Journal of Biomedical Informatics 52 (2014) 329–337

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

# Error-correction learning for artificial neural networks using the Bayesian paradigm. Application to automated medical diagnosis

Smaranda Belciug [a], Florin Gorunescu [b,*]

[a] Department of Computer Science, University of Craiova, Craiova 200585, Romania
[b] Department of Biostatistics and Informatics, University of Medicine and Pharmacy of Craiova, Craiova 200349, Romania

ARTICLE INFO

*Article history:*
Received 9 April 2014
Accepted 11 July 2014
Available online 21 July 2014

*Keywords:*
Automated medical diagnosis
Bayesian-trained neural networks
Breast cancer
Lung cancer
Heart attack
Diabetes

ABSTRACT

Automated medical diagnosis models are now ubiquitous, and research for developing new ones is constantly growing. They play an important role in medical decision-making, helping physicians to provide a fast and accurate diagnosis. Due to their adaptive learning and nonlinear mapping properties, the artificial neural networks are widely used to support the human decision capabilities, avoiding variability in practice and errors based on lack of experience. Among the most common learning approaches, one can mention either the classical back-propagation algorithm based on the partial derivatives of the error function with respect to the weights, or the Bayesian learning method based on posterior probability distribution of weights, given training data. This paper proposes a novel training technique gathering together the error-correction learning, the posterior probability distribution of weights given the error function, and the Goodman–Kruskal Gamma rank correlation to assembly them in a Bayesian learning strategy. This study had two main purposes; firstly, to develop a novel learning technique based on both the Bayesian paradigm and the error back-propagation, and secondly, to assess its effectiveness. The proposed model performance is compared with those obtained by traditional machine learning algorithms using real-life breast and lung cancer, diabetes, and heart attack medical databases. Overall, the statistical comparison results indicate that the novel learning approach outperforms the conventional techniques in almost all respects.



## 1. Introduction

Medical diagnosis refers to the act of identifying a certain disease analyzing the corresponding symptoms. From the point of view of biomedical informatics, medical diagnosis assumes a classification procedure involving a decision-making process based on the available medical data. Thus, the utilization of automated medical diagnosis systems aims to minimize the physician's error by taking advantages of both the intrinsic computation power when using a huge amount of data, and the fast processing speed as compared to that of the human. Such an "intelligent" system is fed with different symptoms and medical data of a patient, and, after comparing them with the observations and corresponding diagnoses contained in medical databases, will provide the most probable diagnosis based on the human knowledge embedded in the database.

From the machine learning (ML) point of view, an automated medical diagnosis may be regarded as a classification problem. Neural networks (NNs) have become a popular tool for solving such tasks [1]. In [2], NNs have been applied to predict the severity of acute pancreatitis at admission to hospital. A competitive/collaborative neural computing decision system has been considered [3] for early detection of pancreatic cancer. Different NNs have been applied in breast cancer detection [4].

Recent years have seen a large development of new approaches regarding NNs applied to the medical diagnosis. Hybrid NNs/genetic algorithms and partially connected NNs were used in breast cancer detection and recurrence [5,6]. NNs based on matrix pseudo-inversion have been applied in biomedical applications [7]. Swarm optimized NNs were used for detection of microcalcification in digital mammograms [8], and a fused hierarchical NN was applied in diagnosing cardiovascular disease [9].

The Bayesian paradigm could be used to learn the weights in NNs, by considering the concept of subjective probability instead of objective probability. NNs used as classifiers actually learn to

* Corresponding author. Address: Department of Biostatistics and Informatics, University of Medicine and Pharmacy of Craiova, 2 Petru Rares Str., Craiova 200349, Romania. Fax: +40 251 412 673.
*E-mail addresses:* smaranda.belciug@inf.ucv.ro (S. Belciug), gorunef@gmail.com, fgorun@rdslink.ro, florin.gorunescu@webmail.umfcv.ro (F. Gorunescu).

http://dx.doi.org/10.1016/j.jbi.2014.07.013
1532-0464/

compute the posterior probabilities that an object belongs to each class. Once the training data is presented to the NN, the posterior probabilities provide the measure that different weights are consistent with the observed data [10–12]. Some studies used Bayesian NNs to solve biomedical problems. In [13], a Bayesian framework for feed-forward neural networks to model censored data with application to prognosis after surgery for breast cancer has been proposed. A Bayesian NN has been used to detect the cardiac arrhythmias within ECG signals [14]. In [15], a Bayesian NN was able to provide early warning of EUSIG-defined hypotensive events.

Different from other approaches dealing with the Bayesian paradigm in conjunction with network models, the current work proposes a novel technique to update the synaptic weights in a multi-layer perceptron (MLP). The underlying idea is to use the error-correction learning and the posterior probability distribution of weights given the error function, making use of the Goodman–Kruskal Gamma rank correlation. The synaptic weights belonging to the unique hidden layer are adjusted inspired by the Bayes' theorem. Technically, in a subjective Bayesian paradigm, they are considered as posterior probabilities estimated using priors and likelihoods expressing only the natural association between object's attributes and the network output, or the error function, respectively, through the non-parametric Goodman–Kruskal Gamma rank correlation. The statistical comparison indicates that the novel learning approach outperforms the conventional techniques regarding both the decision accuracy and the computation speed. The main contributions of the paper are twofold: firstly, to develop a novel learning technique for MLP based on both the Bayesian paradigm and the error back-propagation, and secondly, to assess its effectiveness using real-world databases.

The remainder of this paper is organized in five sections. Section 2 is devoted to the presentation of both the design and implementation of the novel model, and the real-world datasets for the benchmark process. Section 3 presents the experimental results of applying the model to six real-world datasets in terms of performance analysis and performance assessment. Section 4 briefly summarizes the main characteristic of the novel approach, while Section 5 deals with the conclusions and future work.

## 2. Materials and methods

### 2.1. Training dataset – a probabilistic approach

The training dataset TS = $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ contains $N$ objects. Each object is coded as a vector $\mathbf{x}_k = (x_1^k, \ldots, x_i^k, \ldots, x_p^k; y_j)$; $x_i^k$, $i = 1, 2, \ldots, p$, represents the $i$th feature of the $k$ object, $k = 1, 2, \ldots, N$, and $y_j$, $j = 1, 2, \ldots, q$, represents the label of the decision class (category) $C_j$ the object $\mathbf{x}_k$ belongs to.

For each $k = 1, 2, \ldots, N$, the attribute values $x_i^k$ belonging to the attribute $A_i$, $i = 1, 2, \ldots, p$, are governed by a random variable (r.v.) $X_i$. Let $F_i(x)$ be the probability distribution of $X_i$, $i = 1, 2, \ldots, p$. Statistically, the set $\{x_i^1, x_i^2, \ldots, x_i^N\}$ represents a random sample of length $N$ corresponding to the r.v. $X_i$. One can consider, without loss of generality, the naïve assumption that all attributes are independent of each other, i.e., the parent r.v.'s $X_i$, $i = 1, 2, \ldots, p$ are independent.

For each object $\mathbf{x}_k$, the labels $y_j$, $j = 1, 2, \ldots, q$, are governed by a categorical r.v. $Y$, whose (categorical) distribution is denoted by $F(y)$. Statistically, the set $\{y_j^1, y_j^2, \ldots, y_j^N\}$ represents a random sample of length $N$ corresponding to the categorical r.v. $Y$.

### 2.2. Discovering knowledge in data

TS contains objects $\mathbf{x}_k$ characterized by input = features and output = category, providing valuable information in data, ready to be used in the learning phase. Since this information (*subjective* prior information) based upon data is available, the (subjective) Bayesian approach suggests its use to improve the way to find an acceptable solution [16].

Two relationships have to be quantified: (a) the connection between attributes and the decision classes, and (b) the connection between attributes and the network error existing in the training process.

A straight way to discover and use the potential information within data is to assess the statistical dependence between the parent r.v.'s $X_i$, $i = 1, 2, \ldots, p$, and either the decision class variable $Y$, or the network error $E(n)$ at step $n$, using measures of association [17]. Assuming a common case in real-world applications, that is a non-linear monotonic relationships between variables and the existence of many tied observations in data, we chose between the traditional non-parametric approaches (e.g., Spearman rank $\rho$, Kendall *Tau*, etc.) the Goodman–Kruskal Gamma rank correlation $\Gamma$, which is based on the difference between concordant pairs ($C$) and discordant pairs ($D$), and computed as $\Gamma = (C - D)/(C + D)$.

### 2.3. Bayesian learning paradigm

From a Bayesian point of view, a decision-making process naturally combines prior knowledge with information extracted from observations. Intuitively, given a hypothesis $h$, the data or evidence $D$, the posterior probability $P(h|D)$ of $h$ given $D$, the likelihood $P(D|h)$, the prior probability $P(h)$, and the evidence $P(D)$, then:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}. \tag{1}$$

Probabilistically speaking, the Bayes' formula is given by:

$$P\{A_i|B\} = \frac{P\{B|A_i\}P\{A_i\}}{\sum_{i=1}^{n} P\{B|A_i\}P\{A_i\}}, \quad P\{B\} > 0, \quad P\{A_i\} > 0, \quad i = 1, 2, \ldots, n, \tag{2}$$

where $B$ is an arbitrary event and $\{A_1, A_2, \ldots, A_n\}$ is a partition of the sample space $\Omega$. In this context, $P\{A_i|B\}$ represents the *posterior probability*, $P\{A_i\}$ represents the *prior probability*, $P\{B|A_i\}$ represents the *likelihood*, and $P\{B\}$ the *evidence*.

The idea behind the Bayesian classification is so that one can predict the class label of an object given its attributes values by the use of the Bayes' rule. Given an object with attributes $\{A_1, A_2, \ldots, A_n\}$, we wish to classify it in class $C$. Accordingly, we choose the class $C = C_k$ that maximizes $P\{A_1, A_2, \ldots, A_n|C_j\}$ [18].

In classification/decision-making problems, given an object with attributes $\{A_1, A_2, \ldots, A_n\}$, belonging to class $C$, one often assume the so-called *naïve Bayes* classification (*Idiot's Bayes*), stating the independence of attributes (obviously, a false assumption most of the time) for a given class $C$, namely:

$$P\{A_1, A_2, \ldots, A_n|C\} = P\{A_1|C\} \cdot P\{A_2|C\} \cdot P\{A_n|C\}. \tag{3}$$

### 2.4. MLP model

The main elements of a feed-forward NN (or a multi-layer perceptron), seen as a classification model, are:

- input vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ formed by $p$ feature components $x_i$;
- related output/response (multivariate) variable $y_j$, $j = 1, 2, \ldots, q$;
- (synaptic) weights $w_{ij}$, connecting the output of neuron $i$ to the input of neuron $j$, where neuron $j$ lies in a layer to the right of neuron $i$;
- activation non-linear function $f$, usually chosen sigmoidal.

Mathematically, MLP as classifier is seen as a computational framework for defining a non-linear mapping between a $p$-dimensional Euclidian input (feature) space and a $q$-dimensional Euclidian output (decision) space. It generally consists of three or more layers: an input layer, an output layer, and one or more hidden layers.

Popular examples of continuously differentiable non-linear activation functions commonly used in MLP are the logistic sigmoid and the hyperbolic tangent "tanh", the latter often preferred because it converges faster in many instances. Suitable values for the parameters of the above functions can be heuristically estimated [11,19,20]. Practitioners recommend the use of normalized inputs, i.e., average of input variables close to zero, instead of the original ones in order to increase the convergence speed [19]. An appropriate synaptic weights initialization allows the training algorithm to produce a good set of weights and may improve the training speed [11]; the non-parametric Goodman–Kruskal Gamma rank correlation between attributes and decision classes was used for the synaptic weights initialization. A key observation in its practical use, based on the *universal approximation theorem* applied to MLP, is that a network with a single hidden layer is sufficient to uniformly approximate any continuous function [11,20].

### 2.5. Bayesian-trained MLP characteristics

We introduce in this paper a novel error-correction learning strategy for MLP based on the Bayesian paradigm. The following features characterize the proposed Bayesian-MLP (B-MLP) model:

- Architecture: one hidden layer with the number of hidden units equaling the number of decision classes.
- Activation function: the hyperbolic tangent:

$$f(u) = 1.7159 \cdot \tanh\left(\frac{2}{3} \cdot u\right), \tag{4}$$

  recommended for its fast convergence [19];
- Presentation of training examples: normalized inputs, shuffled examples and batch training mode [19,20].
- Initialization: using the Goodman–Kruskal Gamma rank correlation between attributes and decision classes.
- Network output computed using the *winner-takes-all* paradigm: the neuron with the largest output value gives the decision class.
- Stopping criterion: the testing/generalization performance is adequate to the problem at hand.

The three-layer perceptron architectural graph is illustrated in Fig. 1.

### 2.6. Bayesian strategy for updating synaptic weights

The original error-correction learning refers to the minimization of a cost function, leading, in particular, to the commonly referred *delta rule*. The standard back-propagation algorithm applies a correction to the synaptic weights (usually, real-valued numbers) proportional to the gradient of the cost function.

In order to use the Bayesian paradigm to update the synaptic weights, we consider them from a different perspective. As it has been shown in literature [21], the prior probability and data from continuous quantities used in the standard Bayesian inference are, in reality, more or less fuzzy. The fuzziness refers, in some respects, also to the synaptic weights and the error involved in the neural network structure.

Inspired by both the fuzziness perspective involved in the Bayesian paradigm and the debate concerning *fuzziness vs.* (*subjective*) *randomness* in the Bayesian field [22], we distinctly tackled the problem. Rather than consider them as fuzzy data, we can reinterpret them from a (subjective) probabilistic point of view.

For each hidden neuron $HN_j$ belonging to the hidden layer, denote by $w_{ij}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, q$, the corresponding synaptic weight of the input attribute $x_i$ belonging to the feature vector $\mathbf{x}$. Assume that the real values of the synaptic weights $w_{ij}$ occurring in the learning process represent, from a statistical point of view, the values of a statistical variable, denoted by $W_{ij}$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, q$. From a probabilistic point of view, behind this statistical variable there is its parent r.v. denoted, naturally, $W_{ij}$. The network error is perceived in a similar way for the same reason.

Standard pre-Bayesian "training" of neural networks involves estimating the values for $w_{ij}$ to minimize the network error $E(D, w_{ij})$ given (training) dataset $D$, which implies some drawbacks [12]. An alternative approach, used in this paper, deals with the conditional probability $P(E|w_{ij})$ considered as 'likelihood' in Bayesian terms. Suppose that the events $A_{ij}$ corresponding to $W_{ij}$ provide a partition of the "weight space" $W$, and let $E(n)$ be the error of the network in iteration $n$. According the *total probability formula*, we have:

$$P\{E(n)\} = \sum_{i,j} P\{E(n)|A_{ij}\}P\{A_{ij}\}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q. \tag{5}$$

The standard synaptic weight refers to the strength of a connection between two units, perceived as a measure of this strength. From a *subjective Bayesianism*, the probability is interpreted as a measure/degree of belief [16,23]. Under these circumstances, assuming that the real-valued synaptic weights belong to the interval [0,1], the synaptic weights might be interpreted as probability-like measure encoding the strength of a connection; the stronger the strength, the larger the corresponding probability. In such a paradigm, we can use the Bayes rule in the updating process, by considering the synaptic weights as *posterior* (*probabilities*).

Accordingly, the values of $w_{ij}(n+1)$ are given by:

$$w_{ij}(n+1) = P\{A_{ij}|E(n)\} = \frac{P\{E(n)|A_{ij}\}P\{A_{ij}\}}{\sum_{i,j} P\{E(n)|A_{ij}\}P\{A_{ij}\}}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q, \tag{6}$$

where $P\{A_{ij}\}$ represents the *prior* (*probability*), $P\{E(n)|A_{ij}\}$ the *likelihood*, and $P\{E(n)\}$ the *evidence*.

A correlation-based approach is proposed to estimate both the *priors*, the *likelihood*, and the initial values of the synaptic weights, even though "*correlation does not always imply causation*". The correlation coefficient, whatever the type, is a measure of the relationship between two variables of interest, i.e., a measure of the strength/intensity of association, with values ranging from $-1$ to +1. Thinking the same way and taking into consideration the modulus $\Gamma$ of the coefficient, it might be interpreted in the Bayesian framework as probability-like measure encoding the strength of the relationship.

#### 2.6.1. Prior and likelihood estimation

Having in mind that the dilemma "*Objective vs. Subjective*" related to the choice of *priors* is still standing [24], we used a subjective-based approach to solve it. Assuming that the synaptic weights are naturally related to the attributes influence on the decision class, the *priors* $P\{A_{ij}\}$ are considered subjective (informative), expressing specific information about the object $\mathbf{x}$ through the correlation between attributes $X_i$ and decision $Y$. $P\{A_{ij}\}$ can be expressed by the rank correlation $\Gamma$ between $X_i$ and $Y$:

$$P\{A_{ij}\} = \Gamma(X_i, Y), \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q. \tag{7}$$

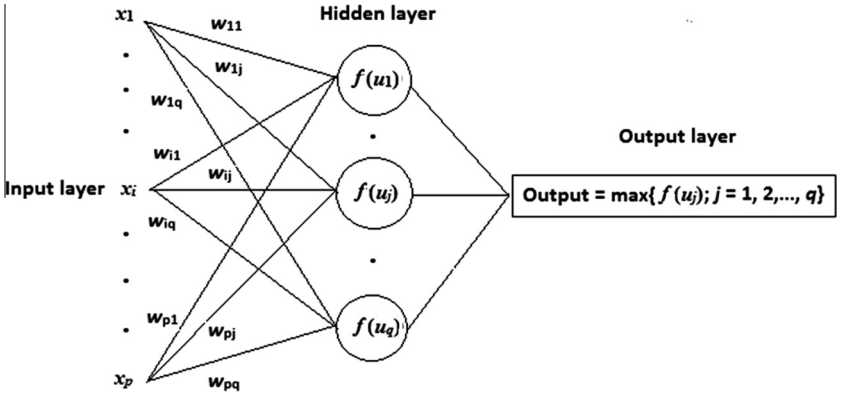


**Fig. 1.** Architectural graph of the three-layer perceptron (B-MLP).

The likelihood can be expressed in the same way by means of the rank correlation $\Gamma$ between $X_i$ and $E(n)$:

$$P\{E(n)|A_{ij}\} = \Gamma(X_i, E(n)). \tag{8}$$

*2.6.2. Backward computation*

The synaptic weights of the network are adjusted inspired by the Bayes' theorem, according to the formula:

$$w_{ij}(n+1) = w_{ij}^{*} = \frac{\Gamma(X_i, E(n)) \cdot \Gamma(X_i, Y)}{\sum_i \Gamma(X_i, E(n)) \cdot \Gamma(X_i, Y)}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q. \tag{9}$$

*2.6.3. Bayesian training algorithm*

We present below the steps of the training process of MLP inspired by the Bayesian paradigm.

*Algorithm (training B-MLP)*

1. For each decision class $C_j$, $j = 1, 2, \ldots, q$, and for each attribute $A_i$, $i = 1, 2, \ldots, p$, compute the corresponding mean attribute value $m_i^j$.
2. For each hidden neuron $HN_j$, $j = 1, 2, \ldots, q$, compute the synaptic weights $w_{ij}$ (initialization), given by:

$$w_{ij} = \Gamma((x_i^k - m_i^j), y_k), \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q, \quad k = 1, 2, \ldots, N. \tag{10}$$

3. For each hidden neuron $HN_j$, $j = 1, 2, \ldots, q$, compute the linear discriminant $u_j$, given by:

$$u_j = \sum_{i=1}^{p} \left( x_i^k \cdot w_{ij} \cdot \frac{1}{(x_i^k - m_i^j)^2} \right), \quad j = 1, 2, \ldots, q, \quad k = 1, 2, \ldots, N. \tag{11}$$

4. For each hidden neuron $HN_j$, $j = 1, 2, \ldots, q$, consider the non-linear activation function given by hyperbolic tangent:

$$f(u_j) = 1.7159 \cdot \tanh\left(\frac{2}{3} \cdot u_j\right), \quad j = 1, 2, \ldots, q. \tag{12}$$

5. For each decision class $C_j$, $j = 1, 2, \ldots, q$, encode the corresponding label $y_j$ using the "*1-of-q*" rule for nominal/categorical data, i.e., $y_1 \sim (0, 0, \ldots, 1)$, $y_2 \sim (0, 0, \ldots, 1, 0)$, $\ldots$, $y_q \sim (1, 0, \ldots, 0)$.
6. The hidden layer can be seen as a discrete random variable, whose distribution is characterized by a probability mass function $g$, which values are given by $g_j = g(f(u_j))$, via the formula:

$$g_j = \frac{\exp(f(u_j)) - \max_i\{f(u_i)\}}{\sum_{i=1}^{q}[f(u_i) - \max_i\{f(u_i)\}]}, \quad j = 1, 2, \ldots, q. \tag{13}$$

7. For each input item $\mathbf{x}_k$ of the training set TS, compute the corresponding error as follows:

$$error_k = \sum_{j=1}^{q} (y_j - g_j), \quad k = 1, 2, \ldots, N. \tag{14}$$

8. Build the error array $E = (error_1, error_2, \ldots, error_N)$, using the error at each step.
9. Update the synaptic weights according to the formula (9):

$$w_{ij}(n+1) = w_{ij}^{*} = \frac{\Gamma((x_i^k - m_i^j), E) \cdot \Gamma((x_i^k - m_i^j), y_j)}{\sum_i ((x_i^k - m_i^j), E) \cdot \Gamma((x_i^k - m_i^j), y_j)}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, q, \tag{15}$$

where $w_{ij}^{*}$ denotes the updated synaptic weight.

Repeat steps 3–9 for a certain number of epochs until the stopping criterion is satisfied.

Return the synaptic weights $w_{ij}^{*}$, which will be used in real-world applications.

*2.7. Medical datasets*

The proposed B-MLP model has been applied on six real-world medical datasets related to: breast cancer (3), lung cancer (1), diabetes (1), and heart attack (1), described below.

1. *Breast Cancer Wisconsin* (*Diagnostic*) – *BCWD* (*UCI Machine Learning repository*). BCWD consist of 569 cases, with two decision classes: benign 357 (62.74%) instances and malign 212 (37.25%) instances. From the total of thirty-two attributes, ten numerical attributes have been considered as the most relevant from medical point of view: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension (detailed description of the BCWD database at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer Wisconsin+%28Diagnostic%29).
2. *Breast Cancer – BC* (*UCI Machine Learning repository*). BC consists of 286 cases with two decision classes: non-recurrent-events 201 (70.27%) instances and recurrent-events 85 (29.72%) instances. The database contains nine mixed attributes, with three numerical attributes and six categorical attributes: age, tumor-size, inv-nodes, menopause, node-caps, deg-malig,

breast, breast-quad, irradiat (detailed description of the BC database at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer).

3. *Breast Cancer Wisconsin* (*Prognostic*) – *BCWP* (*UCI Machine Learning repository*). BCWP consists of 198 cases with two decision classes: non-recurrent-events 151 (76.26%) instances and recurrent-events 47 (23.73%) instances. From the total number of thirty-four attributes contained by the database, ten numerical attributes have been considered to be the most relevant from medical point of view: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension (detailed description of the BCWP database at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29).
4. *Lung Cancer – LC* (*UCI Machine Learning repository*). LC consists of 32 cases with three decision classes, three types of pathological lung cancers: 9 cases of type I, 12 cases of type II, 11 cases of type III. The database contains 56 ordinal (categorical) attributes (detailed description of the LC database at: http://archive.ics.uci.edu/ml/machine-learning-databases/lung-cancer/lung-cancer.names).
5. *Echocardiogram survival rate after heart attack – ECHO* (*UCI Machine Learning repository*). ECHO consists of 132 cases (one censored data) with two decision classes: died 88 (66.66%) instances and survived 43 (32.57%) instances. The database contains 10 numerical attributes: still-alive, uniformity of cell size, age-at-heart-attack, pericardial-effusion, fractional-shortening, E-point septal separation, left ventricular end-diastolic dimension, wall-motion-score, wall-motion-index, mult (detailed description of the ECHO database at: http://archive.ics.uci.edu/ml/machine-learning-databases/echocardiogram/echocardiogram.names).
6. *Pima Indian Diabetes – PID* (*UCI Machine Learning repository*). PID consists of 768 cases with two decision classes: tested negative for diabetes 500 (65%) instances and tested positive for diabetes (35%). The database contains 7 numerical attributes: number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, age (detailed description of the PID database at: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.names).

## 3. Results

The experiments on the six medical datasets, originating from different major diseases, aim to assess the new learning paradigm proposed for NNs against the traditional training approach. Moreover, the direct comparison of the performance of the novel algorithm with results obtained by other state-of-the-art techniques on the same datasets, proved without doubt its effectiveness and advantages regarding simplicity, computation speed and decision accuracy.

### 3.1. Performance analysis

In this subsection, the experimental results on the six medical datasets are presented, described and discussed.

Because the algorithm proposed in this study is of stochastic nature, it has to be independently run a certain number of times to obtain a reliable result regarding both its robustness and effectiveness. Statistically, the classification accuracy obtained during the multiple independent computer runs constitutes a sample of decision performance. In this respect, 100 independent computer runs provide an adequate statistical power (two-tailed type of null hypothesis with default statistical power goal $P \geqslant 95\%$, and type I error $\alpha = 0.05$). The model assessment has been achieved by using the standard 10-fold cross-validation [11]. Concretely, B-MLP has been independently executed 100 times (the model has been run 100 times in a complete 10-fold cross-validation cycle), and the average accuracy (training/testing) computed as the percentage of correctly classified cases (in both phases) has been considered as the decision performance. The experimental results are displayed in Table 1.

Depending on each dataset, the best training/testing accuracy obtained as the average over 100 independent complete cross-validation cycles ranges from 66.48% (LC) to 84.07% (BCWD), and 62.88% (LC) to 81.31% (BCWP), respectively. It is worth mentioning that the small difference between the training and the testing performance, regardless the dataset, proves the robustness of this algorithm.

A semi-graphical way to quantify the performance variability over the 100 independent computer runs is to consider the "box-and-whisker" plot of the testing accuracies, depicted in Fig. 2, clearly illustrating the performance of the network when used in the testing phase.

The above plot highlights low standard deviations values for all the breast cancer datasets, ranging from 3.65% (BCWD) to 4.11% (BCWP), confirming the stability of the algorithm in multiple computer runs, regardless the data type. The model behavior is changing when it is applied on the diabetes dataset and, especially, on the heart attack dataset, the standard deviation reaching 10.07%. This fact does not affect the above conclusion, but we can conclude that ML models could not offer an *omnibus* robustness regardless the problem at hand. Both the boxes (mean ± 1 times standard deviation) and the "whiskers" (mean ± 1.96 times standard deviation) have similar size, regardless the breast cancer dataset. It is worth mentioning in this context that, even BCWD and BCWP datasets refer to the same kind of attributes, their size are significantly different (569 vs. 198), and, much more important, the decision classes refer either to malign/benign type of tumor (BCWD-cancer detection), or to recurrent/non-recurrent events (BCWP-identification of cancer recurrence). Note that the lung cancer case does not differ significantly. Even if none of the performance samples is normally distributed (both Kolmogorov–Smirnov & Lilliefors/Shapiro–Wilk normality tests, $p$-level < 0.05), however the "whiskers" can somehow substitutes the "95% confidence interval", offering thus a synthetic illustration regarding the "compactness" of the decision performance (the scattering around the average accuracy).

### 3.2. Performance assessment

In order to make the entire performance evaluation clear, this sub-section is divided into two parts: (i) comparison with a relative similar Bayesian classifier, and (ii) comparison with standard classifiers. While the first approach focused on some important statistical measures of the performance in order to ensure no bias, the second is devoted to advanced statistical comparison tests able to clearly discriminate between the competitors' performances.

#### 3.2.1. Comparison with Naïve-Bayes classifier

Among some well-known Bayesian network classifiers, the Naïve-Bayes (NB) model is a good choice for comparison with B-MLP, since no structure learning procedure is required, it is particularly suited when the dimensionality of the inputs is high, and it has surprisingly outperformed many sophisticated classification models [25]. We used an implementation within Statistica 7 package – StatSoft Inc. The comparison results are shown in Table 2. To give more insight into the comparison process, we have provided both the corresponding classification accuracy (ACC) – highlighted in bold- and four important statistical measures of the performance: sensitivity (TPR), specificity (SPC), positive predictive value

**Table 1**
Experimental results: training/testing decision accuracy (%).

| Datasets | | | | | | |
|---|---|---|---|---|---|---|
| BCWD Train/test | BC Train/test | BCWP Train/test | LC Train/test | ECHO Train/test | PID Train/test |
| 84.07/81.14 | 77.53/76.99 | 83.03/81.31 | 66.48/62.88 | 81.34/79.04 | 78.10/76.17 |

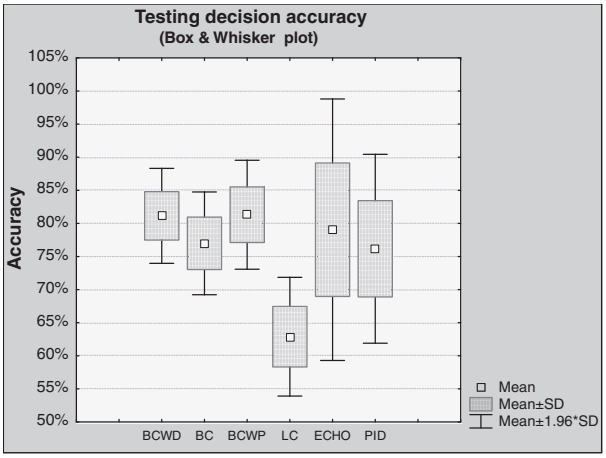


**Fig. 2.** Testing accuracy (box and whisker plot).

(PPV), and negative predictive value (NPV). Classification should be both sensitive and specific as much as possible, on the one hand, and, in addition, it is worth knowing both the proportion of cases with 'positive' test results who are correctly diagnosed (PPV), and the proportion of cases with 'negative' test results who are successfully labeled (NPV) [18]. Note that the LC dataset was partially considered since it involves three decision classes.

From the above table it can be seen that:

- Overall, the diagnosis accuracy of the novel approach exceeds the NB performance, irrespective of dataset.
- The sensitivity is relatively similar, slightly increased for the NB, showing that they have comparable capabilities to indicate disease among those with the disease.
- The specificity is significantly larger for B-MLP in the majority of cases, showing that B-MLP is better to indicate the lack of disease in case of negative screening test.
- For both PPV and NPV there is a similar situation to the specificity, showing that B-MLP manages better the balance between the screening tests probabilities and the existence or not of the disease.

*3.2.2. Comparison with standard classifiers*

Besides the comparison with NB, we also considered the comparison with other standard classification techniques. Several classifiers, such as: RBF, PNN, HPNN, MLP, kNN, optimal discriminant plane, RDA, SVM, PCNN, hybrid MLP/GA, and Cox regression had been tested against the UCI Machine Learning datasets regarding breast cancer, lung cancer, heart attack and diabetes used in this study. To evaluate the performance of the new approach, we compared it with the performance of these advanced models applied on the same datasets, described and reported in literature [26–34]. However, their performance displayed in Table 3 cannot be directly compared for all cases with the ones obtained by B-MLP (highlighted in bold) since the 10-fold cross-validation has not always been used, and ones of them have removed samples with missing data. Note that, due to the lack of consistent information in the case of ECHO and PID datasets, and for a standardization of the decision performance of MLP, PNN and RBF models, we solved this case by using the implementation within Statistica 7 package – StatSoft Inc.

According to this assessment, the novel algorithm provided a better performance compared with other ML approaches in most cases, excepting RBF, kNN and SVM (on BCWD dataset), and hybrid MLP/GA (on BCWD and BC datasets).

To assess the computational effort/time consumption of the new model vs. some basic NNs types (MLP, RBF, and PNN), we compared their CPU time during 100 computer runs (complete cross-validation cycle) when applied on the largest cancer dataset (BCWD), using a mid-level computer system, characterized by Intel(R) core(tm) i3-2330M CPU 2.20 GHz, 2 GB (RAM) – Table 4.

Inheriting the general architecture and the error-correction training procedure from MLP, but taking advantage of the new learning paradigm, the computational effort/time consuming of the new approach is lower than the MLP trained with the back-propagation algorithm. Based on this particular case, one can estimate that the computing speed has increased by about 28% due to the novel learning technique.

In conjunction with Table 3, the visual comparison of the decision performance of five ML techniques (B-MLP, RBF, PNN, MLP, and kNN), for which there is available information related to all the six medical datasets, is illustrated in Fig. 3.

The figure synthetically shows that the decision performance of the novel algorithm is comparable to other well-established techniques in the literature and exceeds in most of the cases the performance of its competitors.

To quantify statistically the magnitude of the contrast between the performance of the new approach and the performance of the competitors, the one-way ANOVA technique [35] along with the Tukey's honestly significant difference (Tukey HSD) *post hoc* test [36], and the Marascuillo procedure [37,38] were considered.

**Table 2**
Comparison between B-MLP and NB performances.

| Dataset | B-MLP | | | | | NB | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ACC (%)** | TPR (%) | SPC (%) | PPV (%) | NPV (%) | **ACC (%)** | TPR (%) | SPC (%) | PPV (%) | NPV (%) |
| BCWD | **81.14** | 71.86 | 87.31 | 79.04 | 82.33 | **78.98** | 74.19 | 88.88 | 93.24 | 62.50 |
| BC | **76.99** | 73.26 | 85.71 | 92.50 | 57.14 | **70.83** | 75.51 | 60.86 | 80.43 | 53.84 |
| BCWP | **81.31** | 80.27 | 84.31 | 93.65 | 59.72 | **69.38** | 80.55 | 38.46 | 78.37 | 41.66 |
| LC | **62.88** | – | – | – | – | **62.50** | – | – | – | – |
| ECHO | **79.04** | 74.00 | 92.00 | 94.87 | 63.88 | **78.94** | 76.92 | 83.33 | 90.90 | 62.50 |
| PID | **76.17** | 75.72 | 76.42 | 64.30 | 84.87 | **70.83** | 76.42 | 60.86 | 77.68 | 59.15 |

**Table 3**
Classification performance of other ML models compared vs. B-MLP.

| Model | Datasets | | | | | |
|---|---|---|---|---|---|---|
| | BCWD | BC | BCWP | LC | ECHO | PID |
| | Testing performance (%) | | | | | |
| **B-MLP** | **81.14** | **76.99** | **81.31** | **62.88** | **79.04** | **76.17** |
| RBF | 87.42 | 72.35 | 54.63 | 37.50 | 72.22 | 70.83 |
| PNN | 71.08 | 24.66 | 74.43 | 50.00 | 78.00 | 65.62 |
| MLP | 81.39 | 76.19 | 60.21 | 62.60 | 77.77 | 74.47 |
| kNN | 94.12 | 72.15 | 63.21 | 53.10 | 61.00 | 71.90 |
| Opt. dis. plane | – | – | – | 59.40 | – | – |
| RDA | – | – | – | 62.50 | – | – |
| HPNN | – | – | – | – | 76.00 | – |
| Cox regression | – | – | – | – | 60.00 | – |
| SVM | 96.92 | – | 74.21 | – | – | 76.30 |
| PCNN | 81.07 | 77.64 | 72.10 | – | – | – |
| MLP/GA | 93.58 | 80.42 | 81.11 | – | – | – |

**Table 4**
Comparing CPU time (BCWD dataset: MLP, RBF, PNN).

| B-MLP | MLP | RBF | PNN |
|---|---|---|---|
| 5′07″ | 7′03″ | 0′23″ | 0′58″ |

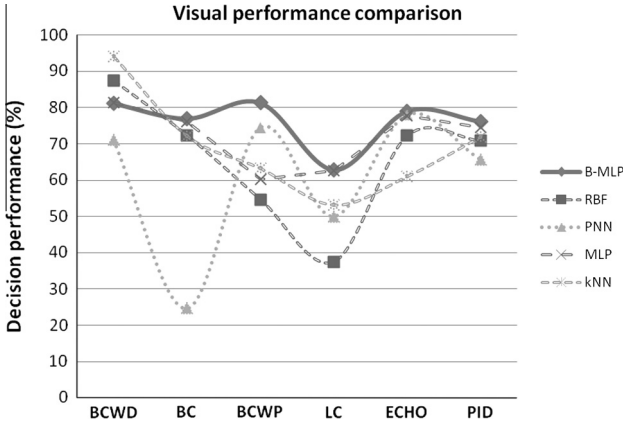


**Fig. 3.** Visual comparison of five algorithms performance.

The three tests have been used to investigate the existence of significant differences regarding the diagnosis performance of five ML techniques (B-MLP, RBF, PNN, MLP, and kNN) applied on all the six medical datasets under investigation, since complete information about the testing performance is available in these cases. Note that the underlying assumptions are fulfilled since the samples are independent and with equal (fairly) large size (balanced experiment). The one-way ANOVA along with Tukey HSD were performed using SPSS 16.0 package – SPSS.Inc, while the Marascuillo procedure was implemented by authors in MS Excel.

The ANOVA output consisted of sums of squares (SS), degrees of freedom (df), mean squares (MS), *F*-value, and *p*-level, and is presented in Table 5.

The *post hoc* Tukey HSD test has revealed statistically significant differences in classification performance ($p$-level < 0.05), with the following exceptions:

- *BCWD* dataset: B-MLP vs. MLP ($p$-level = 0.99).
- *BC* dataset: B-MLP vs. MLP ($p$-level = 0.62).
- *LC* dataset: B-MLP vs. MLP ($p$-level = 0.99).
- *ECHO* dataset:
  - B-MLP vs. PNN ($p$-level = 0.95);
  - B-MLP vs. MLP ($p$-level = 0.89);
- *PID* dataset: B-MLP vs. MLP ($p$-level = 0.38).

**Table 5**
Benchmark results: one-way ANOVA.

| Dataset | SS | df | MS | *F*-value | *p*-Level |
|---|---|---|---|---|---|
| BCWD | 2.91 | 4 | 0.73 | 476.40 | 0.000 |
| BC | 19.99 | 4 | 5.00 | 3145.37 | 0.000 |
| BCWP | 4.73 | 4 | 1.18 | 825.05 | 0.000 |
| LC | 4.39 | 4 | 1.09 | 551.28 | 0.000 |
| ECHO | 2.27 | 4 | 0.568 | 58.91 | 0.000 |
| PID | 0.65 | 4 | 0.163 | 36.41 | 0.000 |

**Table 6**
Benchmark results: one-way ANOVA: Marascuillo procedure.

| Dataset | Contrast | Value/test-statistics | Critical ranges | Significance |
|---|---|---|---|---|
| BC | B-MLP vs. PNN | 0.523 | 0.185 | Yes |
| BCWP | B-MLP vs. RBF | 0.266 | 0.194 | Yes |
| | B-MLP vs. MLP | 0.211 | 0.192 | |
| LC | B-MLP vs. RBF | 0.591 | 0.159 | Yes |

The above exceptions, excluding the case of B-MLP vs. PNN on ECHO dataset, confirm once again the MLP heritage of the novel model.

The Marascuillo procedure output consisted of the $k \cdot (k-1)/2$ test-statistics – the absolute value of the differences $\{|p_i - p_j|, i \neq j\}$, along with the corresponding critical ranges $r_{ij} = \sqrt{\chi^2_{1-\alpha,k-1}} \cdot \sqrt{\frac{p_i \cdot (1-p_i)}{n_i} + \frac{p_j \cdot (1-p_j)}{n_j}}$, where $k$ is the number of pairs, $p_i$ represent the proportions, $n_i$ are the sample sizes, and $\chi^2_{1-\alpha,k-1}$ represents the chi-square distribution with $\alpha$ level (default value $\alpha = 0.05$). The Marascuillo output displaying the statistically significant differences (i.e., test-statistics exceeding critical range) of B-MLP vs. its four competitors is presented in Table 6. Note that no statistical difference was observed on the other three datasets (BCWD, ECHO, and PID).

The results of the two tests, though different, are consistent with the raw information from Table 3, confirming statistically the contrast.

Both from the comparison with NB and with the eleven other classifiers it resulted that the expected behavior of the novel model was confirmed by the experimental evaluation, and the subsequent statistical analysis. Motivated by the promising way of the subjective Bayesian approach to support the identification of real solutions by introducing sufficient additional information, the proposed algorithm shows its capabilities in comparison to other well-established techniques in the literature, and exceeds in most of the cases their performance.

## 4. Discussion

A subjective Bayesian paradigm is proposed as learning methodology for neural networks, as opposed to the traditional techniques, and is subsequently used for the automated medical diagnosis. Bringing additional information to bear on a real-world problem, this approach may significantly improve the way to find an acceptable solution. In order to verify the significance of the results and to validate the model, a thorough statistical analysis has been performed.

The above study focused on two goals. First, we proposed a novel learning scheme for MLP, based on the Bayesian paradigm used in conjunction with the error-correction technique, seen as an alternative to the standard back-propagation training algorithm. Second, we assessed and validated it in real-world applications regarding breast cancer, lung cancer, heart attack and diabetes.

Different from other approaches dealing with the Bayesian paradigm, the current work proposes a novel technique to update the synaptic weights of MLP using the posterior probability distribution of weights through the error-correction learning. In the updating process, the synaptic weights are considered as posterior probabilities estimated using priors and likelihoods expressing the association between attributes and the network output, or the error function, respectively, through the non-parametric Goodman–Kruskal Gamma rank correlation.

For the performance assessment, the one-way ANOVA technique along with the Tukey HSD *post hoc* test, and the Marascuillo procedure were used. The novel approach proved to successfully combine the effectiveness of a feed-forward NN structure with the advantages of the Bayesian paradigm in providing a high performance diagnosis.

Apart from various theoretical arguments in favor of and against the algorithms used in the benchmarking process, the results presented in Tables 2 and 3 show clearly that the classification performance depends straight on the dataset through the corresponding attributes. Discussions on this topic have been the object of a significant amount of research in Statistics, Machine Learning and Data Mining [39]. Anyway, "*there is no free lunch theorem*" for this issue.

The proposed approach has shown its efficiency in its application to several medical datasets from different domains (breast and lung cancer, diabetes, and heart attack). Nevertheless, it is important to point out some characteristics of this approach that confers to it the quality of reliable and competitive challenger. Below, some remarks in this respect:

- The subjective Bayesian paradigm involves the use of a simple rank correlation coefficient in the learning process. On the contrary, other classifiers that are reported use much more complex approaches (e.g., gradient-descent, evolutionary techniques, etc.).
- The synaptic weights initialization is not at all an easy task when using the standard backpropagation algorithm [11,20]. The B-MLP model uses instead the simple Goodman–Kruskal Gamma rank correlation.
- The process of updating the synaptic weights is laborious for other competitors (MLP, RBF, PNN). The B-MLP algorithm uses just the Bayes' formula adapted for this particular case.
- No smoothness assumption regarding the activation function is needed as in BP.
- In comparison to other approaches, including the Naïve Bayes too, B-MLP is not sensitive to the data type (categorical/continuous predictors) since it involves the use of a rank correlation coefficient only.
- No assumptions on the distributions of features as in NB model are needed.
- In the learning process it no longer exists learning rate, momentum, convergence to a local minima, etc., as in BP.

Overall, the idea to use the Bayesian paradigm to train a MLP is straightforward, advantageous and handy in several aspects:

- The Bayesian approach for updating the synaptic weights is transparently presented.
- The assessment and validation process allows a neat and clear discrimination between competitors.
- The corresponding algorithm is easy to understand and implement.
- The model proved to be adaptable to a wide variety of medical decision problems, containing both numerical and categorical attributes.

## 5. Conclusions

Automated medical diagnosis, developed as a collaborative paradigm involving both medical knowledge and Artificial Intelligence methods, has become a very important interdisciplinary technology in health care, yielding fast accurate diagnoses obtained with low costs. The effectiveness of a novel ML algorithm, based on a MLP trained using the Bayesian paradigm in conjunction with the error-correction learning was investigated on the task of providing a reliable real-time decision support for the medical diagnosis. The model was validated in real-world applications regarding breast cancer, lung cancer, heart attack and diabetes. Its performance equaled or exceeded the results reported in literature.

Future research may lie in:

- The use of alternative approaches to the Goodman–Kruskal Gamma rank correlation.
- The use of alternative non-linear activation function.
- The use of alternative network output computation to the *winner-takes-all* paradigm.

## Acknowledgments

This work was supported by the strategic Grant POSDRU/159/1.5/S/133255, Project ID 133255 (2014), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007-2013. The BCWD and BCWP databases were obtained from the University of Wisconsin Hospitals, Madison. Thanks go to Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian for providing the data. The BC database was obtained from the Institute of Oncology, University Medical Centre, Ljubljana, Slovenia. Thanks go to M. Zwitter and M. Soklic for providing the data. For the LC database thanks go to Stefan Aeberhard, James Cook University of North Queensland, the donor of the data. For the ECHO database thanks go to Steven Salzberg, John Hopkins University, the donor of the data. For the PID database thanks go to Vincent Sigillito, John Hopkins University, the donor of the data.

## References

[1] Amato F, Lopez A, Pena-Mendez EM, et al. Artificial neural networks in medical diagnosis. J Appl Biomed 2013;11:47–58.
[2] Andersson B, Andersson R, Ohlsson M, Nilsson JJ. Prediction of severe acute pancreatitis at admission to hospital using artificial neural networks. Pancreatology 2011;11:328–35.
[3] Gorunescu F, Gorunescu M, Saftoiu A, Vilmann P, Belciug S. Competitive/collaborative neural computing system for medical diagnosis in pancreatic cancer detection. Expert Syst 2011;28(1):33–44.
[4] Kalteh AA, Zarbakhsh P, Jirabadi M, Addeh J. A research about breast cancer detection using different neural networks and K-MICA algorithm. J Cancer Res Ther 2013;9(3):456–66.
[5] Belciug S, Gorunescu F. A hybrid neural network/genetic algorithm system applied to the breast cancer detection and recurrence. Expert Syst 2013;30(3):243–54.
[6] Belciug S, El-Darzi E. A partially connected neural network-based approach with application to breast cancer detection and recurrence. In: Proc 5th IEEE conference on intelligent systems – IS, 7–9 July 2010, London, UK; 2010. p. 191–6.
[7] Cai B, Jiang X. A novel artificial neural network method for biomedical prediction based on matrix pseudo-inversion. J Biomed Inform 2014;48:114–21.
[8] Dheeba J, Selvi ST. A swarm optimized neural network system for classification of microcalcification in mammograms. J Med Syst 2012;36(5):3051–61.
[9] Sekar BD, Ming CD, Jun S, Xiang YH. Fused hierarchical neural networks for cardiovascular disease diagnosis. IEEE Sens J 2012;12(3):644–50.
[10] Rojas R. Neural networks. A systematic introduction. Berlin: Springer-Verlag; 1996.
[11] Bishop C. Neural networks for pattern recognition. Oxford University Press; 1995.
[12] Titterington DM. Bayesian methods for neural networks and related models. Stat Sci 2004;19(1):128–39.

[13] Lisboa PJG, Wong H, Harris P, Swindell R. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artif Intell Med 2003;28(1):1–25.
[14] Gao D, Madden M, Chambers D, Lyons G. Bayesian ANN classifier for ECG arrhythmia diagnostic system: a comparison study. In: Proc IEEE intl conference on neural networks, 2005, July 31 2005–August 4, Montreal, Canada, vol. 4; 2005. p. 2383–8.
[15] Donald R, Howells T, Piper I, et al. Early warning of EUSIG-defined hypotensive events using a Bayesian artificial neural network. Acta Neurochir Suppl 2012;114:39–44.
[16] Press J. Subjective and objective bayesian statistics: principles, models, and applications. 2nd ed. Wiley; 2003. <http://onlinelibrary.wiley.com/doi/10.1002/9780470317105.fmatter/pdf> [17.05.10].
[17] Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. 3rd ed. Chapman and Hall/CRC; 2004.
[18] Gorunescu F. Data mining. Concepts, models and techniques. Berlin, Heidelberg: Springer-Verlag; 2011/2013.
[19] LeCun Y, Bottou L, Orr G, Müller K-L. Efficient BackProp. Neural networks: tricks of the trade. Lect Notes Comput Sci 2012;7700:9–48.
[20] Haykin S. Neural networks. A comprehensive foundation. Prentice-Hall; 1999.
[21] Viertl R, Sunanta O. Fuzzy Bayesian inference. Forschungsbericht SM-2013-2, 1-11. Technische Universitat Wien; 2013.
[22] Kosko B. Fuzziness vs. probability. Int J Gen Syst 1990;17:211–40.
[23] Hajek A. In: Edward NZ, editor. Interpretation of probability. The Stanford encyclopedia of philosophy. Winter; 2012. <http://plato.stanford.edu/archives/win2012/entries/probability-interpret/>.
[24] Wagenmakers E-J, Lee M, Lodewyckx T, Iverson G. Bayesian evaluation of informative hypotheses (statistics for social and behavioral sciences). In: Hoijtink H, Klugkist I, Boelen P, editors. Bayesian versus frequentist inference. Springer; 2008. p. 181–207 [chap. 9].
[25] Cheng L, Greiner R. In: Stroulia E, Matwin S, editors. Learning Bayesian belief networks classifiers: algorithms and systems. LNAI, vol. 2056. Berlin, Heidelberg: Springer-Verlag; 2001. p. 141–51.
[26] Gorunescu F, Belciug S. Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization. J Biomed Inform 2014;49:112–8.
[27] Aeberhard S, Coomans D, de Vel O. Comparison of classifiers in high dimensional settings. Tech. rep. 92-02. Dept of Computer Science and Dept of Mathematics and Statistics, James Cook University of North Queensland; 1992.
[28] Aeberhard S, Coomans D, de Vel O. The performance of statistical pattern recognition methods in high dimensional settings. In: IEEE signal processing workshop on higher order statistics, Ceasarea, vol. 4; 1994. p. 14–6.
[29] Wilson RD, Martinez RT. Improved center point selection for probabilistic neural networks. In: Proc intl conference on artificial neural networks and genetic algorithms (ICANNGA'97); 1997. p. 514–7.
[30] Kan G, Visser C, Kooler J, Dunning A. Short, long prediction value of wall motion score in acute myocardial infarction. Brit Heart J 1986;56:422–7.
[31] Kinney E. Cox regression application communication report with TR-10-88. Harvard University, Center for Research in Computing Technology, Aiken Computation Laboratory; 1988.
[32] Stoean R, Preuss M, Stoean C, El-Darzi E, Dumitrescu D. Support vector machine learning with an evolutionary engine. J Oper Res Soc 2009;60(8):1116–22 [Special issue: data mining and operational research: techniques and applications. Kweku-Muata Osei-Bryson, Vic J Rayward-Smith Guest Eds.].
[33] Stoean R, Preuss M, Stoean C, Dumitrescu D. Concerning the potential of evolutionary support vector machines. In: Proc IEEE congress on evolutionary computation – CEC 2007, Singapore; 2007. p. 1436–43.
[34] Carpenter GA, Markuzon N. ARTMAP-IC and medical diagnosis: instance counting and inconsistent cases. Neural Netw 1998;11:323–36.
[35] Seltman H. Experimental design and analysis. <http://www.stat.cmu.edu/~hseltman/309/Book/chapter7.pdf> [chap. 7].
[36] Lowry R. One way ANOVA-independent samples. Vassar.edu; 1999–2013. <http://vassarstats.net/textbook/ch14pt2.html> [chap. 14/part 2].
[37] Glass GV, Hopkins BK. Statistical methods in education and psychology. 3rd ed. Boston: Allyn and Bacon; 1996.
[38] NIST SEMATEC. <http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm>.
[39] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.