Expert Systems with Applications 39 (2012) 10889–10897

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Elucidating clinical context of lymphopenia by nonlinear modelling

Ljiljana Majnaric[a], Marijana Zekic-Susac[b,*]

[a] University of J.J. Strossmayer in Osijek, Department of Family Medicine, Medical School Osijek, Strossmayerova 105, Osijek, Croatia
[b] University of J.J. Strossmayer in Osijek, Faculty of Economics, Gajev trg 7, 31000 Osijek, Croatia

ARTICLE INFO

Keywords:
Lypmphopenia
Outcome
Prediction
Neural networks
Multilayer perceptron
Decision tree

ABSTRACT

A nonlinear approach for detecting relative lymphopenia is suggested by using a health data record based on simple clinical parameters. Two classification methods, neural networks and decision trees, were applied to detect whether a patient has a positive or a negative lymphopenia outcome. Due to a large dimension of input space, a feature selection method was used in the pre-processing stage. All tested models were validated on the same out-of-sample dataset, and a 10-fold cross-validation procedure for testing generalization ability of the models was conducted. The models were compared according to their classification accuracy in the sense of the average hit rate, specificity and sensitivity. The results show that (1) the best neural network model slightly outperforms the decision tree model, (2) the reduced model provides even higher accuracy than the models with all available data, and (3) both methods similarly rank five important predictors of lymphopenia. The paper discusses the relevance of extracted features, and suggests some guidelines for further research.

1. Introduction

A decline in the absolute number of peripheral blood lymphocytes (absolute lymphopenia), in elderly people, has for a long time been recognized as being related with subsequent mortality (Bender, Nagel, Adler, & Andres, 1986). Lymphocytes are pivotely involved in the immune reaction development (Gray & Ahmed, 1996). Assessing their quantitative and qualitative changes may, in elderly people, provide information not only on the efficacy of their immune system, but also on disease conditions, as well as on responses to infection, or a therapy (Brown, 1997; Castle, 2000). For example, it was shown that the elderly are much more prone for developing a significant lymphopenia during the course of common bacterial infections, then younger subjects, and that the severity of lymphocyte depletion is in a good correlation with the prognosis (Proust, Rosenzweig, Debouzy, & Moulias, 1985). Except for some rare, well defined clinical conditions, the number of lymphocytes in the peripheral blood of younger subjects is generally constant, indicating the tight control of the rates of their generation and clearance (Berezne, Bono, Guillevin, & Mouthon, 2006; Caruso et al., 1997). Physiologic, interindividual variations, are likely to be influenced by hormones, homing factors and cytokines, whose production is highly genetically modified (Caruso et al., 1997; Hall et al., 2000).

Slightly decreased lymphocyte number, in elderly population, although proved as being of a major prognostic significance, is however still poorly clarified. In addition to genetic factors, aging of the immune system (immunosenescence), and accumulation of chronic diseases in people with advanced age, might also give a substantial contribution (Brown, 1997; Castle, 2000; Miller, 1996). Understanding clinical context of lymphopenia could fortify further research in aging diseases, including both, prognostic (prospective) studies, as well as basic and translational research engaged into mechanisms of aging. In this paper, we suggest using a systematic health data record based on simple clinical parameters collection, and nonlinear methods for data analysis, as an approach for the first step knowledge extraction on the issue. In this context, we used relative lymphopenia, percent (%) of lymphocytes in White Blood Cell Differential (WBCD), as a more simple surrogate outcome measure for lymphopenia.

The issue of recognizing relative lymphopenia is formulated here as a binary classification problem with the aim to classify patients into one of the two categories: negative (denoted as 0), if the percentage of lymphocytes in White Blood Cell Differential (WBCD) is greater than 35, and positive (denoted as 1), if it is less than or equal to 35. Regarding methodology, a nonlinear approach is selected such that two advanced classification methods were used that have proven their advantage over classical statistical methods in previous research (Apte & Weiss, 1997; Paliwal & Kumar, 2009): neural networks (NN) and decision trees (DT). Both methods are robust, they can deal with uncertain and noisy data, and are widely used in data mining in different domains, although there is lack of research on their usage in clinical medicine. For example, NNs outperformed discriminant analysis in categorizing firms according to wealth creation (St. John, Balakrishnan, & Fiet, 2000), while DT methodology was found successful in conjuction

[*] Corresponding author. Tel.: +385 31 224 400; fax: +385 31 211 604.
E-mail addresses: ljiljana.majnaric@hi.t-com.hr (L. Majnaric), marijana@efos.hr (M. Zekic-Susac).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.003

with data envelopment analysis (Lee, 2010). Heckerling et al. (2007) used NNs and genetic algorithms to predict urinary tract infection by using five variable sets. Their research showed that extracted variable set accurately predicted urine infection, and revealed some novel relationships between symptoms, findings, and infection. In recent literature, some authors suggest to use ensemble-based systems compounded of several classification methods, such as the decision support system for diagnosing cardiovascular disease based on aptamer chips proposed by Eom, Kim and Zhang (2009). Their system combines a set of four different classifiers (support vector machines, neural networks, decision trees, and Bayesian networks) with ensembles, and achieves a high diagnosis accuracy.

The main purpose of this paper was to investigate if neural networks and decision trees can discover hidden relationship among input and output variables within health data, and therefore be exploited in modelling purposes as a decision support tool to human experts in recognizing lymphopenia. The rest of the paper describes data and sampling procedure, followed by the methodology review and description of the modelling strategy used in our experiments. Results of the NN and DT models are reported and compared by statistical tests and in sense of specificity and sensitivity. The accuracy of models and some important predictors are discussed in order to provide guidelines for further research.

2. Materials and methods

2.1. Data collection

The research is based on original data collected in a primary health care in Croatia. The primary aim of the study was to show that by collecting health data systematically, that is, selected in a way to determine many aspects of the health-status of patients, and by using advanced computer-based techniques for data analysis – it could be possible to identify health disorders relevant, in elderly and chronically ill people, for the lymphopenia, which is still unresolved complex medical problem (Majnaric-Trtica & Vitale, 2011; Trtica-Majnaric, Zekic-Susac, Sarlija, & Vitale, 2010). Health parameters were collected on the basis of a total of 93 patients – those who gave their consent, out of 150 individuals requiring the influenza vaccine in the season 2003/2004. There were 35 male and 58 female patients, 50–89 years old (median 69), all having chronic medical conditions. Study protocol was approved by the local ethics committee.

Available input space consisted of a large number of variables (49 initial input variables), including: (1) physician's diagnoses of the main groups of chronic diseases, (2) anthropometric measures and (3) a large set of hematological and biochemical laboratory tests. Blood tests were chosen on the basis of two criteria: (a) to determine the main age-related pathogenetic changes and (b to be available in the primary health care system setting. Based on these criteria and after taking a careful search across the scientific literature, we performed blood tests indicating: (1) inflammation, (2) the nutritional status, (3) the metabolic status, (4) chronic renal impairment, (5) latent infections, (6) humoral immunity, and (7) the neuroendocrine status (Table 1) (Majnaric-Trtica & Vitale, 2011).

In this paper, we wanted to show that the same database can be used for solving different research tasks. For this purpose, we used other health parameter, from the collected database, as the target attribute. This time, the aim of the study was to elucidate the clinical context of lymphopenia, another unresolved complex medical uncertainty, showed to have negative impact on older people's health outcomes. The output variable denoted the percentage of lymphocytes in White Blood Cell Differential (WBCD), indicating relative lymphopenia. It was binary expressed in the form of two categories, where the 0 value denoted the category of lymphopenia (if the value was less than or equal to 35), while the value of 1 denoted normal lymphocytes proportion (if the value was greater than 35). The cut-off value of 35% was used on the basis of our previous results as being relevant for the antibody response to influenza vaccination, i.e. separating good from impaired immune system function (Majnaric-Trtica & Vitale, 2011).

All available input variables and their descriptive statistics (mean and standard deviation for continuous variables, frequencies for categorical variables) are presented in Table 1.

Model 1 was created on the basis of all available input variables, while model 2 was produced on the preselected set of variables extracted by the feature selection procedure based on the chi-square test and its corresponding *p*-value. This procedure extracted only five input variables with a significant *p*-value of the chi-square test. The selected variables were: (1) **NEU** (indicating% of neutrophils in WBCD), (2) **CLEAR** (indicating chronic renal impairment), (3) **MALIG** (indicating subjects with malignant diseases, selected according to the defined criteria), (4) **HbA1c** (indicating the average blood glucose concentration in the last three months, or, in a larger context – the insulin resistance degree), and (5) **HTC** (comparable with blood viscosity).

2.2. Sampling procedure

The total dataset was divided into three subsamples for purposes of NN training, cross-validation, and final testing. Due to the nature of NN learning, while estimating NN model, it is desirable to keep an equal distribution of each class in the training set. Therefore, the number of patients with negative and positive vaccine reaction was equal in the training and in the cross-validation set. The final test subsample is set aside to determine how well the model performs on the holdout data. Such sampling procedure was used while building the master NN model. The structure of subsamples is presented in Table 2.

In order to test the model generalization ability, the 10-fold cross-validation procedure was conducted on each NN, DT, and SVM model, where the random sampling is used such that randomly selected 10% of the data is used for testing, while the remaining 80% of data was used for training and cross-validating phases of the NN modeling.

3. Methods

3.1. Neural network methodology

Together with genetic algorithms, clustering algorithms, decision trees, and some other methods, artificial neural networks (NNs) are widely used in datamining for discovering hidden nonlinear relationships among data (St. John et al., 2000). They have been successfully used for classification, prediction, and association in different problem domains (Paliwal & Kumar, 2009). The main advantage of NNs is the ability to approximate any nonlinear mathematical function (Masters, 1995). A number of research revealed that NNs could serve as a valuable decision support tool in different areas including medicine due to their other advantages such as the ease of optimisation, cost-effective and flexible nonlinear modelling of large datasets, and accuracy for predictive inference (Yeh, Chi, & Hsu, 2010). The most common type of NN was used in this research – the multilayer perceptron (MLP), a feed forward network that is able to minimize the objective function by various algorithms, such as backpropagation, conjugate gradient, or Levenberg–Marquardt. The backpropagation algorithm is based

Table 1
Input variables and their descriptive statistics.

Variable no.	Variable code	Variable description	Descriptive statistics

1	DM	Diabetes mellitus	yes = 22.22%
			no = 65.55%
			IGT (Impaired glucose tolerance) = 12.22%

2	CVD	Cardiovascular diseases
		(myocardial infarction, angina, history of revascularisation, stroke, transient ischaemic cerebral event, peripheral vascular disease)
			yes = 26.67%
			no = 73.33%

3	HYPER	Hypertension	yes = 83.33%
			no = 16.67%

4	GASTR	Gastroduodenal disorders (gastritis. ulcer)	yes = 35.56% no = 64.44%

5	URIN	Chronic urinary tract disorders
		(recurrent cystitis in women. symptoms of prostatism in men)
			yes = 53.33%
			no = 46.67%

6	COPB	Chronic obstructive pulmonary disease	yes = 14.44%
			no = 85.56%

7	ALLER	Allergy (Rhinitis and/or Asthma)	yes = 10.00%
			no = 90.00%

8	OA	Osteoarthritis	yes = 24.44%
			no = 75.56%

9	DERM	Chronic skin disorders
		(chronic dermatitis dermatomycosis)
			yes = 55.56%
			no = 44.44%

10	MALIG	Malignancy	yes = 14.44%
			no = 85.56%

11	OSTEOP	Osteoporosis	yes = 32.22%
			no = 67.78%

12	PSYCH	Neuropsychiatric disorders
		(anxiety/depression Parkinson's disease cognitive impairments)
		(0 = no. 1 = yes) ?????
			yes = 43.33%
			no = 56.67%

13	CLEAR	Creatinine clearance	mean = 1.70
			stdev = 0.45

14	W/H	Waist/hip ratio	mean = 0.95
			stdev = 0.07

15	GLU	Fasting blood glucose	mean = 6.33
			stdev = 1.83

16	GLU2H	Blood glucose 2 hours after Oral Glucose Tolerance Test	mean = 7.33
			stdev = 3.63

17	HbA1c	Glycosilated Haemoglobin
		(showing average blood glucose during last three months)
			mean = 4.89
			stdev = 4.76

18	CHOL	Total cholesterol	mean = 6.17
			stdev = 1.38

19	TG	Triglycerides	mean = 1.83
			stdev = 1.30

20	HDL	HDL-cholesterol	mean = 1.45
			stdev = 0.38

21	SKINF	Triceps skinfold thickness	mean = 33.44
			stdev = 7.47

22	ALB	Albumin	mean = 46.11
			stdev = 3.17

23	CMV	Cytomegalovirus specific IgG	mean = 6.31
			stdev = 3.63

24	HPG	Helicobacter pylori specific IgG	mean = 68.02
			stdev = 63.90

25	HPA	Helicobacter pylori specific IgA	mean = 33.38
			stdev = 52.06

26	LE	Leukocytes	mean = 6.64
			stdev = 1.53

27	NEU	Neutrophils % in White Blood Cell differential	mean = 52.26
			stdev = 9.83

28	MO	Monocytes % in White Blood Cell differential	mean = 8.07
			stdev = 2.25

29	CRP	C-reactive protein	mean = 5.41
			stdev = 3.52

30	E	Erythrocytes	mean = 4.32
			stdev = 0.42

31	HB	Haemoglobin	mean = 134.56
			stdev = 12.60

32	HTC	Haematocrite
		(erythrocyte volume blood fraction)
			mean = 0.39
			stdev = 0.03

33	MCV	Mean cell Volume	mean = 91.09
			stdev = 5.07

34	FE	Iron	mean = 14.72
			stdev = 5.21

35	HOMCIS	Homocysteine	mean = 12.39
			stdev = 3.86

36	ALFA1	Serum protein electrophoresis	mean = 2.11

(*continued on next page*)

Table 1 (*continued*)

Variable no.	Variable code	Variable description	Descriptive statistics
			stdev = 0.25

37	ALFA2	Serum protein electrophoresis	mean = 6.23
			stdev = 1.02

38	BETA	Serum protein electrophoresis	mean = 8.43
			stdev = 0.94

39	GAMA	Serum protein electrophoresis	mean = 12.49
			stdev = 2.31

40	VITB12	Vitamin B12	mean = 286.29
			stdev = 160.90

41	FOLNA	Folic acid	mean = 20.51
			stdev = 8.48

42	INS	Insulin	mean = 23.38
			stdev = 17.34

43	CORTIS	Cortisol in the morning	mean = 374.09
			stdev = 122.76

44	PRL	Prolactin in the morning	mean = 124.86
			stdev = 122.36

45	TSH	Thyroid-stimulating hormone	mean = 2.05
			stdev = 2.65

46	FT3	Free triiodothyronine	mean = 5.46
			stdev = 0.54

47	FT4	Free thyroxine	mean = 13.99
			stdev = 2.24

48	IGE	IgE	mean = 138.25
			stdev = 249.14

49	ANA	Antinuclear antibodies	mean = 29.69
			stdev = 35.41

50	Output	Lymfophenia – % of lymphocytes in White Blood Cell Differential (WBCD) (0 = negative, if it is greater than 35) (1 = positive, if it is less or equal to 35)	0 = 51.11% 1 = 48.89%

Table 2
Sampling procedure in the master NN* and DT** models.

Sample	Lymfophenia 0 (negative)		Lymfophenia 1 (positive)		Total	
	Frequency	Proportion (%)	Frequency	Proportion (%)	Frequency	Proportion (%)
Train	27	50	27	50	54	100
Cross-validation	8	50	8	50	16	100
Test	14	60.87	9	39.13	23	100
Total	49	52.69	44	47.31	93	100

* = Neural network.
** = Decision tree.

on the deterministic gradient descent algorithm originally developed by Werbos in 1974, extended by Rumelhart, McClelland and the PDP Research Group (1986).

Typically, the architecture of the MLP neural network consists of at least three layers of processing units. The input layer of a NN consists of n input units $x_i \in R$, $i = 1, 2, \ldots, n$, and of randomly determined initial weights w_i usually from the interval $[-1, 1]$. Each unit in the hidden (middle) layer receives the weighted sum of all x_i values as the input. The output of the hidden layer denoted as y_c is computed by Masters (1995):

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where f is the activation function selected by the user, which can be logistic, tangent hyperbolic, exponential, linear, step or other. The computed output is compared to the actual output y_a, and the local error ε is computed. The error is then used to adjust the weights of the input vector according to a learning rule, usually the Delta rule (Masters, 1995). The above process is repeated in a number of iterations (epochs), where the gradient descent or other algorithm is used to minimize the error. In order to produce probabilities in the output layer, a softmax activation function is added for classification purposes. The output layer of all NN models in our experiments consisted of a binary variable (valued as 1 for the positive lymphopenia, and 0 for the negative lymphopenia). The backpropagation and conjugate gradient algorithms were tested, by varying the activation function in the hidden layer (sigmoid and tangents hyperbolic). The number of hidden units varied from 2 to 50. The NN structure and training time was optimized by a cross-validation procedure. The maximum number of training epochs was set to 1000. Overtraining is avoided by a split-sample process which alternatively trains and tests the network (using a separate test sample) until the performance of the network on the test sample does not improve for n number of iterations. The generalization ability of all three NN models is determined by a 10-fold crossvalidation procedure. The sensitivity analysis was performed on the test sample in order to determine the significance of input variables to the model.

3.2. Decision tree methodology

Decision trees i.e. classification trees are frequently used in datamining, due to its ability to find hidden relationships among data. Benchmarking NNs to decision trees is also present in previous research (Lee, 2010; Bensic, Sarlija, & Zekic-Susac, 2005). The aim of this method is to build a binary tree by splitting the input vectors at each node according to a function of a single input. The two algorithms are the most popular for building a decision tree: discriminant-based univariate splits, and classification and regression trees (CART or C&RT). CART algorithm was pioneered in 1984 by Breiman, Friedman, Olshen and Stone (1984). Questier, Put, Coomans, Walczak, and Vander Heyden (2005) summarized CART steps as: (1) assign all objects to root node, (2) split each explanatory variable at all possible split points, (3) for each split point, split the parent node into two child nodes by separating the objects with values lower and higher than the split point for the considered explanatory variable, (4) select the variable and split point with the highest reduction of impurity, (5) perform the split of the parent node into the two child nodes according to the selected split point, (6) repeat steps 2–5, using each node as a new parent node, until the tree has maximum size, and (7) prune the tree back using cross-validation to select the right-sized tree. The evaluation function used in this research for splitting is the Gini index defined as (Questier et al., 2005):

$$Gini(t) = 1 - \sum_i p_i^2 \tag{2}$$

where t is a current node and p_i is the probability of class i in t. The CART algorithm considers all possible splits in order to find the best one by Gini index. The CART style exhaustive search for univariate splits was used in our experiments, with Gini index, equal prior probabilities, and equal misclassification costs. Prune of misclassification error was used as the stopping rule, with minimum $n = 5$, and standard error rule = 1. The 10-fold CV procedure was used during the training phase in order to find the right-sized tree with the minimal CV cost.

3.3. Modelling strategy and objective function

As one of the aims of the paper was to extract important predictors of lymphopenia among a high-dimensional input space, two different modeling strategies were conducted in the paper and tested by both methods (NNs and DT): (1) strategy of including all available variables, resulting in NN-model 1 and DT-model 1, and (2) strategy of including only variables selected by a feature selection procedure based on the *chi-square* statistics and p value for each predictor variable ($p < 0.05$ was used as the criterion for selecting important variables in our experiments). The second strategy resulted in NN-model 2 and DT-model 2.

The above mentioned four models were initially developed on the single train set and tested on the same out-of-sample data (i.e. the test set) in order to compare the test performance of NNs and DT. The performance of all models is measured by the hit rate of class 0 (i.e. the "negative lyphopenia" – hit_0")), hit rate of class 1 (i.e. the "positive lyphopenia"), and the average hit rate (*ave hit*) according to:

$$hit_0 = \frac{c_0}{t_0}, \quad hit_1 = \frac{c_1}{t_1} \quad, ave\ hit = \frac{hit_0 + hit_1}{2} \tag{3}$$

where c_0 is the number of patients accurately predicted to have output 0, t_0 is the number of patients with actual (target) 0 output, c_1 is the number of patients accurately predicted to have output 1, and t_1 is the number of patients with actual output 1. The sensitivity and specificity ratios were computed according to (Simon & Boring, 1990):

$$sensitivity = \frac{c_1}{(c_1 + d_0)}, \quad specificity = \frac{c_0}{(c_0 + d_1)} \tag{4}$$

where d_0 is the number of false negatives (the number of patients falsely predicted to have output 0), and d_1 is the number of false positives (the number of patients falsely predicted to have output 1). The type I and type II errors were calculated in order to compare the cost of misclassification produced by the two models.

Many authors, such as Liu (1995) and Tourassi and Floyd (1997), emphasize that model selection should be performed on the basis of a generalization error. Some of the well-known methods for testing the generalization ability of the models are n-fold cross validation, jackknifing, bootstrapping (Finn, 2008) and round robin techique (Liu, 1995). All of them have the purpose to reduce the small-sample estimation bias and variance contributions (Liu, 1995; Tourassi & Floyd, 1997).

The cross-validation is used in this paper because it produces no statistical bias of the result since each tested sample is not the member of the training set. A 10-fold cross-validation procedure (or leave k cases out, where $k = 1/10$ of the total sample) is performed, and the average of the classification rate is computed, which is used to estimate the generalization error. This estimate is also used as the model selection criterion. The procedure of cross-validation is performed according to a slightly modified description of Masters (1995)), also used in Trtica-Majnaric et al. (2010) including the following steps: (1) the in-sample data were divided into 10 equally-sized independent subsamples, (2) each NN model is estimated 10 times, each time using the different set of 9 subsamples for training, and tested on 1 sample that was left out of training, (3) 10 different results were obtained for each NN model, (4) the average of 10 obtained results i.e. the average hit rate is computed. The generalization ability in our study is measured by the average hit rate, and the model with the highest average hit rate is selected as the best model. The described 10-fold CV procedure was performed on each of the four models: NN-model1, NN-model2, DT-model1 and DT-model2. The average hit rate, as well as the hit rate of negative and positive lymphopenia are computed on each of the 10 test samples in the 10-fold CV procedure. The average hit rate of all 10 NNs is used to select the best NN model.

4. Results

The MLP neural networks with logistic and tangent hyperbolic activation function were trained and tested on each model by using the sampling procedure presented in Table 2. The decision tree models were tested on the same samples as the NN models, in order to enable the comparison. Those results are noted here as "master" results due to the fact that they were obtained on a single data sample vs. the results of the 10-fold cross-validation procedure that were obtained on 10 samples. Due to a large number of variables comparing to the sample size, the DT model was obtained only for the selected set of variables, i.e. model 2. The master results of NN and DT models are presented in Table 3.

It can be seen from Table 3 that the highest average hit rate (85.32%) is obtained by the DT-model 2 using discriminant based univariate split algorithm. This model accurately classified 92.86% of negative patients (belonging to category 0), while the hit rate of category 1 (positive patients) was 77.78%. The average hit rate of the CART algorithm was 79.76%, and this algorithm had a lower hit rate of the positive patients (66.67%).

When the NN results are analyzed, it can be noticed that the NN model 2 – with selected variables has higher accuracy than the NN model 1 with all available variables. Both NN architectures (with logistic and with tangents hyperbolic activation function) produce the same average hit rate of 79% on the selected set of input variables. However, if the all available input space is used, the accuracy of the NN models decreases drastically to 54% with logistic, and 66% with tangents hyperbolic activation function.

It is obvious from the above that the usage of preselected set of input variables produces higher average hit rates than the usage of all input space.

Therefore, in further analysis, it is relevant to compare the prediction power of the NN model 2 and DT model 2. Since the results

Table 3
Master NN* and DT** results of model 1 and model 2.

NN algorithm	Model 1 (all available variables)			Model 2 (only selected variables)		
	Hit rate of negative group – 0 (%)	Hit rate of positive group – 1 (%)	Ave.hit rate (%)	Hit rate of negative group – 0 (%)	Hit rate of positive group – 1 (%)	Ave.hit rate (%)
MLP – logistic function	64.00	44.00	54.00	92.00	66.00	79.00
MLP – tangens hyperbolic function	64.00	66.00	66.00	92.00	66.00	79.00
Decision tree – discriminant based univariate splits	–	–	–	92.86	77.78	85.32
Decision tree – CART algorithm	–	–	–	92.86	66.67	79.76

* = Neural network.
** = Decision tree.

in Table 3 were obtained on the single data set, it would be interesting to explore the generalization ability of the NN and DT models. One of the standard methods of the generalization ability is the cross-validation. For that purpose, 10 random samples were generated from the initial data sample, following the procedure described in the Methodology section. In order to examine the generalization ability of the NN models tested, the 10-fold cross-validation procedure is conducted for the best NN and the best DT model from Table 3 (since the two NN architectures produced the same average hit rate, the MLP network with logistic function was used in the cross-validation procedure due to is simpler structure (lower number of hidden units), and the DT with discriminant based univariate splits was used). The results of the 10-fold cross-validation are presented in Table 4.

It can be seen from Table 4 that higher average hit rate across 10 samples is obtained by the NN model 2 (average hit rate of 88.65%). The average hit rate of the DT model 2 was only slightly lower (88.35%) indicating that the two methodologies do not differ much in their generalization abilities. The NN model 2 also produce higher average negative hit rate (92.60% comparing to 90.90%) showing that the NNs are more accurate in recognizing patients that do not have lymphopenia. The situation is different in recognizing patients with positive lymphopenia, since the average positive hit rate of DT model 2 (85.79%) is higher than the average positive hit rate of the NN model 2 (84.70%).

In order to further compare the performance of the best NN and DT models according to the way they classify patients, it is valuable to examine their sensitivity and specificity, as well as type I and type II errors. In order to calculate those ratios, the number of patients correctly and incorrectly classified into each category is observed in each of the 10 test samples used in the 10-fold cross-validation procedure, and the values of c_1, c_0, d_1 and d_0 are

Table 4
NN* and DT** models' results of the 10-fold cross-validation procedure.

Test sample in CV procedure	MLP with logistic function – NN model 2			DT with discriminant based algorithm – DT model 2		
	Ave. hit rate (%)	Hit rate of negative group – 0 (%)	Hit rate of positive group – 1 (%)	Ave. hit rate (%)	Hit rate of negative group – 0 (%)	Hit rate of positive group – 1 (%)
1	100.00	100.00	100.00	100.00	100.00	100.00
2	100.00	100.00	100.00	90.00	80.00	100.00
3	83.00	66.00	100.00	100.00	100.00	100.00
4	78.50	100.00	57.00	71.43	100.00	42.86
5	65.00	80.00	50.00	77.50	80.00	75.00
6	90.00	80.00	100.00	90.00	80.00	100.00
7	70.00	100.00	40.00	70.00	100.00	40.00
8	100.00	100.00	100.00	100.00	100.00	100.00
9	100.00	100.00	100.00	92.86	85.71	100.00
10	100.00	100.00	100.00	91.67	83.33	100.00
Ave. hit rate of 10 samples	88.65	92.60	84.70	88.35	90.90	85.79

* = Neural network.
** = Decision tree.

Table 5
Sensitivity, specificity, type I and type II errors of the best NN* and DT** models.

Test sample in CV procedure	MLP with logistic function – NN model 2				DT with discriminant based algorithm – DT model 2			
	Sensitivity	Specificity	Type I error (α)	Type II error (β)	Sensitivity	Specificity	Type I error (α)	Type II error (β)
1	1	1	0	0	0.67	1	0	0.33
2	1	0	1	0	1	0	1	0
3	1	0.67	0.33	0	1	1	0	0
4	0.5	0.8	0.2	0.5	0.43	1	0	0.57
5	1	0.8	0.2	0	0.75	0.8	0.2	0.25
6	1	0.8	0.2	0	1	0.8	0.2	0
7	0.4	1	0	0.6	0.4	1	0	0.6
8	1	1	0	0	1	1	0	0
9	1	1	0	0	1	0.86	0.14	0
10	1	1	0	0	1	0.83	0.17	0
Average of 10 samples	0.89	0.81	0.19	0.11	0.82	0.83	0.17	0.18

produced and used in formula (2) to compute the sensitivity, specificity, type I and type II errors in each of the 10 samples, as well as their average values, as shown in Table 5.

Assuming that the average ratios obtained on 10 sample are those that could be expected in model deployment phase, the further analysis will be focused on those average ratio presented in the last row of Table 5.

It can be seen that the NN model's sensitivity is 0.89, while the specificity is 0.81, indicating that this model classifies 89% of patients that were actually positive into the class of positive ones. The specificity represents the ratio of actually negative (0 category) patients that the NN model recognized as the negative ones. It implies that the NN model is more sensitive than specific. When the type I and type II errors are observed, it could be noticed that the average type I of the NN model (0.19) is higher than the average type II error (0.11), indicating that this model produces more false positives than false negative patients, i.e. tends to misclassify more patients that actually had negative lypmhopenia into the category of positive group than vise versa. Such behavior of the model is desirable if the aim of the prediction process is to identify more positive patients, which is the case in almost all medical diagnosis problems, where the cost of not recognizing a positive patient is higher than accepting an actually negative patient as a positive one.

If the ratios of the DT model in Table 5 are observed, it can be noticed that the average specificity of the DT model (0.83) is higher than the specificity of the NN model (0.81), indicating that the DT model is more specific than sensitive, and that it is even more specific than the NN model. The average type I error of the DT model is 0.17, which is lower than the average type I error of the NN model (0.19), while the situation is the opposite concerning the type II errors, since the DT model has a higher type II error than the NN model (0.18 comparing to 0.11). It can be concluded when comparing the two models that the NN model is more sensitive, while the DT model is more specific.

The model with a high sensitivity could be used for screening for the disease, since it has tendency to misclassify more patients into the group of positive ones. The model with a high specificity could be used for confirming the test results, since it is more specific in recognizing the actual positive patients. From the medical point of view it is more useful to have model that has smaller type II error, therefore it would be reasonable to suggest the NN model as the more successful one.

In order to investigate the importance of each input variable to the output, the sensitivity analysis is performed on each NN model in the 10-fold cross-validation procedure, and the average sensitivity ratios of each input variable are presented in Fig. 1.

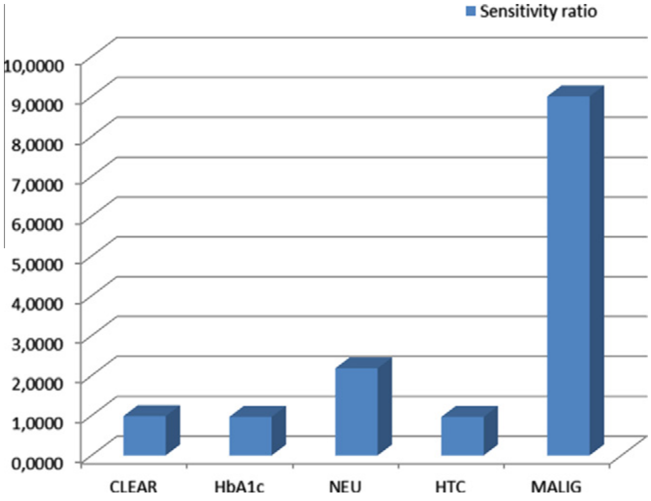

Fig. 1. Average sensitivity ratio of the best NN model obtained on 10 samples.

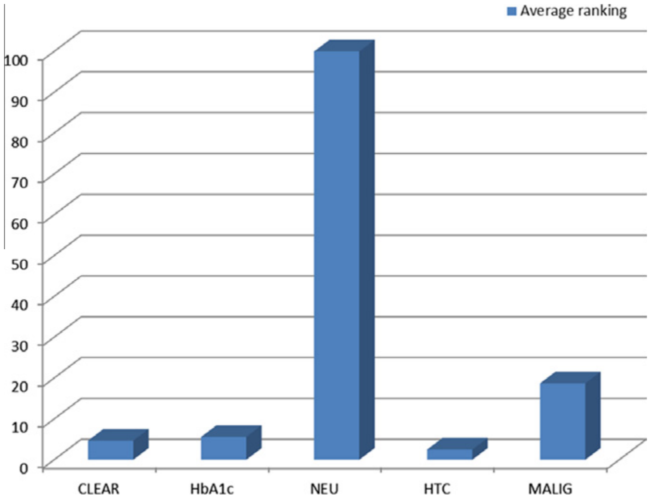

Fig. 2. Average input variable importance of the best DT model obtained on 10 samples.

It can be seen from Fig. 1 that the highest sensitivity ratio in obtained for the input variable named **MALIG** which indicates subjects with malignant diseases, and its influence to the output is significantly higher than the influence of other four preselected inputs. The second most important predictor is **NEU** (% of neutrophils in WBCD), followed by **HbA1c** (average blood glucose concentration in the last three months, or the insulin resistance degree), **CLEAR** (chronic renal impairment), and **HTC** (comparable with blood viscosity).

The selected DT model resulted with slightly different coefficients of predictors' influence. Fig. 2 presents the average ranking of input variables obtained by the best DT model 2 on 10 samples in the 10-fold cross-validation procedure.

The DT model gives the highest importance to the **NEU** variable (% of neutrophils in WBCD), while the second-ranked variable is **MALIG** (subjects with malignant diseases). Thus, the NN and DT models emphasized the same two most important predictors, but in different order. Also, the next group of predictors ranked by the DT model 2 consists of the same three variables as identified by the NN model 2, and they are equally ordered by the both methodologies. They are: **HbA1c** (average blood glucose concentration in the last three months, or the insulin resistance degree), **CLEAR** (chronic renal impairment), and **HTC** (comparable with blood viscosity).

The above analysis reveals that both NN and DT models recognize and highlight a high influence of the same two variables (**MALIG** and **NEU**) comparing to the lower influence of the **Hb1A1c**, **CLEAR**, and **HTC**.

5. Discussion on clinical relevance of the results

The results of the NN model 1, based on using all parameters in the performed dataset as the input, generally show that chronic diseases probably exert their effects on the lowering the number of lymphocytes in the peripheral blood by operating via many intermediate, biochemical mechanisms, since the majority of the data, used for modelling, provide the same and small contribution to the model's validity (Fig. 1). However, operating with too much parameters could be inconvenient from the practical point of view. In addition, a large input dimension, especially when combined with the small sample, can produce noisy effects which can reduce the generalization ability of the model, as it is shown on our example (Table 3, Model 1). This approach, however, by using a systematic health data record and a small sample, is likely to be appropriate for the computer-based simulation purposes, since it is possible, by using this approach, to identify, in relatively easy and the short-time manner, health disorders with the highest relevance for the issue. In relation to this, preselection of parameters from the dataset, prior entering the process of modelling, resulted in much better model's performances (Table 4, Model 2).

As with respect to the selection abilities of two different nonlinear methods used, MLP algorithms and decision trees, our results indicate a high comparability of these methods (Table 3 and Table 4, NN models and DT models). Namely, the both methods selected five identical predictors as most informative for the defined outcome measure. Moreover, these selected predictors are ranked in the same order; that is as follows: (1) **NEU** (indicating% of neutrophils in WBCD), (2) **MALIG** (indicating subjects with malignant diseases, selected according to the defined criteria), (3) **HbA1c** (indicating the average blood glucose concentration in the last three months, or, in a larger context – the insulin resistance degree), (4) **CLEAR** (indicating chronic renal impairment), and (5) **HTC** (comparable with blood viscosity).

The top-ranked variable, **NEU**, when comparing our results with the existing knowledge, is likely to indicate increased percent of neutrophils in the peripheral blood. In this sense, decreased percent of lymphocytes (relative lymphopenia) and increased percent of neutrophils in WBCD, can be coupled and reflective of the switch from the specific to non-specific and cellular immune response – a condition frequently found in older population (Castle, 2000; Franceschi, Bonafe, & Valensin, 2000). In relation to this, this might be a marker of the weakening of the immune system efficiency. This assumption is supported by the recent finding showing lymphopenia and neutrophil-lymphocyte count ratio as better predictors of the onset of sepsis, in patients with acute bacterial infection, than routine infection parameters, such as WBC and neutrophil count, or CRP concentration (de Jager et al., 2010).

Other selected parameters are likely to indicate clinical conditions most closely associated with lymphocyte count depletion. One of them is cancer (indicated by the parameter **MALIG**), already recognized as a cause of a significant lymphopenia, due to aggressive therapy, malnutrition and strong engagement of the immune system in a host defense against cancer (Caruso et al., 1997; Proust et al., 1985). Although a reverse situation, when lymphopenia precedes the development of cancer, is possible, it is less known, and should be considered in future research. Chronic renal failure, in a stage of haemodialysis, is also known as a major cause of lymphopenia, mostly due to increased susceptibility of lymphocytes for apoptosis (Matsumoto et al., 1995). In addition, our results point out mild chronic renal impairment in older patients with common chronic diseases, as to significantly contribute to lymphopenia (parameter **CLEAR**) (Fig. 1). Another selected parameter, **HbA1c**, probably indicates the insulin resistance state. This is, in our results, additionally supported by the upper-ranked parameters, **HYPER** and **GLU2h**, in the NN model 1, generated by using all parameters in the dataset (not shown in the results). Namely, the insulin resistance state, a blockade of insulin action in peripheral tissue, is a common background disorder which compromises several cardiovascular risk factors, including also hypertension and impaired glucose tolerance, indicated by these selected parameters (DeFronzo & Ferrannini, 1991). These risk factors might contribute to lymphopenia by affecting aggregability of blood cells, leading to marginalisation of lymphocytes from the blood flow (Fusman et al., 2001). Activation of neutrophils, which can simultaneously occur, might, by realising reactive oxidative and proteolytic substances, facilitate senescence, or apoptosis of lymphocytes (Kristal et al., 1998). These all events can elevate blood viscosity, which, in our results, is likely to be indicated by the parameter **HTC** (Meiselman, 1999). These results, by pointing out insulin resistance and elevated blood viscosity as disorders cordially involved in generating lymphopenia, are completely new findings, not mentioned in the scientific literature ever before. These achievements argue in favor of the usefulness of the approach that we suggest here, that is, a systematic health data record and a computer-based data modelling. In this way, by using simple, low-cost and widely available health parameters, many hidden relationships, not easily detectable in clinical studies, can be identified. Are these disorders, insulin resistance and elevated blood viscosity, joint to impaired renal function, as existing knowledge is likely to suggest, it is not quite clear from our results and deserve further investigation (Landray et al., 2001).

6. Conclusion

The paper deals with recognizing relative lymphopenia by using a systematic health data record based on simple clinical parameters collection, and nonlinear methods such as artificial neural networks and decision trees. Relative lymphopenia is expressed as the percent of lymphocytes in White Blood Cell Differential (WBCD). Both methods were used to classify patients into a positive or a negative category, while the modelling was performed in two stages: (1) on the whole input space, and (2) on the preselected set of input variables extracted by a feature selection procedure. Different neural network and decision tree algorithms were trained, tested, and validated on the same out-of-sample dataset, and a 10-fold cross-validation procedure for testing generalization ability of the models was conducted. The best NN and DT models were selected on the basis of the average classification rate. Although the NN model slightly outperformed the decision tree model, both models similarly ranked five important predictors that could be suggested to take into most consideration while deciding about the lymphopenia existence. This research reveals that there is a potential of intelligent classification methods in supporting medical diagnosis, while the future research could be focused on confirming obtained results on more independent datasets.

References

Apte, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems, 13*, 197–210.

Bender, B. S., Nagel, J. E., Adler, W. H., & Andres, R. (1986). Absolute peripheral blood lymphocyte count and subsequent mortality of elderly men. *Journal of the American Geriatrics Society, 34*(9), 649–654.

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modeling small business credit scoring using logistic regression, neural networks, and decision trees. *Intelligent Systems in Accounting, Finance and Management, 13*(3), 133–150.

Berezne, A., Bono, W., Guillevin, L., & Mouthon, L. (2006). Diagnosis of lymphocytopenia. *Abstract Presse Medicale, 35*(5Pt2), 895–902 [in French].

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wardsworth Inc.

Brown, K. A. (1997). Nonmalignant disorders of lymphocytes. *Clinical and Laboratory Science, 10*(6), 329–335.

Caruso, C., Bongiardina, C., Candore, G., Cigna, D., Romano, G. C., Colucci, At, et al. (1997). HLA-B8, DR3 haplotype affects lymphocyte blood levels. *Immunological Investigations, 26*(3), 333–340.

Castle, C. S. (2000). Clinical relevance of age-related immune dysfunction. *Clinical Infectious Diseases, 31*, 578–585.

de Jager, C. P. C., van Wijk, P. T. L., Mathoera, R. B., de Jongh-Leuvenink, J., van der Poll, T., & Wever, P. C. (2010). Lymphocytopenia and neutrophil-lymphocyte count ratio predict bacteriemia better than conventional infection markers in an emergency care unit. *Critical Care, 14*(R192), 108–115.

DeFronzo, R. A., & Ferrannini, E. (1991). Insulin resistance. A multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia and atherosclerotic cardiovascular disease.. *Diabetes Care, 14*(3), 173–194.

Eom, J.-H., Kim, S.-C., & Zhang, B.-T. (2009). AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Systems with Applications, 34*, 2465–2479.

Finn, O. J. (2008). Cancer immunology. *New England Journal of Medicine, 358*, 2701–2715.

Franceschi, C., Bonafe, M., & Valensin, S. (2000). Human immunosenescence: The prevailing of innate immunity, the failing of clonotypic immunity and the filling of immunological space. *Vaccine, 18*, 1717–1720.

Fusman, R., Rotstein, R., Berliner, S., Elishkewich, K., Rubinstein, A., Izkhacov, E., et al. (2001). The concomitant appearance of aggregated erythrocytes, leukocytes and platelets in the peripheral blood of patients with risk factors for atherothrombosis. *Clinical Hemorheology and Microcirculation, 25*, 165–173.

Gray, D., & Ahmed, R. (1996). Immunological memory and protective immunity: understanding their relation. *Science, 272*, 54–60.

Hall, M. A., Ahmadi, K. R., Norman, P., Snieder, H., MacGregor, A. J., Vaughan, R. W., et al. (2000). Genetic influence on peripheral blood T lymphocyte levels. *Genes and Immunity, 1*, 423–427.

Heckerling, P. S., Canaris, G. J., Flach, S. D., Tape, T. G., Wigton, R. S., & Gerber, B. S. (2007). Predictors of urinary tract infection based on artificial neural networks and genetic algorithms. *International Journal of Medical Informatics, 76*(4), 289–296.

Kristal, B., Shurtz-Swirski, R., Chezar, J., Manaster, J., Levy, R., Shapiro, G., et al. (1998). Participation of peripheral polymorphonuclear leukocytes in the oxidative stress and inflammation in patients with essential hypertension. *AJH, 11*, 921–928.

Landray, M. J., Thambyrajah, J., McGlynn, F. J., Jones, H. J., Baigent, C., Kendall, M. J., et al. (2001). Epidemiological evaluation of known and suspected cardiovascular risk factors in chronic renal impairment. *American Journal of Kidney Diseases, 38*(3), 537–546.

Lee, S. (2010). Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls. *Decision Support Systems, 49*, 486–497.

Liu, Y. (1995). Unbiased estimate of generalization error and model selection in neural network. *Neural Networks, 8*, 215–219.

Majnaric-Trtica, Lj, & Vitale, B. (2011). Systems biology as a conceptual framework for research in family medicine; use in predicting response to influenza vaccination. *Primary Health Care Research & Development*, 1–12. http://dx.doi.org/10.1017/S146342361100008.

Masters, T. (1995). *Advanced algorithms for neural networks. A C++ sourcebook*. New York: John Wiley & Sons.

Matsumoto, Y., Shinzato, T., Amano, I., Takai, I., Kimura, Y., Morita, H., et al. (1995). Relationship between susceptibility to apoptosis and Fas expression in peripheral blood T cells from uremic patients – a possible mechanism for lymphopenia in chronic renal failure. *Biochemistry and Biophysics Research Communications, 215*(1), 98–105.

Meiselman, H. J. (1999). Hemorheologic alterations in hypertension: Chicken or egg? *Clinical Hemorheology and Microcirculation, 21*(3-4), 195–200.

Miller, R. A. (1996). The aging immune system. *Primer and Prospectus Science, 273*, 70–73.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications, 36*, 2–17.

Proust, J., Rosenzweig, P., Debouzy, C., & Moulias, R. (1985). Lymphopenia induced by acute bacterial infections in the elderly: A sign of age-related immune dysfunction of major prognostic significance. *Gerontology, 31*(3), 178–185.

Questier, F., Put, R., Coomans, D., Walczak, B., & Vander Heyden, Y. (2005). The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems, 76*, 45–54.

Rumelhart, D., McClelland, J., & the PDP Research Group. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Simon, D., & Boring, III J. R. (1990). Sensitivity, specificity, and predictive value. In H. K. Walker, W. D. Hall, & J. W. Hurst (Eds.), *Clinical methods: The history, physical, and laboratory examinations* (pp. 49–54). Boston: Butterworths.

St. John, C. H., Balakrishnan, N., & Fiet, J. O. (2000). Modeling the relationship between corporate strategy and wealth creation using neural networks. *Computers & Operations Research, 27*, 1077–1092.

Tourassi, G. D., & Floyd, C. E. (1997). The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Medical Decision Making, 17*, 186–192.

Trtica-Majnaric, Lj, Zekic-Susac, M., Sarlija, N., & Vitale, B. (2010). Prediction of influenza vaccination outcome by neural networks. *Journal of Biomedical Informatics, 43*, 774–781.

Yeh, C. C., Chi, D. J., & Hsu, M. F. (2010). A hybrid approach of DEA, rough set and support vector machines for business failure prediction. *Expert Systems with Applications, 37*, 1535–1541.