local minima. Besides, manual setting of parameters was also required in the gradient-descent methods.

Moreover, by comparing Tables 4, 5 and 6 with Table 3, we can easily find that the accuracy performance obtained by both ELM and RELM degrades a lot after random permutation of 'age', 'tumor stage' and 'preoperative serum albumin level'. The accuracy was decreased by 11.96%, 16.67% and 14.49% respectively for ELM; and 10.42%, 13.54% and 6.04% respectively for RELM. Therefore, we can conclude by using the machine learning approach that these three attributes are predictors of survival after radical cystectomy. The findings here are in line with the claim in [2], demonstrating the feasibility of using ELM and RELM to identify the predictors of survival after radical cystectomy.

Finally, referring to Table 7, it is found that the proposed ELM and RELM models have better accuracy than the adopted nomogram prediction method, i.e. greater than 0.7 for ELM and RELM versus 0.63 for the nomogram prediction method. In other words, the results demonstrate the advantage of using machine learning technology in predicting mortality after radical cystectomy.

In summary, the findings in the experiments suggested that ELM and RELM were well-suited for the prediction of mortality after radical cystectomy concerned in this study.

7. Conclusion

Seven machine learning methods were explored in this study for the prediction of mortality of radical cystectomy for bladder cancer. Real clinicopathological data of 117 patients who had undergone the surgery were adopted. Prediction models were developed using the machine learning methods and the performance was evaluated in terms of accuracy, sensitivity, specificity and precision.

The experimental results indicated that ELM and RELM demonstrated competitive performance among the machine learning methods investigated. The running time was also among the lowest. The findings suggest that the ELM-based algorithms are relatively more effective in the prediction of mortality of radical cystectomy. In addition to superiority in accuracy and speed, ELM is also advantageous over the gradient-descent algorithms in machine learning that require manual parameter tuning and suffer from the problem of local minima. Moreover, the feasibility of using ELM and RELM to identify the predictors of survival after radical cystectomy was validated. The proposed ELM-based algorithms exhibited better accuracy than one common-used nomogram prediction method.

In addition to these seven models, it is also worthwhile to evaluate the performance of other regression models like least square support vector regression on the bladder cancer data set, which is a future work of the study. Furthermore, it is also noted that the classification accuracy of the seven models investigated in the study was indeed not high enough (less than 0.8) to offer a reliable reference for clinical decision support. This may be caused by the small sample size in the study, which calls for the need to establish a centralized data repository to collect the clinicopathological data obtained from the urology units of multiple hospitals for bladder cancer research and to facilitate the sharing of the data for the development of reliable prediction models. Further work on external validation will be conducted through prospective collection of the test data with the attributes required by the methods of the Bladder Cancer Consortium or the SEER-Medicare database, and also the machine learning models presented here. To facilitate data collection, a clinical data repository is being set up to consolidate the data from the urology unit of multiple hospitals.

Conflicts of interest statement

None declared.

Acknowledgments

This work was supported in part by the Research Grants Council of the Hong Kong SAR (PolyU5134/12E), the Hong Kong Polytechnic University (G-UC93), and the scholarship donated by Nelson Y.C. Yu.

References

[1] E.S.Y. Chan, et al., Current management practice for bladder cancer in Hong Kong: a hospital-based cross-sectional survey, Hong Kong Med. J. 20 (3) (2014) 229–233.
[2] E.S. Chan, et al., Age, tumor stage, and preoperative serum albumin level are independent predictors of mortality after radical cystectomy for treatment of bladder cancer in Hong Kong Chinese, Hong Kong Med. J.= Xianggang yi xue za zhi/Hong Kong Acad. Med. 19 (5) (2013) 400.
[3] J. Reynard, S. Brewster, S. Biers, Oxford Handbook of Urology, Oxford University Press, Oxford, England, 2013.
[4] G. DALBAGNI, et al., Cystectomy for bladder cancer: a contemporary series, J. Urol. 165 (4) (2001) 1111–1116.
[5] D. Rosario, M. Becker, J. Anderson, The changing pattern of mortality and morbidity from radical cystectomy, BJU Int. 85 (4) (2000) 427–430.
[6] J.P. Stein, et al., Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1054 patients, J. Clin. Oncol. 19 (3) (2001) 666–675.
[7] W.S. McDougal, et al., Campbell-Walsh Urology (Review), 10th ed., Elsevier Health Sciences, Philadelphia, W.B. Saunders, 2011.
[8] E. Rubin, H.M. Reisner, Essentials of Rubin's Pathology, Lippincott Williams & Wilkins, Baltimore, Maryland, 2009.
[9] J.A. Witjes, et al., EAU guidelines on muscle-invasive and metastatic bladder cancer: summary of the 2013 guidelines, Eur. Urol. 65 (4) (2014) 778–792.
[10] J.R. Egner, AJCC cancer staging manual, JAMA 304 (15) (2010) 1726–1727.
[11] J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, Cancer Inform. 2 (2006) 59.
[12] F. Millan-Rodriguez, et al., Multivariate analysis of the prognostic factors of primary superficial bladder cancer, J. Urol. 163 (1) (2000) 73–78.
[13] P. Bassi, et al., Prognostic accuracy of an artificial neural network in patients undergoing radical cystectomy for bladder cancer: a comparison with logistic regression analysis, BJU Int. 99 (5) (2007) 1007–1012.
[14] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, MIT Press, Cambridge, Massachusetts, 2012.
[15] J.F. MCCARTHY, et al., applications of machine learning and high – dimensional visualization in cancer detection, diagnosis, and management, Ann. N. Y. Acad. Sci. 1020 (1) (2004) 239–262.
[16] A.D. Weston, L. Hood, Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine, J. Proteome Res. 3 (2) (2004) 179–196.
[17] A.M. Vukicevic, et al., Evolutionary assembled neural networks for making medical decisions with minimal regret: application for predicting advanced bladder cancer outcome, Expert Syst. Appl. 41 (18) (2014) 8092–8100.
[18] W. Ji, R.N.G. Naguib, M.A. Ghoneim, Neural network-based assessment of prognostic markers and outcome prediction in bilharziasis-associated bladder cancer, IEEE Trans. Inf. Technol. Biomed. 7 (3) (2003) 218–224.
[19] K.N. Qureshi, et al., Neural network analysis of clinicopathological and molecular markers in bladder cancer, J. Urol. 163 (2) (2000) 630–633.
[20] K.-R. Müller, et al., Classifying 'drug-likeness' with kernel-based learning methods, J. Chem. Inf. Model. 45 (2) (2005) 249–253.
[21] R.J. Erb, Introduction to backpropagation neural network computation, Pharm. Res. 10 (2) (1993) 165–170.
[22] P.D. Heermann, N. Khazenie, Classification of multispectral remote sensing data using a back-propagation neural network, IEEE Trans. Geosci. Remote Sens. 30 (1) (1992) 81–88.
[23] A.Abraham, Artificial neural networks, Handbook of Measuring System Design, 2005.
[24] J. Park, I.W. Sandberg, Approximation and radial-basis-function networks, Neural Comput. 5 (2) (1993) 305–316.
[25] S. Chen, B. Mulgrew, P.M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Netw. 4 (4) (1993) 570–590.
[26] S. Panda, D. Chakraborty, S. Pal, Flank wear prediction in drilling using back propagation neural network and radial basis function network, Appl. Soft Comput. 8 (2) (2008) 858–871.
[27] M.J. Er, et al., Face recognition with radial basis function (RBF) neural networks, IEEE Trans. Neural Netw. 13 (3) (2002) 697–710.
[28] G.-B.Huang, Q.-Y. Zhu, and C.-K. Siew , Extreme learning machine: a new learning scheme of feedforward neural networks, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2004.
[29] M.-B. Li, et al., Fully complex extreme learning machine, Neurocomputing 68 (2005) 306–314.
[30] Q.-Y. Zhu, et al., Evolutionary extreme learning machine, Pattern Recogn. 38 (10) (2005) 1759–1763.
[31] J.M. Martínez-Martínez, et al., Regularized extreme learning machine for regression problems, Neurocomputing 74 (17) (2011) 3716–3721.
[32] W. Deng,Q. Zheng, L. Chen, Regularized extreme learning machine, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM, 2009.
[33] E. Byvatov, et al., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inform. Comput. Sci. 43 (6) (2003) 1882–1889.
[34] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (3) (1992) 175–185.
[35] D.T.Larose, k – nearest neighbor algorithm. Discovering Knowledge in Data: An Introduction to Data Mining, 2005, pp. 90–106.
[36] K.P. Murphy, Naive Bayes classifiers, University of British Columbia, Vancouver, B.C, 2006.
[37] I. Rish , An empirical study of the naive Bayes classifier, in: Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.
[38] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the IJCAI, 1995.
[39] J.-B. Yang, et al., Feature selection for mlp neural network: the use of random permutation of probabilistic outputs, IEEE Trans. Neural Netw. 20 (12) (2009) 1911–1922.

[40] G. Lughezzani, et al., A population-based competing – risks analysis of the survival of patients treated with radical cystectomy for bladder cancer, Cancer 117 (1) (2011) 103–109.
[41] Nomogram predicting the probability of mortality due to bladder cancer versus other causes, 2006. Available from: ⟨http://labs.fccc.edu/nomograms/nomogram.php?id=48&audience=1⟩.
[42] M.J. Isla, et al., Investigating the performance of naive-Bayes classifiers and k-nearest neighbor classifiers, in: Proceedings of the IEEE International Conference on Convergence Information Technology, 2007.
[43] G.-B. Huang, D.H. Wang, Y. Lan, Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern. 2 (2) (2011) 107–122.