Associate Professor Catalina Liliana ANDREI, PhD
University of Medicine and Pharmacy "Carol Davila", Bucharest
E-mail: ccatalina97@yahoo.com
Professor Bogdan OANCEA, PhD
„Nicolae Titulescu" University of Bucharest
E-mail: bogdanoancea@univnt.ro
Professor Monica NEDELCU, PhD
The Bucharest University of Economic Studies
E-mail: mona.nedelcu@yahoo.com
Associate Professor Ruxandra Diana SINESCU, PhD
University of Medicine and Pharmacy "Carol Davila", Bucharest
E-mail: ruxandrasinescu@gmail.com (corresponding author)

PREDICTING CARDIOVASCULAR DISEASES PREVALENCE USING NEURAL NETWORKS

Abstract.*Healthcare costs recorded an important increase in volume during the past years in almost every country and represents and important burden on each government. The reasons of increased healthcare costs are various: hospital cost, healthcare providers' prices, medical technology costs, unhealthy lifestyles, aging population. One important category of health problems are cardiovascular diseases because of the high mortality risk and the high cost of medical care. In this paper we addressed the problem of predicting the cardiovascular diseases based on general health determinants, using a neural network. For this purpose we used the data collected within the European Health Interview Survey (EHIS) for Romania. The data set consists of 18173 records, 75% being used for training the network and 25% for testing. The neural network was a multilayer perceptron (MLP) with one input layer, one hidden layer and one output layer and it was trained using different algorithms.*

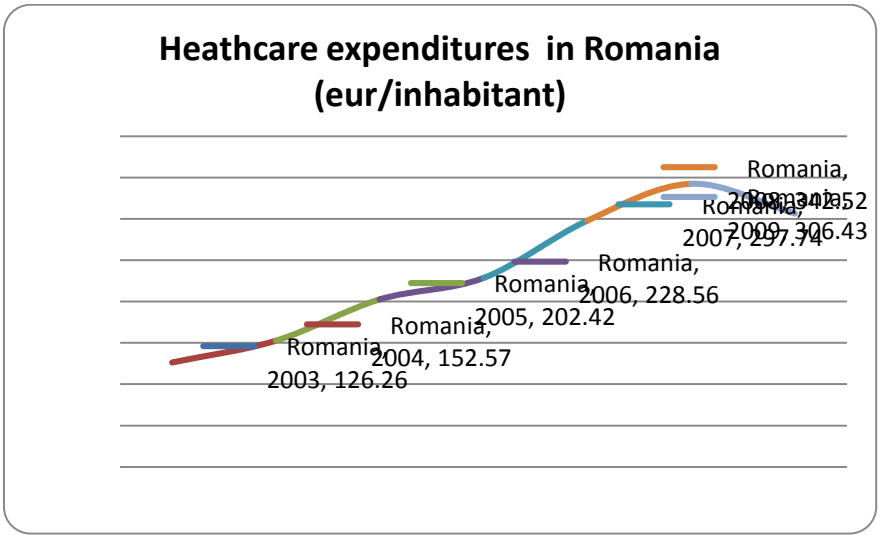
Keywords: healthcare, cardiovascular diseases prevalence, neural networks, prediction.

JEL Classification: I10, C02

1. Introduction

Healthcare cost has recorded an important increase during the past years. The reasons of the increased healthcare costs are various: hospital costs has increased, healthcare providers' cost has also increased, modern medical technology is very expensive, the average age of the population has increased almost all over the world, the lifestyle of the population tends to be unhealthy. The estimation of the costs of healthcare are based on various surveys and administrative (register-based) data sources, as well as estimations made by each state, reflecting country-specific ways of organising the healthcare processes and different reporting and recording systems for the statistics pertaining to healthcare. For example, in Romania the healthcare cost per inhabitant increased from 126 EUR/inhabitant in 2003 to 346 EUR/inhabitant in 2009 as presented in EUROSTAT Yearbook (2013). Figure 1 shows the evolution of the healthcare costs in Romania.

Figure 1. The healthcare expenditures in Romania (EUR/inhabitant)



Currently, almost half of the noncommunicable diseases are represented by cardiovascular diseases that have become the world's major disease burden in the last years (Laslett, 2012). Cardiovascular diseases remain the leading global cause of death: in 2008 a number of 17.3 million deaths were recorded because of these kinds of diseases. This number is expected to grow to over 23.6 million by 2030 (Collins, 2012; Mendis, 2011). Although in Western and Central European

countries the mortality declined over the past 20 years, in Romania there were an increase of the mortality caused by cardiovascular diseases, which was about 10 deaths per 1000 inhabitants in 2008 (Andrei, 2011).

Identifying the determinants of cardiovascular diseases and predicting the health status of the population are important tasks that can help any national healthcare system.

In this paper we used the data from the EHIS survey taken in Romania in 2008 to predict the cardiovascular diseases prevalence using an artificial neural network. Feed forward artificial neural networks (ANN) are applied in many fields like medical diagnosis, financial forecasting, pattern recognition, OCR. ANNs are being used for regression or classification purposes because they are one of the best functional mappers. The good results of applying neural networks in classification problems reported thus far in the literature lead us to use them for predicting the prevalence of cardiovascular diseases.

The paper is organized as follows. The next section reviews the related work in the field of using artificial neural networks for prediction and classification purposes in medicine and healthcare. Next, we describe the data used in our study, the variables used as inputs for the ANN and the structure of the ANN. A short review of the basic principles of ANNs is also given here. In the next section we present the results of using the neural network emphasizing the good capabilities for classification of the ANN. The final section of the paper presents some conclusions of the study and some guidelines for further development.

2.Related work

Artificial neural networks have long been used by many researchers in medicine and healthcare, especially for medical diagnosis systems.

Cancer survival prediction is one of the fields where neural networks were successfully used (Burke, 1997). The authors used an ANN to predict the survival period of cancer diagnosed patients using data from the Commission on Cancer's Breast and Colorectal Carcinoma Patient Care Evaluation in USA and they showed that the ANN has better prediction accuracy then other statistical prediction systems used in cancer diagnosis.

Prediction of metastases in breast cancer patients was successfully implemented by Choong (1994) using an entropy maximization network (EMN). The authors used an EMN to build discrete models that predict the occurrence of axilliary lymph node metastases in breast cancer patients.

Sordo (1994) presents a comparison between the performance of different neural networks and training techniques used in the diagnosis of Down's syndrome in unborn babies. The results obtained by neural networks (84% correct classification rate) were better than other statistical methods used in this field.

In Dangare (2012) the authors presents a data mining approach for predicting heart diseases using a neural network. The system developed by the authors of the paper for prediction uses 13 medical parameters like sex, blood pressure, cholesterol, obesity, lack of physical activity etc. The network used was a multilayer perceptron trained by the classical backpropagation algorithm. The accuracy of the network was impressive: 99.25%.

Vanisree (2011) presents a decision support system for diagnosis of congenital heart disease. The system was implemented using a multilayer feedforward neural network trained by a supervised delta learning rule. The system is capable of predicting heart diseases with an accuracy of 90% using data regarding signs, symptoms and the results of physical evaluation of the patients.

Image processing and pattern recognition is another field where artificial neural networks are successfully applied. Daschlein (1994) describes an implementation of a neural network system for segmentation of the CT images of the abdomen separating the images of various tissues. A review of the application of neural networks in medical image processing is presented in Shi (2010) emphasizing the major strengths and weakness of applying neural networks for medical image processing.

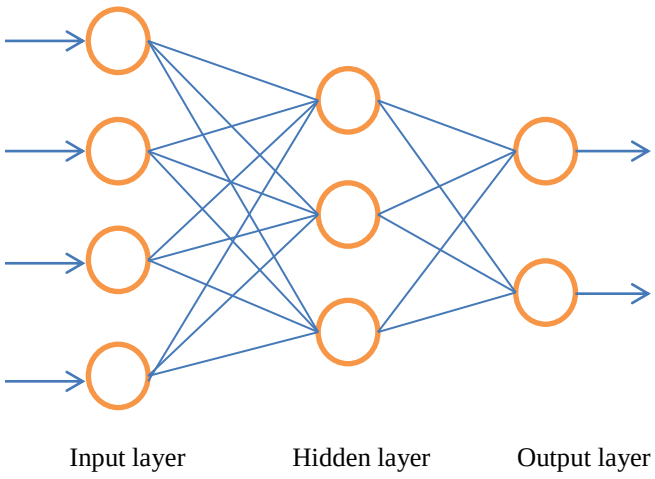
Prediction of life expectancy using neural networks and survival analysis is described in Iliadis (2012). The authors stressed the contribution of the determinants to the health status of the population in the countries of OECD. They used Cox regression and ANNs to analyse the life expectancy at birth based on health spending, pollution, education, income, alcohol consumption, tobacco and diet.

3. Methodology

An artificial neural network is a computational system consisting of processing elements called neurons which are interconnected via synapses. Figure 1 depicts the structure of a simple neural network. It can be seen that neural networks are structured in layers. The network in figure 1 has 3 layers: one input layer, one hidden layer and one output layer. Such a network is often called a multilayer perceptron (MLP). The information flows through each neuron in an input-output way: each neuron receives an input signal, processes it and outputs the transformed signal to the other connected neurons in the next layer. Each

connection between two neurons has an associated weight. The input layer has the only task to receive the external signal and forward it to the next layer. The neurons in the hidden layer receive a weighted sum of signals sent by the input layer and process this sum by means of applying an activation function.

Figure 2. A neural network with 3 layers



Several activation functions are used in practice but the most common the sigmoid function, hyperbolic tangent, softmax function. The signal is than sent to the next layer which can be another hidden layer or an output layer. The neurons in the output layer compute a weighted sum of its inputs and apply an activation function. From this description it can be noted that the information flows forward through the successive layers of the network. That's why this kind of networks is also called feedforward networks. The learning process consists in adapting the weights of the connections in order to optimize the network output to a given input data.

In supervised learning, the network is trained with pairs (input, output) data. For each input data the network produces an output. The accuracy of the output results is measured by an error function defined as the difference between the actual output a_p and the desired one d_p:

$$E = \frac{1}{2}\sum_{j=1}^{N}(a_{pj} - d_{pj})^2 \quad (1)$$

where N is the number of output neurons, and *p* designates the training pair. The network weights are updated to minimize the output error given above. The most used training algorithm is called *backpropagation* and is described in detail in (Rumelhart, 1986).

We used a feed forward neural network for predicting the cardiovascular diseases prevalence using data form European Health Interview Survey (EHIS) for Romania (Rasmussen, 2008). The interview covered 10410 houses from all counties in Romania and it used a two stages survey.

From the 10410 houses selected in the sample, data were collected from 9963 houses, the rest being seasonal houses or no longer exist at the date of the interview. The response rate was 97.2%: data were collected from 18173 adults and 2616 children. In our experiments we used only the data for adult persons. In Andrei (2010) it is shown that socio-economic and lifestyle factors are important determinants of the cardiovascular diseases.

We used the following variables from the EHIS database as input data:

- Age
- Sex
- Degree of urbanization
- Education level
- Labour status
- Body mass index
- The number of days per week with vigorous physical activity
- The number of days per week with moderate physical activity
- The number of days per week on which the patient walked at least 10 minutes
- Fruits consumption
- Vegetables consumption
- Fruit or vegetable juice consumption
- Exposure to noise

- Air pollution
- Exposure to bed smells from industry, agriculture, sewer, waste
- Smoking (frequency)
- Exposure to tobacco smoke indoors at home or at workplace
- Frequency of alcohol consumption
- Drugs (cannabis, cocaine, amphetamines, ecstasy or other similar substances) consumption

For the output of the network we used the existence of a longstanding cardiovascular disease (myocardial infarction, coronary heart disease, high blood pressure, cerebral haemorrhage, cerebral thrombosis as they are encoded in HS04A-HS04U variables) diagnosed by a medical doctor. This variable was computed using HS02, HS04A-HS04U, HS05A-HS05U from the EHIS data and has two values: 1 – the person has a cardiovascular disease, 0 otherwise.

We build a neural network with one input layer, one hidden layer and one output layer. Our task was to predict the existence of a longstanding cardiovascular disease given the input data. The first layer of the neural network comprises 111 neurons, 110 corresponding to different values of the input variables and one neuron used as a bias signal. We conducted a series of experiments in order to establish the optimal number of hidden layers and the optimal number of neurons in each layer.

Our experiments give us that the best results are obtained with one hidden layer with 166 neurons. The hidden layer also receives a bias signal of 1. The output layer has two neurons, one for each of the two classes. Every neuron in a layer is fully connected to every neuron in the next layer.

Each neuron in the hidden layer accumulates the input from the neurons in the preceding layer, computes the weighted sum of these inputs, adds the bias and calculates the output signal according to the following expression:

$$y_i = f\left(\sum w_{ij} x_j + b\right) \tag{2}$$

where b is the bias input which in our case has the value of 1, and $f$is the activation function of the neuron. ANN literature reports many activation functions (Hassoun, 1995): hyperbolic tangent, sigmoid, linear, step, logistic, softmax etc. We used the *hyperbolic tangent* as an activation function for the neurons in the hidden layer and the *softmax* function for the neurons in the output layer. *Softmax* function is widely used in classification or clustering systems because it converts a raw value into a

posterior probability that provides a measure of certainty of classification. The network error function used in our experiments was the mean square error (MSE). MSE error function applied on the validation data set can be used as an estimate of the variance and can be also used to compute the confidence interval of the network output assuming a normal distribution of the output values.

We experimented several supervised training algorithms: classical backpropagation, scaled conjugate gradient, quick propagation, resilient propagation (RPROP+) and some versions of it: RPROP-, iRPROP+, iRPROP-. The best convergence rate was obtained using a version of the resilient backpropagation called iRPROP+. This result is consistent with other works presented in Oancea, (2013). iRPROP+ is described in detail in Igel, (2000) and is one of the best performing first-order supervised learning methods for feed forward neural networks.

The classical resilient propagation supervised training algorithm controls the weight update of each connection maximizing the update step size and minimizing the oscillations. The sign of the partial derivative of the error function of the network $\frac{\partial E}{\partial w_{ij}}$ is being used to compute the direction of the weight update. Each weight is updated with a different value which is independent of the absolute value of the partial derivative. If the partial derivative $\frac{\partial E}{\partial w_{ij}}$ has the same sign for consecutive steps, the step size of the update is increased, otherwise the step size is decreased. The change in sign of the partial derivative $\frac{\partial E}{\partial w_{ij}}$ gives the weight updates direction: if the sign is unchanged the weight update is done normally but if the sign has changed the previous weight update is reverted (Riedmiller, 1993).

The iRPROP+ training algorithm reverts only weight updates that have caused sign changes of the partial derivative and an error increase. This rule combines information about the sign of the error function which is error surface information with the magnitude of the network error when the decision of reverting an update step is taken.

4. Results

Our neural network was implemented in Java using the Encog framework (Heaton, 2011). Also it is a common believe that Java is still behind other languages like C or C++ regarding the performances, recent works (Oancea, 2011) showed that Java has good computing capabilities even for numerically intensive applications. We used 18317 records from the EHIS database: 75% of them were used for training the network while 25% of them were used for testing purposes.

A difficult problem was to choose the right number of neurons and layers because choosing a small number of neurons and layers will lower the mapping power of the network and, on the other hand, choosing a large number of hidden layers and neurons could give the network the power to fit very complex data but it will slow down the training process.

While there are some methods of choosing the best/optimal network structure (Ploj, 2014), we proceeded with a trial and error process. The number of neurons in the input layer and the output layer was dictated by the size of input and output data. We had to choose the optimum number of hidden layers and the number of neurons in the hidden layer. We started with a network with one hidden layer and a small number of neurons and gradually increase its size. We found that the best results regarding the convergence rate are obtained for a network with the following structure: 111I-166H-2O, i.e. 111 input neurons, a hidden layer with 166 neurons and an output layer with 2 neurons. Adding more neurons to the hidden layer or a second hidden layer decreased the convergence rate. These results are in line with other theoretical findings (Cybenko, 1989).

We trained the network with different training algorithms imposing as a stop criterion a value of 0.1% for MSE. The resilient propagation class of training algorithms outperformed the other algorithms considered for training the network. Table 1 gives the number of training epochs needed to achieve the desired error for 4 training algorithms belonging to the resilient backpropagation class available in Encog framework.

Table 1. The number of training epochs needed to achieve an error of 0.1%

Training algorithm	Number of training epochs
RPROP classic	22133
RPROP-	16234
iRPROP+	9605
iRPROP-	10343

In the EHIS data set, 18.5% from the persons who participated at the interview had a cardiovascular disease. Our network predicted the existence of a cardiovascular disease with an accuracy of 96.2% (computed as the ratio between the correct predicted values and total number of records in the test data set).

5. Conclusions

Prediction is an important tool and represents a first step in every healthcare system. Cardiovascular diseases are one of the world's major disease burden. In

Romania the annual rate of mortality from heart failure is approximately 60% (Andrei, 2010). This high mortality rate and the high costs of treating patients with cardiovascular diseases make a prediction tool a must in every healthcare management system. We used the classification power of an artificial neural network to predict the occurrence of a cardiovascular disease based on certain input data regarding the level of education, labour status, body mass index, diet, and lifestyle. We achieved an accuracy rate of over 96% showing that the ANN can be successfully used in prediction of cardiovascular diseases.

Our network was implemented using the Encog framework and it was a multilayer perceptron with one input layer, one hidden layer and one output layer. The activation function of the hidden layer was *tanh* and those of the output layer was the *softmax* function. We tested several supervised training algorithms and we found out that the best results were obtained using the iRPROP+ algorithm.

This average predictability rate is comparable with other results presented in the beginning of the paper. A direction for a future research would be to identify those variables from the input data that influence in a greater extent the occurrence of a cardiovascular disease. Knowing the most important health determinants could be helpful for healthcare policy makers to design proper policies in order to reduce the risks of cardiovascular diseases, reducing thus the overall healthcare costs.

REFERENCES

[1]**Andrei, C. L., Sinescu, C. J., Oancea, B., and Iacob, A. I. (2010), Methods *Used by Students for Development of Socio-economics Barriers in Heart Failure Management***; *Procedia Social and Behavioural Sciences* 9, pp. 1272-1276;
[2]**Andrei, C. L., Oancea, B., Iacob, A. I., and Sinescu, C. J. (2012), *Quantitative Techniques Used for Analyzing the Geographical Particularities of Patients with Acute Coronary Syndromes*;** *Procedia Social and Behavioural Sciences* 31, pp. 198-201;
[3]**Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., Marks, J. R., Winchester, D. P. and Bostwick, D. G. (1997), *Artificial Neural Networks Improve the Accuracy if Cancer Survival Prediction***; *CANCER*, 79(4), pp. 857-862;
[4]**Choong, P. L., deSilva, C. J. S. (1994), *Breast Cancer Prognosis Using EMN Architecture***; *Proceedings of IEEE International Conference on Neural Networks*, Orlando, Florida, USA;
[5]**Cybenko G. (1989), *Approximations by Superpositions of Sigmoidal Functions***; *Mathematics of Control, Signals and Systems* 2(4), pp. 303-314;

[6]**Danfgare, C. S. and Apte, S.S. (2012),** ***A Data Mining Approach for Prediction of Heart Disease Using Neural Networks***; *International Journal of Computer Engineering &Technology*, 3(3) pp. 30-40;
[7]**Daschlein R., Waschulzik, T. and Brauer, W. (1994),** ***Segmentation of Computer Tomographies with Neural Network Based on Local Features*****.** In: Ifeachor E., Rosen K. eds., International Conference on Neural Networks and Expert Systems in Medicine and Healthcare, pp. 174-180;
[8]**Eurostat, (2013),** ***Europe in figure*** – Eurostat Yearbook;
[9]**Hassoun, M. H. (1995),** ***Fundamentals of Artificial Neural Networks***, MIT Press;
[10]**Heaton, J. (2011),*****Programming Neural Networks with Encog3 in Java***, 2nd Edition, Heaton Research;
[11]**Igel, C. and Hüsken, M. (2000),** ***Improving the Rprop Learning Algorithm***; *Proceedings of the Second International Symposium on Neural Computation*, pp. 115–121, ICSC Academic Press;
[12]**Iliadis, L., Kitidou, K., Skoularik, K. (2012),** ***Combining Survival Analysis and Neural Networks to Predict Life Expectancy***; *International Journal of Artificial Intelligence*, 9(A12), pp. 140-151;
[13]**Laslett, L.J., Alagona, P., Clark,B.A., Drozda, J. P., Saldivar, F., Wilson, S. R., Poe, C., Hart, M. (2012),** ***The Worldwide Environment of Cardiovascular Disease: Prevalence, Diagnosis, Therapy, and Policy Issues: A Report From the American College of Cardiology***; *Journal of American College of Cardiology*, 60(25_S):S1-S49;
[14]**Mendis, S., Puska, P., and Norrving, B. eds., (2011),** ***Cardiovascular Disease: Global Atlas on Cardiovascular Disease Prevention and Control***; World Health Organization;
[15]**Oancea, B., Ciucu, S. (2013),** ***Time Series Forecasting Using Neural Networks*****;** *Proceedings of the "Challenges of the Knowledge Society" Conference*, pp. 1402-1408;
[16]**Oancea, B., Rosca, I. G., Andrei, T., Iacob, A. I. (2011),** ***Evaluating Java Performance for Linear Algebra Numerical Computations***; *Procedia Computer Science*, vol 3, pp. 474-478;
[17]**Ploj, B., Harb, R. and Milan Zorman, M. (2014),** ***Border Pairs Method – Constructive MLP Learning Classification Algorithm.*** *Neurocomputing*, 126, pp. 180-187;
[18]**Rasmussen, N., Erikson, J., Grigore, D. C., Branzio, C. and Ciocanel, B.(2008),** ***Starea de sanatate a populatiei din Romania***;
[19]**Riedmiller, M., H. Braun, H. (1993),** ***A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP algorithm*****.** In: *Proceedings of the IEEE International Conference on Neural Networks*, pp. 586–591, IEEE Press;

[20]**Rumelhart, D., McClelland J. (1986),** ***Parallel Distributed Processing. Explorations in the Microstructure of Cognition; Vol. 1: Foundations***. MIT Press;
[21]**Shi, Z., He, L. (2010),** ***Application of Neural Networks in Medical Image Processing***; *Proceedings of the Second International Symposium on Networking and Network Security*, pp. 23-26;
[22]**Smith S.C., Collins A. and Ferrari R. (2012),** ***Our Time: A Call to Save Preventable Death from Cardiovascular Disease (heart disease and stroke).*** *Journal of American College of Cardiology*, 60(23), pp. 2343-2348;
[23]**Sordo M. (1995),** ***Neural Networks for Detection of Down's Syndrome.*** Master's thesis, Department of Artificial Intelligence, University of Edinburgh;
[24]**Vanisree K. and Singaraju, J. (2011),** ***Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks***, *International Journal of Computer Applications*, 19(6), pp.8875-8887.