Expert Systems with Applications 41 (2014) 8092–8100

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Evolutionary assembled neural networks for making medical decisions with minimal regret: Application for predicting advanced bladder cancer outcome

Arso M. Vukicevic [a], Gordana R. Jovicic [a,*], Miroslav M. Stojadinovic [b], Rade I. Prelevic [c], Nenad D. Filipovic [a]

[a] Faculty of Engineering, University of Kragujevac, Serbia
[b] Clinic of Urology and Nephrology, Kragujevac, Serbia
[c] Military Medical Academy, Belgrade, Serbia

ARTICLE INFO

Article history:
Available online 11 July 2014

Keywords:
Bladder cancer
Expert systems
Artificial Neural Network
Genetic Algorithms
Regret Theory

ABSTRACT

Development of reliable medical decision support systems has been the subject of many studies among which Artificial Neural Networks (ANNs) gained increasing popularity and gave promising results. However, wider application of ANNs in clinical practice remains limited due to the lack of a standard and intuitive procedure for their configuration and evaluation which is traditionally a slow process depending on human experts. The principal contribution of this study is a novel procedure for obtaining ANN predictive models with high performances. In order to reach those considerations with minimal user effort, optimal configuration of ANN was performed automatically by Genetic Algorithms (GA). The only two user dependent tasks were selecting data (input and output variables) and evaluation of ANN threshold probability with respect to the Regret Theory (RT). The goal of the GA optimization was reaching the best prognostic performances relevant for clinicians: correctness, discrimination and calibration. After optimally configuring ANNs with respect to these criteria, the clinical usefulness was evaluated by the RT Decision Curve Analysis. The method is initially proposed for the prediction of advanced bladder cancer (BC) in patients undergoing radical cystectomy, due to the fact that it is clinically relevant problem with profound influence on health care. Testing on the data of the ten years cohort study, which included 183 evaluable patients, showed that soft max activation functions and good calibration were the most important for obtaining reliable BC predictive models for the given dataset. Extensive analysis and comparison with the solutions commonly used in literature showed that better prognostic performances were achieved while user-dependency was significantly reduced. It is concluded that presented procedure represents a suitable, robust and user-friendly framework with potential to have wide applications and influence in further development of health care decision support systems.

1. Introduction

According to the clinical reports, bladder cancer (BC) represents the most common urologic cancer in men and the fifth most common malignancy worldwide (Jemal et al., 2013). Since recently reported bladder tumours are superficial, the possibility of their progression and their muscle invasive (MI) nature make BC treatment a very challenging task with a profound influence on health care. The most important factors for predicting pathological stages are clinical staging based on physical examination, transurethral resection (TUR) pathology and imaging (Compérat & Van der Kwast, 2013; Ramy & Yair, 2011). However, predictions in BC remain nontrivial both in staging of primary tumour as well as in nodal staging (Bostrom et al., 2010). Since cancer classification commonly relies on clinical and histopathological information, clinicans' decision may turn out to be incomplete or misleading. As a result, clinical prediction has evolved from physician judgment alone to the use of various medical decision support systems and predictive models (Lughezzani et al., 2010). In literature, two most frequently used predictive models for prediction of BC outcome are Artificial Neural Networks (ANNs) and Logistic Regression (LR). After extensive work performed on comparing and analysing their performances, ANNs have gained increasing popularity over LR in research community (Bassi, Sacco, De Marco, Aragona, & Volpe, 2007; Dreiseitla & Ohno-Machadob, 2002; Hu et al., 2013). ANNs are data mining technique developed on the

* Corresponding author. Address: Sestre Janjic 6, Kragujevac 34000, Serbia. Tel.: +381 34334379; fax: +381 34333192.
E-mail addresses: arso_kg@yahoo.com (A.M. Vukicevic), gjovicic.kg.ac.rs@gmail.com (G.R. Jovicic), midinac@EUnet.rs (M.M. Stojadinovic), drrprelevic@hotmail.com (R.I. Prelevic), fica@kg.ac.rs (N.D. Filipovic).

http://dx.doi.org/10.1016/j.eswa.2014.07.006
0957-4174/

basis of biological process of learning (Looney, 1993). In cancer research, they have been used due to their ability to learn and recognize complex data patterns and identify nonlinear interactions between input (dependent) and output variables (Lisboa & Taktak, 2006). In oncological urology ANNs have been applied with promising results for the diagnostic (Çınar, Engin, Engin, & Ateşçi, 2009), staging (Anagnostou, Remzi, & Djavan, 2003) and prognostic (Saritas, Ozkan, & Sert, 2010) problems of prostate cancer. In BC patients ANNs have been used to predict outcome following screening (Finnea et al., 2000), prostate biopsy (Meijer et al., 2009) and radical cystectomy (El-Mekresh et al., 2009), as well as disease progression and tumour recurrence of non-invasive transitional cell carcinoma (Remzi, Anagnostou, & Ravery, 2003).

However, the application of ANNs in clinical practice remains limited due to a few reasons. It is assumed that users are able to choose an optimal combination of various ANN configuration parameters (such as: available activation functions, number of neurons and layers, learning algorithms, momentum, how much data to use for training, validation, testing, which objective-function to use for measuring quality of training and other settings). It may be noticed that for the efficient usage of ANNs clinicians should be familiar with a complex evolving foundation of ANN framework. As a result of such limitations, a wider usage of ANNs among clinicians remains unpopular despite promising results obtained by research community. In order to reduce user-dependency of ANNs based expert systems, evolutionary ANNs (EANNs) were proposed (Castellani, 2013; Yao & Liu, 1998). The basic idea of EANNs is to use an optimization procedure for increasing ANN performances iteratively with respect to some criteria (Castellani, 2013; Rivero, Dorado, Rabuñal, & Pazos, 2010; Tallón-Ballesteros & Hervás-Martínez, 2011). Independently from the progress in the field of neurocomputing, authors focused on the evaluation of medical decision support systems found that measuring clinical usefulness of decisions in medicine is not equal to measuring accuracy (according to which traditional ANN-based expert systems were configured) (Baker, 2009; Baker & Kramer, 2012; Holmberg & Vickers, 2013). It is explained that traditional way of measuring predictive accuracy does not capture important characteristics of intelligent behavior necessary for making rational decisions in medicine (for example taking into account clinical implications such as harms and benefits of making wrong and correct decisions – see Section 2.3 later) (Mallett, Haligan, Thompson, Collins, & Altman, 2012). To our best knowledge, obtaining reliable predictive models with respect to clinical needs remains an open question the answer to which could advance knowledge discovery in medicine (Esfandiari, Babavalian, Moghadam, & Tabar, 2014). Taking this into account, the aim of this study was to develop a robust procedure for obtaining ANN predictive models with high prognostic performances for the prediction of advanced BC in patients undergoing RC.

2. Materials and methods

For simplicity, this section is divided into three parts. First, the traditional expert-dependent configuration of ANNs is briefly introduced and the problems which this study aims to solve are higlighted. Next, in order to reach performances of expert ANN users with minimal user-system interaction, an evolutionary approach for configuring of ANN was presented. Finally, a regret approach was applied for evaluating clinical usefulness of ANN predictive models in order to estimate best ANN model for a particular problem.

2.1. Common ANN configuration and training procedure for the purposes of classification and prediction of advanced BC

ANN (Wallis, 1999) may be described as a mathematical model which on a much smaller scale mimics the way a biological neural network works (Fig. 1(a)). Transmission of electrical signals over neuron connections (axon and dendrites) is mathematically modeled as a sum of n weighted scalar inputs p_i and constant b (called bias): $s = \sum_i^n p_i w_i + b$. The result is then used as an argument of an activation function f, which produces the output $t = f(s)$.

The central idea of the ANN framework is that by adjusting the scalar parameters b_i and w_i an artificial neuron can exhibit a desired intelligent behavior (such as classification, prediction or estimation). The most frequently used types of transfer functions are the hard-limit (or step), linear, sigmoid (or logistic) and soft max, to name just a few (Karlik & Olgac, 2010). A common ANN architecture consists of many neurons organized in layers. The ANN architecture considered in this paper is a feed-forward multilayer perception (FFN) with a single hidden layer as the most suitable for the purposes of binary classification and survival prediction (Zhang, 2000) (Fig. 1(b)).

Regarding the ANN training, there are two main principles of learning: supervised and unsupervised. The subject of this paper was supervised learning with backpropagation learning algorithm (Yu & Chen, 1997), when the network is provided with a set of examples (pair of inputs and known correct outputs). For the purpose of learning, the input data are commonly divided into three sets: training, validation and testing. A training data set was used only for learning (adjusting weights and biases). Validation set was used to decide when to stop the training process – to avoid overfitting (a situation when the ANN memorizes the training data rather than learning the rules that govern them). A testing set was used for independently measuring performance of the trained network. In the beginning of the training, the network was initialized with randomly chosen weights. After the inputs were applied to the ANN, for every qth iteration (learning epoch), the prediction was compared with the true category by calculation of classification error E^q. Since E^q is a continuous and differentiable function of l weights between neurons, it could be minimized by using an iterative gradient descent procedure. After calculating $\nabla E^q = \left(\frac{\partial E^q}{\partial w_1^q}, \frac{\partial E^q}{\partial w_2^q}, \ldots, \frac{\partial E^q}{\partial w_l^q}\right)$, each weight in the next $(q+1)$ epoch was updated using the increment Δw_i^q: $w_i^{q+1} = w_i^q - \Delta w_i^q = w_i^q - \gamma \frac{\partial E^q}{\partial w_i^q}$ for $i = 1 \ldots l$, where γ represents the learning constant which defines the step length (in weights space). Therefore, the whole process of ANN learning is in the end reduced to iteratively adjusting the neurons' interconnections ("weights") until the classification error converges ($\nabla E = 0$) with respect to some criteria and configuration parameters. For the purposes of minimizing error-function, different algorithms may be used: Levenberg–Marquardt, Gradient Descent, Polak–Ribiére Conjugate Gradient and BFGS Quasi-Newton, to name just a few (Ghaffari et al., 2006; Lethaus, Baumann, Köster, & Lemmer, 2013; Mukherjee & Routroy, 2012).

To sum up, training of the ANN to do a particular task may be intuitively described as choosing various above mentioned parameters. Since a significant number of parameters is required to be set correctly, a deeper understanding of ANN framework remains necessary which represents an obstacle for the wider application of ANN predictive models among clinicians (or nonexperts in general).

2.2. Managing the process of ANN training by using Genetic Algorithms

Genetic Algorithm (GA) is an iterative method for solving both constrained and unconstrained optimization problems (Yang, (2014, chap. 5 – Genetic algorithms)). The process of optimization starts from an initial guess of parameters (called population), which are the subject of optimization. At each iteration (called generation), the GA selects some portion of best individuals from the current population and uses them as parents to produce the

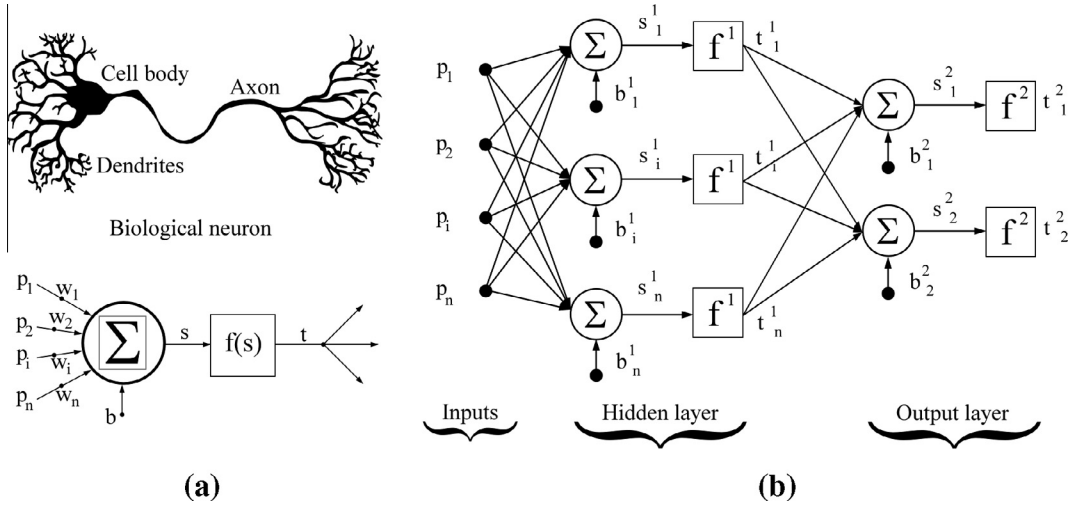

Fig. 1. ANN framework (a) single neuron, (b) considered architecture.

candidates (called children) for the next generation. Over successive generations, this process leads to the evolution of populations of individuals that are better and better adapted to their environment than the individuals that they originated from (just as in natural adaptation).

In this paper, GA were used to solve the problem of optimal and automated configuring of ANNs. The algorithm of the overall proposed procedure is shown in Fig. 2. The subject of GA optimization was the following ANN parameters: a type of ANN objective function, number of neurons in hidden layer, type of activation functions in layers, portion of data used for training, testing and validation, learning algorithm and learning momentum. Those parameters were passed to the GA objective function, within which ANNs were trained and evaluated iteratively. Therefore, the value of GA objective-function was the performance of the ANN, so that with every generation a better ANN performance was achieved until GA converged or reached a maximum number of iterations (generations). Regarding the ANN assessment, the following criteria were considered: Hosmer–Lemeshow goodness of fit test (HL test), accuracy, Brier score (BS) and Area Under receiver operating Characteristic (AUC) Hosmer & Lemeshow, 2000; Metz, 1978. However, any other criterion may be also used.

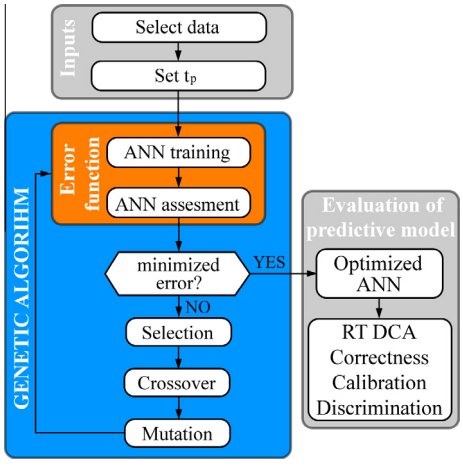

Fig. 2. Procedure for optimizing ANNs by using GA.

2.3. Regret Theory and Decision Curve Analysis approach for the assessment and evaluation of ANN predictive models

The output of the ANN classifications were probabilities t with the range approximately from 0 to 1. Elicitation of the threshold probability t_p was performed with respect to the Regret Theory (RT) (Djulbegovic, Hozo, Beckstead, Tsalatsanis, & Pauker, 2012; Hozo & Djulbegovic, 2008) as: $t_p = \left(1 + \frac{\text{consequence of FN}}{\text{consequence of FP}}\right)^{-1}$, where consequences are described on the scale 0 to 100 (0 means no consequence of wrong decision). Therefore, t_p represents the probability at which a clinician is indifferent between two strategies (treat or do not treat a patient). After setting t_p, it was assumed that probabilities higher than t_p were positive (TRUE), while probabilities less than t_p were assumed to be negative (FALSE). ANNs predictive performances were assessed by estimating: AUC, sensitivity, specificity, positive and negative predictive value, accuracy, Hosmer–Lemeshow statistic and the Brier score (mean square error). Following the previous steps, the four (in general it may be more or less) optimized ANNs were obtained and evaluated. However, the question which arises for clinicians is: "Which of the developed predictive models to select and within which ranges of probabilities?". For example, a new ANN predictive model increased specificity by 5% but decreased sensitivity by 3% compared to the currently used model – which one to use and how high an AUC is enough high for justifying clinical usefulness? Decision Curve Analysis (DCA) methods provide a solution by considering the clinical consequences and benefits of decision (Vickers & Elkin, 2006). The more the disease is vicious and the more the treatment is expensive or invasive – the consequences of wrong decisions are more harmful, just like in the case of BC. In brief, the RT approach of Decision Curve Analysis (DCA) describes the relationship between *regret* associated with "omissions" (e.g. failure to treat) vs. "commissions" (e.g. treating unnecessary) and decision maker's preferences as expressed in terms of threshold probability (Tsalatsanis, Hozo, Vickers, & Djulbegovic, 2010). In this paper, RT DCA was used for estimation of probabilities under which a decision maker tolerates a potentially wrong decision of a predictive model. It was assumed that after considering ANN model as an option during the treatment, for each patient in a cohort, the clinician faced three alternatives: treat patient, do not treat patient and use the ANN predictive model (treat patient if $p > t_p$). Computing associated Net Expected Regret Difference (NERD) for the three alternatives was performed following the algorithm (Tsalatsanis et al., 2010):

1. *for* the range of probabilities $P_t = p_{\min} \cdots p_{\max}$
2. for n samples in a dataset compute $output \begin{Bmatrix} true\ if\ (P_t > t_p) \\ false\ otherwise \end{Bmatrix}$
3. compute # *TP*, # *FP*, # *TN* and # *FN*
4. compute $NERD(Treat\ none,\ Model) = \frac{\#TP}{n} - \frac{\#FP}{n} \cdot \frac{P_t}{1-P_t}$
5. compute $NERD(Treat\ all,\ Model) = \frac{\#TN}{n} \cdot \frac{P_t}{1-P_t} - \frac{\#FN}{n}$
6. assuming that # *TP* and # *FP* is the number of patiens with and without disease compute $NERD(Treat\ none,\ Treatall) = \frac{\#TP}{n} - \frac{\#FP}{n} \cdot \frac{P_t}{1-P_t}$
7. plot the NERDs computed in steps 3, 4 and 5
8. *endfor*

where n is total number of patiens, *TP* is number of true positive, *FP* is number of false positive, *TN* is number of true negative and *FN* is number of false negative.

The optimal strategy at a particular probability is the one which brings the least regret in case it was proven wrong. An example of measuring reliability of a single predictive model on two different data sets is given in Fig. 3. The predictive model is recommended to be used on threshold probabilities $t_a > p > t_b$ for dataset 1, and $t_c > p > t_d$ for dataset 2. Since these two regions overlap, the overall trusted range of probabilities is estimated as their intersection $t_c > p > t_b$. By calculating and comparing NERDs for each of the developed predictive models and data sets, it could be estimated which ANN model is clinically the most useful and at which particular probabilities. Assuming that, it can be concluded that the most clinically robust and useful ANN predictive models are those which bring less regret and the widest trusted range of probabilities for clinicians.

3. Results

3.1. Population data

After obtaining the institutional review board approval, we retrospectively reviewed the medical records of 248 patients who had undergone radical surgery for BC at Military Medical Academy, Belgrade, Serbia. The data were collected over the 11-year study period from January 2002 until December 2012. The patients had undergone a routine cystoscopic and upper tract assessment, physical examination, transurethral resection of bladder tumour (TURBT), abdominal and pelvic computed tomography (CT) and chest radiography. After exluding patients with non-urothelial

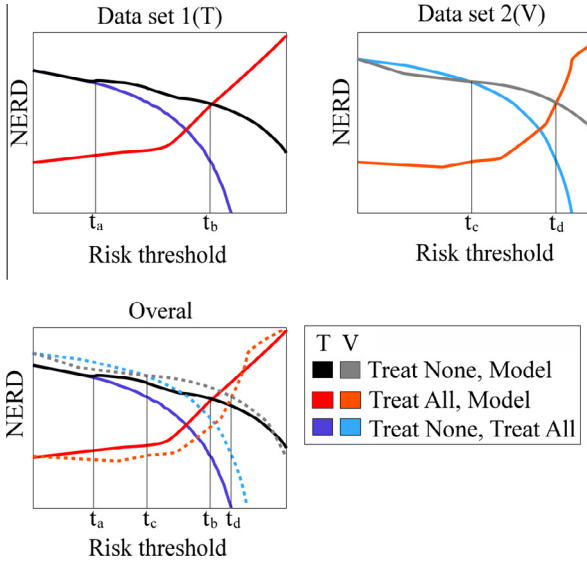

Fig. 3. Example of ANN predictive model evaluation on various data sets by RT DCA.

BC, distant metastatic disease, or those who had received radiotherapy or chemotherapy before RC, or incomplete data, total 183 patients were considered for this study. The following ten predictor variables were chosen for the defined outcome: demographic data (age, sex), TURBT findings (grade, stage, multiplicity of tumours, LVI), hydronephrosis, abdominal and pelvic CT radiography (size of the tumour, lymph node enlargement) and pathological stage after radical cystectomy (RC). Primary or secondary RC were defined as previously described (McLaughlin et al., 2007). Status of LNs on CT was categorized into three main groups. The first group represented patients with no LNs involvement or presence of LNs with a shot axis of <10 mm. The second group was considered when a nodal enlargement ⩾10–20 mm in the long axis was depicted and third group when LNs ⩾ 20 mm was found in the long axis (Eisenhauer et al., 2009). TUR was performed in 149 (81.4%) patients, median (interquartile range [IQR]; range) 2 (1; 1–12) months before RC. Overall, 109 (59.6%) patients had LN metastases, median 3 (IQR; range) (3; 1–11). Generally, 134 (73.2%) patients had advanced BC. In Table 1 complete characteristics of the cohort (ten were used as inputs and the rest were used for estimation of them) are given. The cohort of 183 patients was stratified into a training set (122 patients) and a validation set (61 patients) by random sampling. Baseline clinicopathological characteristics of all the patients, in derivation and validation set, are shown in Table 1. It can be seen that there were no statistically significant differences in derivation and validation sets. The primary interest of statistical analysis was the presence of advanced BC in surgical specimen, that was defined as pT3–4 tumour or presence of lymph node metastases after pathological review.

3.2. Findings

The implementation of presented procedure was performed by using Matlab R2010a (MathWorks, Natick, MA). Since Matlab already has the ANN and GA toolboxes, overall codig was reduced to a few hundred lines of code which reader could repeat following the paper and references. For the purposes of this study, according to the RT approach, the threshold probability during the ANN training was set to $P_t = 0.4$. It was assumed that the consequence of FN was 75 and the consequence of FP was 50 on the scale 0–100. The number of GA generations was set to 50 and the size of population was set to 100. The parameters considered for optimization had the following boundaries. Minimum number of neurons in hidden layer was set to half (5) and maximum was set to 3 × number of inputs (30). Considered activation functions were the following: positive hard limit transfer (hardlim), symetric hard limit (hardlims), logarithmic sigmoid (logsig), positive linear (poslin), linear (purelin), soft max (soft max), symetric sigmoid (tansig), triangular basis (tri bas). Considered training algorithms were: BFGS quasi-Newton backpropagation (BFGS), Bayesian Regulation backpropagation (trainbr), Cyclical order weight/bias training (trainc), Conjugate gradient backpropagation with Powell–Beale restarts (traincgb), Conjugate gradient backpropagation with Polak–Ribiere updates (traincgp), Gradient descent backpropagation (traingd), Gradient descent with adaptive lr backpropagation (traingda), Gradient descent with momentum (traingdm), Levenberg–Marquardt backpropagation (trainlm). Considered ANN error-functions were: Mean absolute error performance function (mae), Mean squared error performance function (mse), Sum absolute error performance function (sae), Sum squared error performance function (sse). GA converged at approximately 30 generations for all four criteria (AUC, Brier score, accuracy and HL statistic) and less than five minutes of execution on the PC with 2 GHz processor and 2 GB RAM memory using our database (Table 1). Configuration of ANNs optimized with respect to the four criteria are shown in Table 2. Proposed 1, Proposed 2, Proposed 3 and Proposed 4 were ANNs

Table 1
Clinicopathological characteristics of patients: all, test and validation set.

Characteristics	All (n = 183)	Test set (n = 122)	Validation set (n = 61)	P value
Age, years	63.4 ± 9	63 ± 9.6	64 ± 7.8	0.503
Gender, female/male, n (%)	17/166 (9.3)	12/110 (9.8)	5/56 (8.2)	0.719
Primary/secondary, n (%)	84/99 (45.9)	56/66 (45.9)	28/33 (45.9)	1.000
Size of tumors, cm[a]	4, 3	4, 3	4, 2.5	0.562
Number of tumors, 1, 2 or ⩾3, n (%)	28/28/127 (15.3/15.3/69.4)	17/23/82 (13.9/18.9/64.6)	11/5/45 (18/8.2/73.8)	0.156
TURBT no/yes (%)	34/149 (18.6/81.4)	25/97 (20.5/79.5)	9/52 (14.8/85.2)	0.347
Initial tumor grade 2 or 3, n (%)	22/156 (12/88)	14/108 (11.5/88.5)	8/53 (13.1/86.9)	0.748
Initial TUR stage Ta/T1/T2, n (%)	6/39/138 (3.3/21.3/75.4)	3/24/95 (2.5/19.7/77.9)	3/15/43 (4.9/24.6/70.5)	0.471
Lymphovascular invasion no/yes, n (%)	51/132 (27.9/72.1)	34/88 (27.9/72.1)	17/44 (27.9/72.1)	1.000
Time to TUR[a]	2, 1	2, 1	2, 1.5	0.618
Hydronephrosis no/unilateral/bill, n (%)	77/61/45 (42.1/33.3/24.6)	47/42/33 (38.5/34.4/27)	30/19/12 (49.2/31.1/19.7)	0.345
Status LN on CT n (%)	76/43/64 (41.5/23.5/35)	51/28/43 (41.8/23/35)	25/15/21 (41/24.6/34.4)	0.970
Pathological stage CIS/T1, T2, T3, T4 n (%)	9/52/73/49 (4.9/28.4/39.9/26.8)	6/38/43/35 (4.9/31.1/35.2/28.7)	3/14/30/14 (4.9/23/49.2/23)	0.328
OC/ABC stage n (%)	61/122 (33.3/66.7)	44/78 (36.1/63.9)	17/44 (27.9/72.1)	0.268
Removed LNs[a]	14, 8	14, 7	14, 8	0.160
Positive removed LNs[a]	2, 4	2, 4	2, 3.5	0.918
LNs positive n (%)	74/109 (40.4/59.6)	52/70 (42.6/57.4)	22/39 (36.1/63.9)	0.394
OC/ABC overall n (%)	49/134 (26.8/73.2)	36/86 (29.5/70.5)	13/48 (21.3/78.7)	0.238

[a] Median, interquartile range; CIS, carcinoma in situ; OC, organ-confined; ABC, advanced bladder cancer.

optimized with respect to BS, AUC, accuracy and HL test respectively.

In order to perform an additional analysis of correctness and relevance of the presented procedure, the obtained results were compared face to face with the results obtained with state of the art solutions. Three commonly used tools were tested on the same data sets: Logistic Regression implemented in SPSS version 20.0 (SPSS Inc., Chicago, IL), automated ANN toolbox in Statistica Automated Neural Networks 8.0 (StatSoft, Tulsa, OK) – 2000 ANNs were generated with all options enabled, and ANN Matlab R2010a Toolbox (MathWorks, Natick, MA) – when experienced user is allowed to configure ANN manually as it was described in Section 2.1. All of the developed predictive models were evaluated on both training and independent validation data sets and their prognostic performances are given in Table 3. ROC curves of the developed predictive models are shown in Fig. 4, where ANN Matlab model developed by an experienced user was considered as a reference for comparing proposed and commonly used solutions. To estimate clinical usefulness and robustness of the developed models, NERD values were calculated and analysed (Fig. 5).

4. Discussion

Considering the obtained results and their correctness, it may be seen that an experienced Matlab ANN (MANN) user is able to outperform Statistica Automated Neural Networks (SANN) and SPSS LR (SPSS LR). In order to achieve such results, we remind that one needs to devote a significant amount of time (since try-and-error approach must be used) as well as to develop good understanding of ANN framework and gain experience with Matlab (which does not present a problem if one uses both SPSS LR and SANN). In this paper, we prefered to use MANN as the reference for measuring improvements in prognostic performances (Fig. 4(b) and (d)). From the obtained ANN configurations (Table 2), it may be confirmed that there is no *golden rule* for configuring ANN predictive models. Moreover, it may be noticed that optimizing ANNs with respect to the different criteria leads to variations of other prognostic performances (Table 3). For example, architecture of Proposed 3 (P3) maximized accuracy but it showed tendency to overfit the model showing a very bad calibration. The most similar to P3 is SANN model (in terms of configuration), but due to the limitations of the SANN licence we were not able to perform HL test and RT DCA analysis. Since the same activation functions were used as the those obtained for the P3 model, we may guess (with reserve) that SANN model would have similar disadvantages. Proposed 1 (P1) model and MANN have a very similar architecture. However, it is shown that by varying the configuration, training algorithm, momentum and portion of data used for test-validation one could additionally increase the performances. Proposed 2 (P2) and Proposed 4 (P4) models showed that configuring ANN with respect to the AUC and HL statistic would result with the most reliable predictive models. Generally, from Table 3 it may be concluded that P2 model outperformed P4. However, it may be noticed in Fig. 5 that P4 has a wider range of trusted probabilities, which means that P4 is a more clinically robust compared to P2. It is found that robustness and calibration of P4, P2 and MANN are probably related to used activation functions (soft max and linear) which produce a wider range of probabilities (compared to the tansig and logsig which produce probabilities closer to 0 and 1). To sum up, for the considered problem we recommend to use P4. Also, we would like to highlight that it is not a *golden rule*. For different datasets or problems we do not suggest one to use P4, but to perform a complete analysis which should result in the best predictive model.

Regarding the novelty, to best of our knowledge, there is no similar study in literature since none was focused on the prediction

Table 2
Configurations of the developed ANN predictive models.

	Configuration	Hidden activation	Output activation	Error-function	GA assessment criterion	Training algorithm	Learning momentum	Train, validation, test
Proposed 1	10-29-2	softmax	tansig	mse	BS	traingd	0.88	80–10-10
Proposed 2	10-16-2	logsig	tansig	mse	AUC	trainscg	0.75	69–15-16
Proposed 3	10-26-2	tansig	tansig	sse	ACC	traingdx	0.33	85–10-5
Proposed 4	10-26-2	softmax	purelin	mse	HL	trainlm	0.59	72–16-12
ANN Matlab	10-25-2	softmax	tansig	mse	–	trainscg	0.66	70–15-15
Statistica AANN	10-12-2	tansig	tansig	sos	–	bfgs 30	–	70–15-15

Table 3
Prognostic performances of the developed predictive models.

	Data set	AUC	Sensitivity	Specificity	Youden's index	PPV	NPV	Accuracy	Brier score	HL-test	
										HL test, χ^2	*P* value
SPSS 20.0 LR	T	86.8	89.5	66.7	56.2	86.5	72.7	82.8	0.06	7.84	0.34
	V	86.5	81.2	84.6	65.9	95.1	55.0	81.9	0.12	6.59	0.09
Matlab R2010a ANN	T	89.4	96.5	66.7	63.2	87.4	88.9	87.7	0.10	4.15	0.14
	V	85.8	87.5	76.9	64.4	93.3	62.5	85.3	0.11	6.41	0.04
Statistica 8.0 AANN	T	90.1	90.6	72.2	62.9	88.6	76.4	85.2	–	–	–
	V	85.1	83.7	80.5	64.2	91.1	67.4	82.7	–	–	–
Proposed 1	T	95.1	95.3	77.7	73.1	91.1	87.5	90.1	0.07	5.5	0.29
	V	88.9	87.5	76.9	64.4	93.3	62.5	85.2	0.130	7.9	0.55
Proposed 2	T	90.6	100	83.3	83.3	93.4	100	95.0	0.054	2.54	0.04
	V	90.8	87.5	84.6	72.1	95.4	64.7	86.8	0.118	5.44	0.29
Proposed 3	T	96.1	97.6	91.6	89.3	96.5	94.2	95.9	0.044	12.1	0.85
	V	81.8	83.3	69.2	52.5	90.9	52.9	80.3	0.158	26.3	0.99
Proposed 4	T	0.95	0.90	0.72	0.62	0.88	0.76	85.1	0.085	1.62	0.01
	V	0.90	0.85	0.76	0.62	0.93	0.58	83.9	0.126	4.5	0.19

HL, Hosmer–Lemeshow; χ^2 Chi square; Data set (from Table 1): T-test; V-validation.

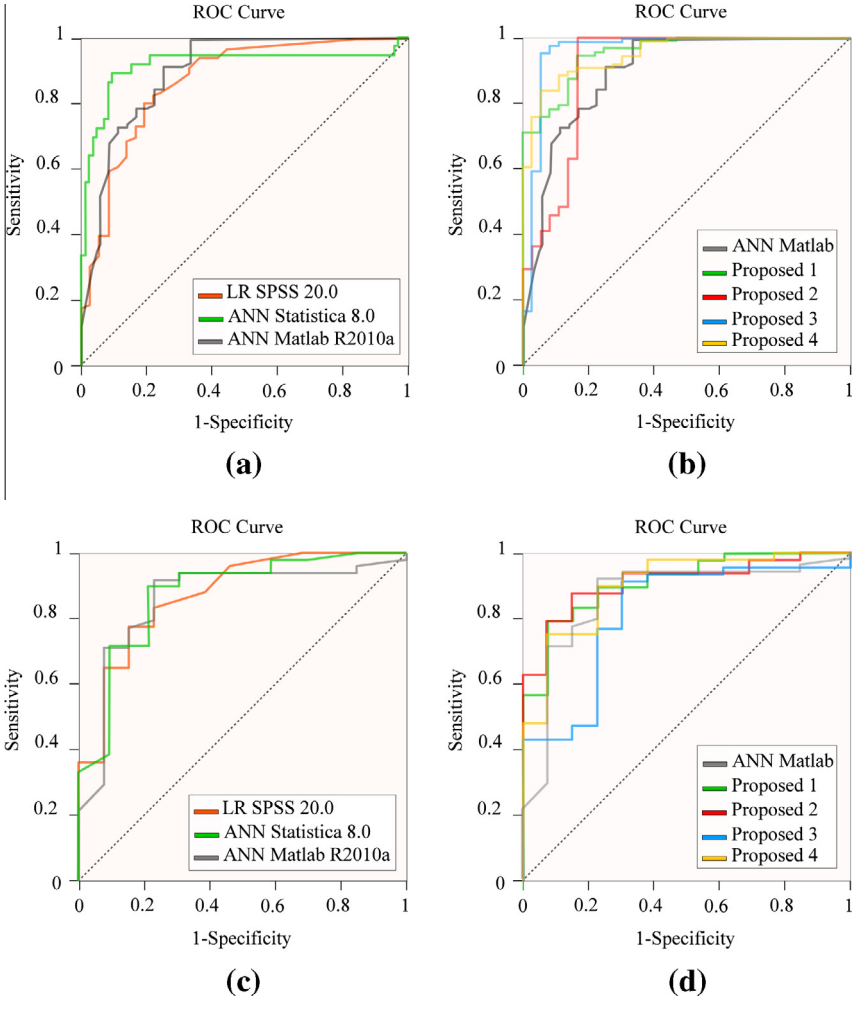

Fig. 4. ROC curves of the developed predictive models.

of BC outcome and clinical considerations of decision making which are covered by RT in the given study. However, we would like to point out that various evolutionary algorithms (EA) have been previously used to train the weights (Eisenhauer et al., 2009), design network architecture (Rivero et al., 2010) and select feature subsets (Askarzadeh & Rezazadeh, 2013). In the similar studies, the most frequently used criteria for measuring ANN performances are those related to maximizing their accuracy (Looney, 1993; Saritas et al., 2010), and compared to them here developed P3 model gained better performances. Taking this into account,

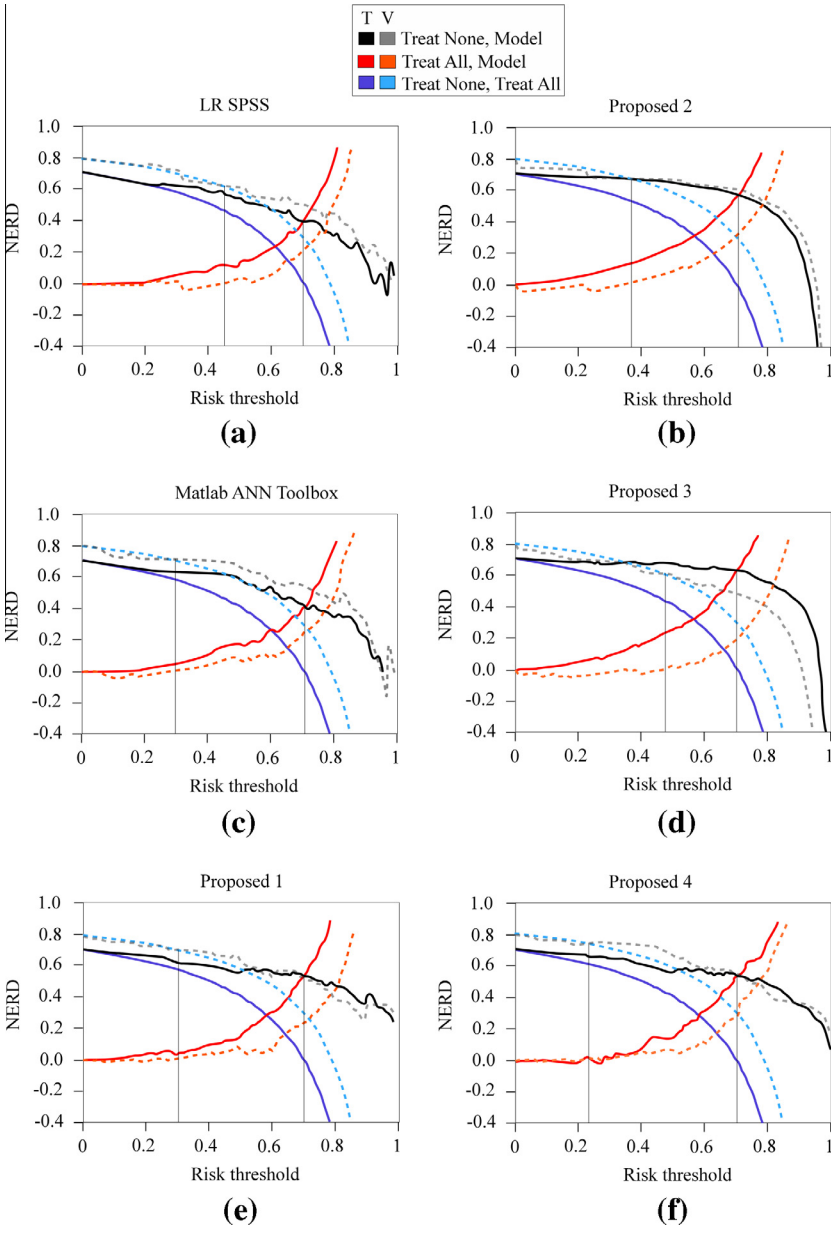

Fig. 5. NERD associated with the developed predictive models (Legend: T – test set, V – validation set).

one may conclude that we obtained better results compared to previous studies. However, since different databases and features were considered, for us, it was ungrateful to make such a conclusion. For that reason, instead of comparing obtained numerical results with those obtained in other studies we performed analysis on our dataset with the three solutions widely used in literature (please find that much of the literature referenced in this study used those solutions). Moreover, the focus of this study was not on maximizing the accuracy but the clinical utility which was not considered in the previous studies on the topic. To illustrate the contribution of the approach used in this study, we would like to consider the following example. Let us assume that one adopted the mean square error as a measure of correctness after setting a threshold probability (usually 0.5) and probabilities higher and lower than the threshold were assumed to be true and false, respectively. Since excepted values are zero or one in binary classification, such criteria have a tendency to train model to produce outputs which are close to the two binary states (trying at the same time to increase the number of correct predictions). Oppositely, for clinicians it is desirable to have predictive models which exhibit clinician behavior – such is minimizing regret when making decision (instead of producing binary outputs that are 0 and 1 which produce most harm in case they are proven wrong). Moreover, the most used criteria for assessing the performances of predictive models may be grouped into: (1) measuring correctness (for example BS or accuracy), (2) area under receiver operating characteristic curve for model discriminative ability and (3) goodness-of-fit for model calibration. Therefore, the question considered in this paper was: which assessment criterion is the most relevant for measuring ANN performances during the learning process and on which threshold probabilities for particular problem and data set the developed predictive model is relevant? In some situations it is more important to maximize sensitivity (minimize false negatives) than to maximize specificity. In prostate cancer screening, however, false positives are common and undesirable (expensive biopsy, harmful treatment, emotional impact). In our case, maximizing specificity was also important (moving toward the lower left corner of the ROC curve). As a solution, apart from

optimizing and evaluation of ANN with respect to the classical correctness analysis, clinical usefulness was assessed by using RT DCA. As a novel method not widely used in literature, RT DCA proved to be a reliable tool for measuring clinical robustness of predictive models. In situations when more predictive models have similar prognostic performance it is recommended to use the most robust one with the widest ranges of trusted probabilities.

Regarding the implications and clinical usefulness, beside this study a significant number of previous studies reported promising results, too (Anagnostou et al., 2003; Finnea et al., 2000; Hu et al., 2013; Lisboa & Taktak, 2006; Looney, 1993; Saritas et al., 2010; Çınar et al., 2009). Despite that, the application of ANN based decision support systems in clinical practice remains limited so far. For example, in situations when ANN does not outperform LR significantly, clinical audience would commonly prefer to use LR due to a few reasons. Commonly called black boxes, ANN models do not carry any real-life interpretation which could provide clinicians with intuitive association between explanatory variable and the outcome of interest. Methodological approaches used for ANN varied widely and there is a lack of a standard rule for optimally configuring ANN for a specific problem and dataset. Choosing the right configuration of ANN for fulfilling clinical needs seems as an open question, leaving to users unfamiliar with ANN to choose on their own among variant available activation functions, number of neurons and layers, learning rates, learning momentum, how much data to use for training-validation-testing, which criteria to use for measuring a quality of training etc. and in the end to choose between developed predictive models. In this paper, the principal contribution was avoiding user-system dependency. By reducing the process of using ANN to only selecting input and output data, the demands for learning complex foundation of ANN are significantly reduced while reliability of ANN is increased (as it was presented in the results). In order to use the proposed framework correctly, beside selecting data, a user (clinicians) needs to be familiar with BC treatment (which is assumed to be the case) in order to set the threshold probability with respect to the RT. In general, the main advantage of this study is the fusing of a clinically intuitive way of decision making with the ANN framework. RT DCA used for analysis of the developed predictive models helps clinical audience to obtain the most robust ANN models. In comparison with the alternative available solutions, the procedure used in this paper required less user-system interaction while better prognostic performances were achieved. Assuming that, it may be concluded that there is no barrier for a wider usage of the proposed procedure among clinical audience.

As limitations of this study, it has been previously mentioned that the database available for our purposes contained around 200 patients. In general, larger and more complete databases are desirable for further clinical studies based on the procedure described in this paper. However, since two independent datasets were used for training and validation, and intensive comparison with alternative solutions was performed, we may expect a similar response with larger data sets. Cross-validation of the ANN was not considered in this study, since two solutions used for benchmarking have not implemented such options. However, the existing procedure could be easily extended with new features. Moreover, the focus of this paper was to obtain a single ANN model and, of course, one could try to use various ensembling techniques (Zhou, Wu, & Tang, 2002) in order to combine different ANNs and create more robust classifiers (Cantú-Paz & Kamath, 2005).

5. Conclusion and future work

In this paper, the procedure for obtaining reliable ANNs predictive models was presented and applied for the purpose of predicting advanced BC in patients undergoing radical cystectomy. In order to enable clinicians to reach performances of an experts in obtaining ANN predictive models, various state of the art decision making methods were integrated. Testing on the ten years cohort study showed that soft max activation functions and good calibration were the most important for obtaining reliable BC predictive models for the given dataset. After comparing the obtained results with solutions commonly used in literature, it was shown that better prognostic performances were achieved while system-dependency of user was significantly reduced. Assuming that, it may be concluded that the the principal goal of the study was achieved. It is important to emphasize that a general procedure was presented and applied for the particular problem, which means that it may be used for different problems and studies as well. Also, developed ANNs represent optimal predictive models for a given dataset and considered problem. Therefore, the ANN models presented in this paper may not be the optimal choice for different problems or different data sets – but by using the described procedure it is expected that one would be able to obtain optimal and robust ANN predictive models for its own purposes with minimal effort (which is the principal contribution of this study). As mentioned in the discussion above, expected future research would be directed to application of the presented procedure to wide range of problems in medicine, data mining and engineering. Second, future extension with a new features such are ANN ensembling could additionally improve performances. Moreover, despite those considered in this study, various types of ANNs and criterions remains to investigate. Finally, an automated choosing an optimal combination of features from a larger input data set could additionally facilitate application of ANNs based expert systems. To sum up, presented procedure represents suitable, robust and a user-friendly framework with potential to have wide applications and influence in further development of health care decision support systems.

Acknowledgments

The authors were supported through a research Grants FP7-ICT-2007 project (Grant agreement 224297, ARTreat) of the EU and N0175014, III41007 and ON174028 of the Ministry of Science and Technological Development of Serbia. Also, we would like to thank two anonymous reviewers for very constructive suggestions which improved the present paper.

References

Anagnostou, T., Remzi, M., & Djavan, B. (2003). Artificial neural networks for decision-making in urologic oncology. *Reviews in Urology, 5*(1), 15–21.
Askarzadeh, A., & Rezazadeh, A. (2013). Artificial neural network training using a new efficient optimization algorithm. *Applied Soft Computing, 13*(2), 1206–1213.
Baker, S. G. (2009). Putting risk prediction in perspective: Relative utility curves. *Journal of the National Cancer Institute, 101*, 1538–1542.
Baker, S. G., & Kramer, B. S. (2012). Evaluating a new marker for risk prediction: Decision analysis to the rescue. *Discovery Medicine, 14*(76), 181–188.
Bassi, P., Sacco, E., De Marco, V., Aragona, M., & Volpe, A. (2007). Prognostic accuracy of an artificial neural network in patients undergoing radical cystectomy for bladder cancer: A comparison with logistic regression analysis. *BJU International, 99*(5), 1007–1012.
Bostrom, P. J., van Rhijn, B. W. G., Fleshner, N., Finelli, A., Jewett, M., Thoms, J., Hanna, S., Kuk, C., & Zlott, A. R. (2010). Staging and staging errors in bladder cancer. *European Urology Supplements, 9*(1), 2–9.
Cantú-Paz, E., & Kamath, C. (2005). An empirical comparison of combinations of evolutionary algorithms and neural networks for classification problems. *IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics), 35*(5), 915–927.
Castellani, M. (2013). Evolutionary generation of neural network classifiers – an empirical comparison. *Neurocomputing, 99*, 214–229.
Castellani, M. (2013). Evolutionary generation of neural network classifiers – an empirical comparison. *Neurocomputing, 99*(1), 214–229.
Çınar, M., Engin, M., Engin, E. Z., & Ateşçi, Y. Z. (2009). Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications, 36*(3), 6357–6361.

Compérat, E., & Van der Kwast, T. H. (2013). Pathological staging of bladder cancer. *Diagnostic Histopathology, 19*(10), 366–375.

Djulbegovic, B., Hozo, I., Beckstead, J., Tsalatsanis, A., & Pauker, A. G. (2012). Dual processing model of medical decision-making. *BMC Medical Informatics and Decision Making, 12*, 94. http://dx.doi.org/10.1186/1472-6947-12-94.

Dreiseitla, S., & Ohno-Machadob, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5–6), 352–359.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., et al. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer, 45*, 228–247.

El-Mekresh, M., Akl, A., Mosbah, A., Abdel-Latif, M., Abol-Enein, H., & Ghoneim, M. A. (2009). Prediction of survival after radical cystectomy for invasive bladder carcinoma: Risk group stratification, nomograms or artificial neural networks? *The Journal of Urology, 182*(2), 466–472.

Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., & Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications, 41*(9), 4434–4463.

Finnea, P., Finneb, R., Auvinenc, A., Juuselae, H., Arof, J., Määttäneng, L., et al. (2000). Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. *Urology, 56*(3), 418–422.

Ghaffari, A., Abdollahi, H., Khoshayand, M. R., Soltani Bozchalooi, I., Dadgar, A., & Rafiee-Tehrani, M. (2006). Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *International Journal of Pharmaceutics, 327*(1–2), 126–138.

Holmberg, L., & Vickers, A. (2013). Evaluation of prediction models for decision-making:7 beyond calibration and discrimination. *PLoS Medicine, 10*(7), e1001491. http://dx.doi.org/10.1371/journal.pmed.1001491.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley & Sons.

Hozo, I., & Djulbegovic, B. (2008). When is diagnostic testing inappropriate or irrational? Acceptable regret approach. *Medical Decision Making, 28*(4), 540–553.

Hu, X., Cammann, H., Meyer, H. A., Miller, K., Jung, K., & Stephan, C. (2013). Artificial neural networks and prostate cancer-tools for diagnosis and management. *Nature Reviews Urology, 10*(3), 174–182.

Jemal, A., Simard, E. P., Dorell, C., Noone, A. M., Markowitz, L. E., Kohler, B., et al. (2013). Annual report to the nation on the status of cancer, 1975–2009, featuring the burden and trends in human papillomavirus (HPV)-associated cancers and HPV vaccination coverage levels. *Journal of the National Cancer Institute, 105*(3), 175–201.

Karlik, B., & Olgac, A. V. (2010). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems, 1*(4), 111–122.

Lethaus, F., Baumann, M. R. K., Köster, F., & Lemmer, K. (2013). A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data. *Neurocomputing, 121*, 108–130.

Lisboa, P. J., & Taktak, A. F. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks, 19*, 408–415.

Looney, C. G. (1993). Neural networks as expert systems. *Expert Systems with Applications, 6*(2), 129–136.

Lughezzani, G., Briganti, A., Karakiewicz, P. I., Kattan, M. W., Montorsi, F., Shariat, S. F., & Vickers, A. J. (2010). Predictive and prognostic models in radical prostatectomy candidates: A critical analysis of the literature. *European Urology, 58*(5), 687–700.

Mallett, S., Haligan, S., Thompson, M., Collins, G. S., & Altman, D. G. (2012). Interpreting diagnostic accuracy studies for patient care. *BMJ, 344*, e3999.

McLaughlin, S., Shephard, J., Wallen, E., Maygarden, S., Carson, C. C., & Pruthi, R. S. (2007). Comparison of the clinical and pathologic staging in patients undergoing radical cystectomy for bladder cancer. *International Brazilian Journal of Urology, 33*, 25–31.

Meijer, R. P., Gemen, E. F. A., van Onna, I. E. W., van der Linden, J. C., Beerlage, H. P., & Kusters, G. C. M. (2009). The value of an artificial neural network in the decision-making for prostate biopsies. *World Journal of Urology, 27*, 593–598.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*(4), 283–298.

Mukherjee, I., & Routroy, S. (2012). Comparing the performance of neural networks developed by using Levenberg–Marquardt and Quasi-Newton with the gradient descent algorithm for modelling a multiple response grinding process. *Expert Systems with Applications, 39*(3), 2397–2407.

Ramy, F. Y., & Yair, L. (2011). Predictors of outcome of non-muscle-invasive and muscle-invasive bladder cancer. *The Scientific World Journal, 11*, 369–381.

Remzi, M., Anagnostou, T., & Ravery, V. (2003). An artificial neural network to predict the outcome of repeat prostate biopsies. *Urology, 62*, 456–460.

Rivero, D., Dorado, J., Rabuñal, J., & Pazos, J. (2010). Generation and simplification of artificial neural networks by means of genetic programming. *Neurocomputing, 73*(16–18), 3200–3223.

Saritas, I., Ozkan, I. A., & Sert, I. U. (2010). Prognosis of prostate cancer by artificial neural networks. *Expert Systems with Applications, 37*(9), 6646–6650.

Tallón-Ballesteros, A. J., & Hervás-Martínez, C. (2011). A two-stage algorithm in evolutionary product unit neural networks for classification. *Expert Systems with Applications, 38*(1), 743–754.

Tsalatsanis, A., Hozo, I., Vickers, A., & Djulbegovic, B. (2010). A regret theory approach to decision curve analysis: A novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making, 10*, 51. http://dx.doi.org/10.1186/1472-6947-10-51.

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574.

Wallis, W. (1999). Fundamentals of neural network modeling: Neuropsychology and cognitive neuroscience. *Brain, 122*(12), 2413–2416.

Yang, X. S. ((2014)). *Nature-inspired optimization algorithms*. Elsevier.

Yao, X., & Liu, Y. (1998). Towards designing artificial neural networks by evolution. *Applied Mathematics and Computation, 91*(1), 83–90.

Yu, X. H., & Chen, G. A. (1997). Efficient backpropagation learning using optimal learning rate and momentum. *Neural Networks, 10*(3), 517–527.

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Human–Machine Systems, 30*(4), 451–462.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence, 137*(1–2), 239–263.