Elsevier Editorial System(tm) for Biomedical Signal Processing and Control
Manuscript Draft

Manuscript Number: BSPC-D-14-00407

Title: Ensemble back-propagation neural networks for predicting the risk of hip bone fracture for elders in Taiwan

Article Type: Research Paper

Keywords: Ensemble artificial neural networks; back-propagation neural networks; sensitivity analysis; connection weights; hip fracture.

Corresponding Author: Prof. Jiann-Shing Shieh,

Corresponding Author's Institution:

First Author: Quan Liu

Order of Authors: Quan Liu; Xingran Cui; Yuan-Chao Chou; Maysam F Abbod; Jinn Lin; Jiann-Shing Shieh

Abstract: Hip bone fracture is one of the most important causes of morbidity and mortality in the elder adults. It is necessary to establish a prediction model to provide suggestions for elders. A total of 725 subjects were involved, including 228 patients with first low-trauma hip fracture and 497 ages-, sex-, and living area-matched controls (215 from the same hospital and 282 from community). All the subjects were interviewed with the same questionnaire, and the answers of the interviewees were recorded to be the database. Three-layer back-propagation Artificial Neural Networks (ANN) models were applied for females and males separately in this study to predict the risk of hip bone fracture for elders. Furthermore, to improve the accuracies and the generalizations of the models, the ensemble ANNs method was applied. To understand variables contributions and find the important variables for predicting hip fracture, sensitivity analysis and connection weights approach were applied. In this study, three ANNs prediction models were tested with different architectures. With the fivefold cross-validation method evaluating the performances, one of the three models turned out to be the best prediction model and achieved a big success of prediction. The best area under the receiver operating characteristic (ROC) curve and the accuracy of the prediction model are 0.91 ± 0.028 (mean ± SD) and 0.85 ± 0.029 for females, while for males are 0.99 ± 0.015 and 0.93 ± 0.020. With the method of sensitivity analysis and connection weights, input variables were ranked according to contributions/importance, and the top 10 variables show great proportion of contribution to predict hip fracture. The top 10 important variables causing hip fracture for both females and males are similar to our previous results got from logistic regression model and other related researches. In conclusion, ANNs has successfully been to establish prediction models for predicting the risk of hip bone fracture for both female and male elder adults respectively and identified the top 10 important variables from 74 input variables to predict hip bone fracture of elders. This study verified the performance of ANNs to be a highly complex prediction model.

Cover Letter

23 Sept 2014

Dear Prof. R. Allen,

I use the electronic version to send this manuscript to you. The manuscript title is: “Ensemble back-propagation neural networks for predicting the risk of hip bone fracture for elders in Taiwan”. We are submitting this material for possible publication in “Biomedical Signal Processing and Control”. This material has not been submitted for publication or published elsewhere in whole or part. We believe this manuscript represents an original and significant contribution to the field of “Neural networks for predicting the risk of hip bone fracture” and therefore would like to be considered for publication in “Original Articles”.

Sincerely yours,

Prof. Jiann-Shing Shieh, PhD
Head of Dept. of Mechanical Engineering
Director of Institute of Mechanical Engineering
Joint Professor of Graduate School of Biotechnology and Bioengineering
Yuan Ze University
Tel: 886-3-4638800 ext. 2470
Fax: 886-3-4558013
Email: jsshieh@saturn.yzu.edu.tw

Ensemble back-propagation neural networks for predicting the risk of hip bone fracture for elders in Taiwan

Quan Liu[a,†], Xingran Cui[a,b,†] , Yuan-Chao Chou[c], , Maysam F. Abbod[d], Jinn Lin[e], and Jiann-Shing Shieh[c,f,g,*]

[a] School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei, China

[b] Department of Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, USA

[c] Department of Mechanical Engineering, Yuan Ze University, Chung-Li, Taiwan

[d] School of Engineering and Design, Brunel University, London, UB8 3PH, United Kingdom

[e] Department of Orthopaedic Surgery, National Taiwan University Hospital, Taipei, Taiwan

[f] Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Chung-Li, Taiwan

[g] Center for Dynamical Biomarkers and Translational Medicine, National Central University, Chung-Li, Taiwan

†These authors contributed equally to the work.

*Corresponding author: Professor Jiann-Shing Shieh, PhD, Department of Mechanical Engineering, and Innovation Center for Big Data and Digital Convergence, Yuan Ze University, 135, Yuan-Tung Road, Chung-Li 32003, Taiwan. Tel: 886-3-4638800 ext. 2470. Email: jsshieh@saturn.yzu.edu.tw

Abstract

Hip bone fracture is one of the most important causes of morbidity and mortality in the elder adults. It is necessary to establish a prediction model to provide suggestions for elders. A total of 725 subjects were involved, including 228 patients with first low-trauma hip fracture and 497 ages-, sex-, and living area-matched controls (215 from the same hospital and 282 from community). All the subjects were interviewed with the same questionnaire, and the answers of the interviewees were recorded to be the database. Three-layer back-propagation Artificial Neural Networks (ANN) models were applied for females and males separately in this study to predict the risk of hip bone fracture for elders. Furthermore, to improve the accuracies and the generalizations of the models, the ensemble ANNs method was applied. To understand variables contributions and find the important variables for predicting hip fracture, sensitivity analysis and connection weights approach were applied. In this study, three ANNs prediction models were tested with different architectures. With the fivefold cross-validation method evaluating the performances, one of the three models turned out to be the best prediction model and achieved a big success of prediction. The best area under the receiver operating characteristic (ROC) curve and the accuracy of the prediction model are 0.91 ± 0.028 (mean ± SD) and 0.85 ± 0.029 for females, while for males are 0.99 ± 0.015 and 0.93 ± 0.020. With the method of sensitivity analysis and connection weights, input variables were ranked according to contributions/importance, and the top 10 variables show great proportion of contribution to predict hip fracture. The top 10 important variables causing hip fracture for both females and males are similar to our previous results got from logistic regression model and other related researches. In conclusion, ANNs has successfully been to establish prediction models for predicting the risk of hip bone fracture for both female and male elder adults respectively and identified the top 10 important variables from 74 input variables to predict hip bone fracture of elders. This study verified the performance of ANNs to be a highly complex prediction model.

Keywords: Ensemble artificial neural networks; back-propagation neural networks; sensitivity analysis; connection weights; hip fracture.

1. Introduction

Hip fracture is a kind of serious injury for elders. In previous studies, they have found out that elder adults with hip bone fracture have a relatively higher risk of death (Johnell et al., 2004; Magaziner et al., 1989). The post-fracture one-year mortality rates for the elders with hip fracture are 18-33% (Magaziner et al., 2003). Even if the patients survive after the fracture, some of them still suffer functional loss in daily activities (Jette et al., 1987). Moreover, the elders with hip fracture and their family need to shoulder much higher health care costs compared with their matched controls (Haentjens et al., 2005). Therefore, hip fracture is not only a considerable health burden but also an increasing economic burden.

To reduce the incidence of this preventable injury and subsequent adverse outcomes, many studies have identified the risk factors for hip fracture (Benetos et al., 2007; Dubey et al., 1999; Wehren et al., 2003). Recently, a matched case-control study carried out a conditional logistic regression to find out the important risk factors with the combined effects of different risk factors (Lan et al., 2010). Another suitable method to analysis biomedical systems is artificial neural network (ANN). According to the advantages of nonlinearity, fault tolerance, universality, and real-time operation, ANNs have been proposed as a quite suitable algorithm for modeling complex non-linear relationships in health care research (Baxt et al., 1995; Cross et al., 1995; Kung and Hwang, 1998). Eller-Vainicher et al. (2011) identified the promising role of ANN in predicting osteoporotic fracture among postmenopause osteoporosis women. For the comparison of the characteristics between ANNs and logistic regression applied to this epidemiological research field, a study has established prediction models for predicting living setting after hip fracture by ANNs and logistic regression, and shown that ANN is slightly better than logistic regression (Ottenbacher et al., 2004). Lin et al. found ANN algorithm could reliably predict the mortality of hip fractured patients and outperforms the logistic regression method (Lin et al., 2010). Although in many studies ANNs have been shown to exhibit superior predictive power compared to traditional approaches (Cui et al., 2012; Liu et al., 2011), they have also been labeled a ‘‘black box’’ because they provide little explanatory insight into the relative influence of the independent variables in the prediction process. This lack of explanatory power is a major concern to ecologists since the interpretation of statistical models is desirable for gaining knowledge of the causal relationships driving ecological phenomena. Besides, the significant ranking of each input is very important for the neural network operation. To “illuminate” the “black box”, Olden et al. (2004) introduced nine methods for quantifying variable importance in artificial neural networks, of which, sensitivity analysis is a generally used method. The sensitivity analysis methodology is able to show the specific contribution of the input variables while ANN has the capability to handle non-linear, complex ecological data and to incorporate causality (Lek and Guegan, 2000; Rechnagel, 2003). Hence, the present study had two primary goals. The first goal was to establish ANNs prediction models to predict the risk of hip fracture for female and male elder adults respectively, and examine them via the ROC curve analysis. With this ANNs models, the second goal of this paper was to use the methods of sensitivity analysis and connection weights to understand the contribution of each input variable and identify the top 10 important variables for predicting hip fracture. These top 10 important variables were also compared with the most influential variables got from conventional logistic regression method (Lan et al., 2010; Tseng et al., 2013).

2. Materials and methods

2.1 Database

The database utilized in this study were collected in the previous case-control matched study for the analysis of risk factors of hip fracture for elder adults aged 60 and older (Lan et al., 2010). The data were collected from the questionnaire surveys interviewed by trained interviewers. The database included a total of 725 subjects, of which, 228 subjects were the patients admitted to the National Taiwan University Hospital with first low-trauma hip fracture, 215 subjects were hospital controls (patients in the same hospital but without hip fracture) and 282 subjects were community controls (randomly selected dwellers) individually matched to the hip fracture patients by age, gender, and living area, and then two control groups were combined together as 497 controls. Since women may have some risk factors different from men, such as reproductive history, etc, female and male models were developed seperatly, so the data were separated by gender. Of the total 725 subjects, 163 hip fracture patients and 345 controls were women, and 65 hip fracture patients and 152 controls were men. Moreover, in Lan's study (2010), they used intraclass correlation coefficient to examine the reliability of the sample data. As a result, the moderate to high agreement suggested that the data was reliable.

2.2 Architecture of ensemble ANNs

There are many types of ANNs with different structures. Typical back-propagation neural networks (BPNN) are commonly adopted for solving classification problems. BPNN include an input layer, a hidden layer (or several hidden layers), and an output layer. Each layer contains at least 1 node (neuron). Activation functions (transfer functions) only exist in the hidden nodes and output nodes, and only the inputs for hidden nodes and output nodes will be processed via weights and biases. Typically, the data for ANNs analysis consist of possible inputs and the corresponding targets, and are divided into three parts: training datasets for training models, validation datasets for checking the over-fitting of models, and testing datasets for testing the generalization of models. To avoid over-fitting is very important to make sure the generalization of an ANN model. Another method for improving the generalization of an ANN model is ensemble ANNs method (Hansen and Salamon, 1990; Zhou et al., 2002). The learning effect of an ANN is decided by random initial weights. The training process of ANNs is an optimization processing of the connections (i.e., weights and biases) between neurons in different layers. Therefore, different initial points (i.e., initial weights and biases) will lead to different optimization results. The idea of the ensemble method is to train a finite number of component neural networks and then combine the component outputs to reduce the errors come from different initial weights and biases (Pulido et al., 2013; Pulido et al., 2014).

Hence, the BPNN and ensemble method were applied in this study. Lan et al. (2010) and Tseng et al. (2013) did a widely analysis of risk factors for hip fracture in elder adults. For comparing to Lan et al. (2010) and Tseng et al. (2013), a total of 74 risk factors of hip bone fracture were selected to be the inputs for the female BPNN models. The 74 risk factors are listed in Table 1. Excluding 7 factors related to female's reproductive history, 67 of above factors were chosen to be the inputs for male BPNN models. Both female and male models have 1 hidden layer and 1 binary output (hip-fractured defined as 1versus non-hip-fractured as 0). The suitable number of hidden neurons (N_{HN}) depends on many conditions, like the number of inputs and outputs (N_{INP} and N_{OUT}), the amount of noise in the targets, the complexity of the function, regularization, etc. (Sarle, 2002). It is usually necessary to train several networks and estimate the generalization error of each network to find out the suitable hidden layer size. In general, it is essential to employ lots of hidden neurons to avoid overfitting with the use of early stopping method, or it will likely be underfitting (Sarle, 1995). Hence N_{HN} is set to be about twice the N_{INP} (i.e., 140 nodes in hidden layer) at first, and got good generalized prediction results. N_{HN} = 140 was set for both female prediction models and male prediction models, since a proportionally small change in N_{HN} will not lead to an obvious effect on the performance. However, some literatures proposed rules to relate N_{HN} to N_{INP}, N_{OUT}, or number of training patterns (Basheer and Hajmeer, 2000). A rule of defining N_{HN} = ($N_{INP} + N_{OUT}$)/2 was selected to be compared with the result of setting N_{HN} = 140 for checking whether the result can be better. The inputs and outputs were normalized into a range between -1 and 1. The activation function of hidden neurons is a tangent sigmoid transfer function, which is defined by:

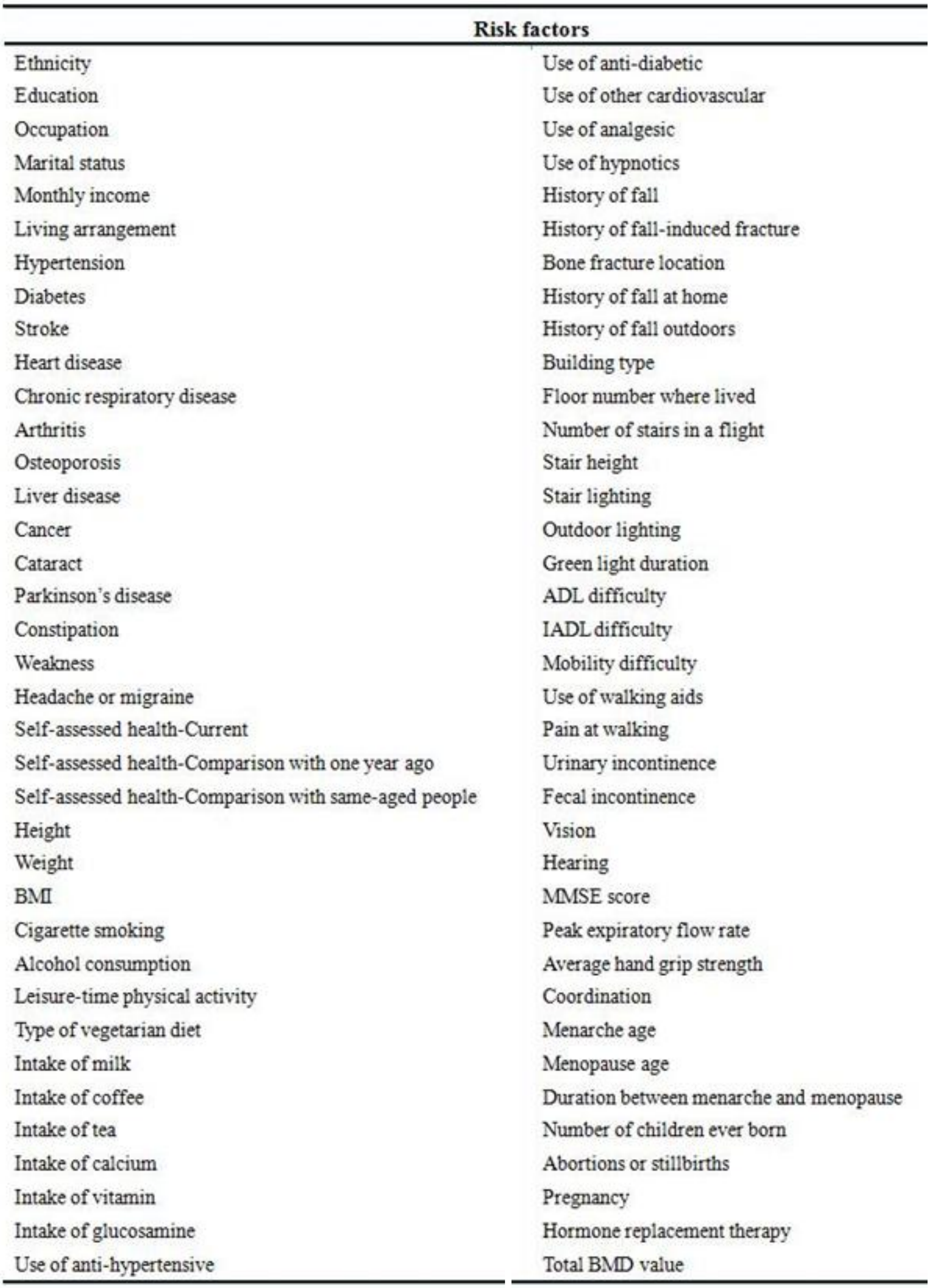
$$f_a = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

where f_a is the output of tangent sigmoid function and x is the input of tangent sigmoid function.

The output of a tangent sigmoid function is limited to a range between -1 to 1 for the normalized output -1 and 1 (0 is normalized to -1 and 1 is 1). The activation function of output neurons is a simple linear transfer function.

In this study, many risk factors were chosen to be the inputs for the prediction models. With large number of N_{HN}, the weights and biases in the networks become huge amount of data. Therefore, for reducing the required memory to do the computations and obtaining better efficiency, conjugate-gradient algorithms (Sarle, 2002) are suitable for the case in this study, and the scaled conjugate-gradient algorithm (Moller, 1993) was chosen to train the models in this study.

Table 1. The risk factors (input variables) for hip fracture analyzed in this study



Risk factors	
Ethnicity	Use of anti-diabetic
Education	Use of other cardiovascular
Occupation	Use of analgesic
Marital status	Use of hypnotics
Monthly income	History of fall
Living arrangement	History of fall-induced fracture
Hypertension	Bone fracture location
Diabetes	History of fall at home
Stroke	History of fall outdoors
Heart disease	Building type
Chronic respiratory disease	Floor number where lived
Arthritis	Number of stairs in a flight
Osteoporosis	Stair height
Liver disease	Stair lighting
Cancer	Outdoor lighting
Cataract	Green light duration
Parkinson's disease	ADL difficulty
Constipation	IADL difficulty
Weakness	Mobility difficulty
Headache or migraine	Use of walking aids
Self-assessed health-Current	Pain at walking
Self-assessed health-Comparison with one year ago	Urinary incontinence
Self-assessed health-Comparison with same-aged people	Fecal incontinence
Height	Vision
Weight	Hearing
BMI	MMSE score
Cigarette smoking	Peak expiratory flow rate
Alcohol consumption	Average hand grip strength
Leisure-time physical activity	Coordination
Type of vegetarian diet	Menarche age
Intake of milk	Menopause age
Intake of coffee	Duration between menarche and menopause
Intake of tea	Number of children ever born
Intake of calcium	Abortions or stillbirths
Intake of vitamin	Pregnancy
Intake of glucosamine	Hormone replacement therapy
Use of anti-hypertensive	Total BMD value

To apply ensemble method, the process for establishing an ensemble ANN model in this study is explained below and illustrated in Fig. 1.

(1) Firstly, the whole database was divided into two parts: 90% and 10%. The 10% part was set to be the testing data for testing the generalization effect.

(2) Training data were randomly chosen 80% of whole database from the 90% part (not 80% of the 90% part), and the remaining 10% were the validation data to supervise over-fitting. This step was repeated 20 times such that 20 training data and validation data sets were generated with different combinations and sequences.

(3) Each training dataset and validation dataset was used to train 15 networks with different initial weights.

(4) The learning effect of each network was tested by the testing data to examine the generalization of the network, and the best networks in each training data and validation dataset were selected to be combined into the ensemble model.

(5) Finally, an ensemble ANN model constructed by 20 networks was established. The output of the ensemble ANN model was the average of 20 best networks.

In this structure, the best networks were selected to be a member of the ensemble for the least generalization error. However, some of the best networks might be selected just by chance, and not as the global optima, i.e., it just especially matched some certain cases. Hence, the median networks were also tested for each training dataset and validation dataset to be the members of the ensemble model. Other processes were the same with the method of choosing best networks as shown in Fig. 1.

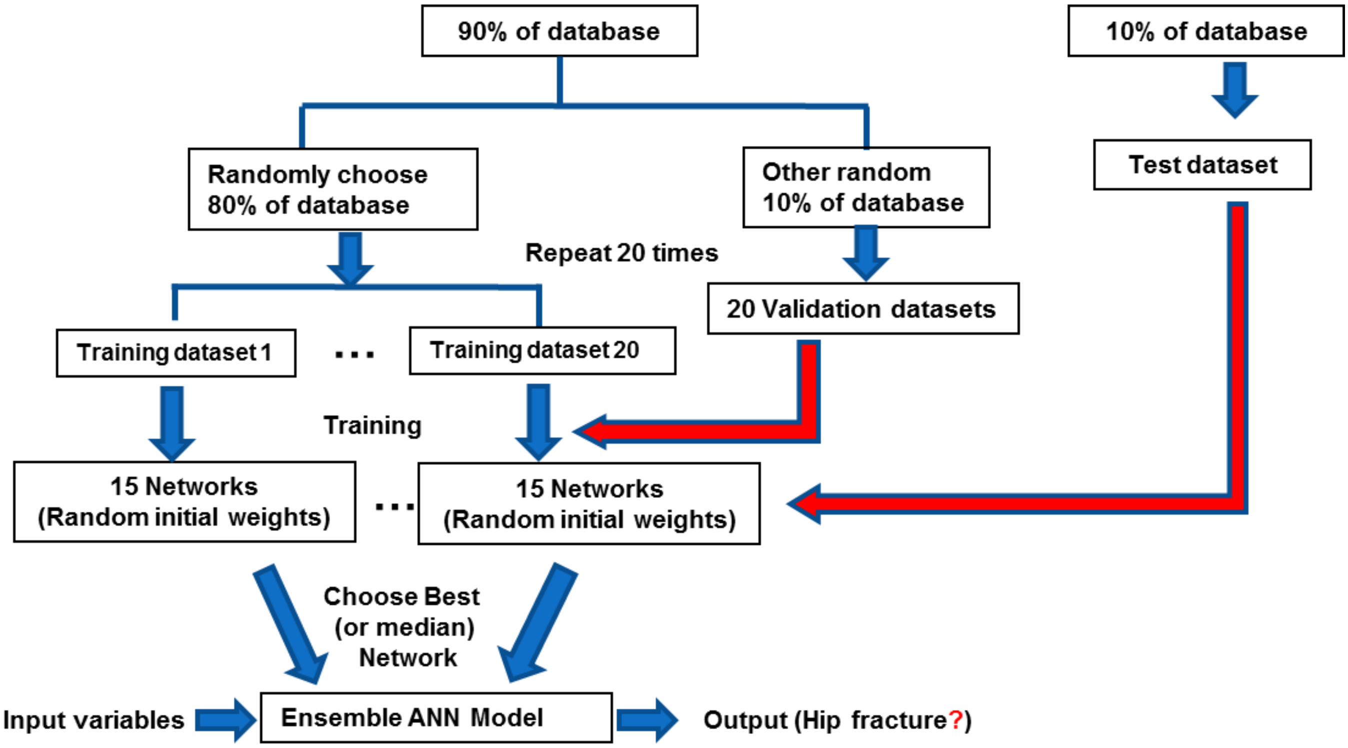


Fig. 1. The flow chart for establishing an ensemble ANN model

2.3 Architecture of the prediction models

The aim of this study is to establish prediction models to predict the risk of hip fracture for elder adults. In Lan's study (2010), the risk factors of hip fracture for female and male elder adults were analyzed respectively, because males do not have the risk factors of reproductive history. For comparison, different ensemble BPNN models were established with different number of inputs for female and male elder adults.

As a result of above considerations in section *2.2*, it has been decided to establish three types of ensemble BPNN models with different structures and compare their performances.

For the females, the three architectures were: (1) 74 inputs / 140 hidden nodes / 1 output and choosing the best networks in the ensemble method, (2) 74 inputs / 37 hidden nodes / 1 output and choosing the best networks in the ensemble method, and (3) 74 inputs / 140 hidden nodes / 1 output and choosing the median networks in the ensemble method.

For the males, the three architectures were: (1) 67 inputs / 140 hidden nodes / 1 output and choosing the best networks in the ensemble method, (2) 67 inputs / 34 hidden nodes / 1 output and choosing the best networks in the ensemble method, and (3) 67 inputs / 140 hidden nodes / 1 output and choosing the median networks in the ensemble method.

The structure of $N_{HN} = (N_{INP} + N_{OUT})/2$ with median networks was not selected since the models with the first structure was selected and examined with good results. If the second and third structures are not better than the first one, the fourth structure type does not need to be considered.

2.4 Coding variables

The raw data in this study are originally from the questionnaire results in Lan's study (2010). Some of the 74 risk factors (the inputs for the BPNN models) in this study are associated with many questions in the original questionnaire. For example, the risk factor of ethnicity is associated with two questions: (1) the ethnicity of father and (2) the ethnicity of mother. However, in the designed model, there is only one input node for the risk factor of ethnicity. To apply the data of the questionnaire surveys into these ensemble BPNN models, the raw multiple data associated with one risk factor are required to be related to one value for each input node. An idea for transferring multiple data into one value came from genetic algorithms. The method of coding multiple data into one value is explained below and illustrated with an example in Fig. 2.

(1) The raw decimal data were transferred to binary values.

(2) The multiple binary data were "combined" together into a binary value.

(3) The combined binary value was transferred into a decimal value.

After the above processing, the multiple questionnaire data can be coded into one value and applied to training, validating or testing ensemble BPNN models.

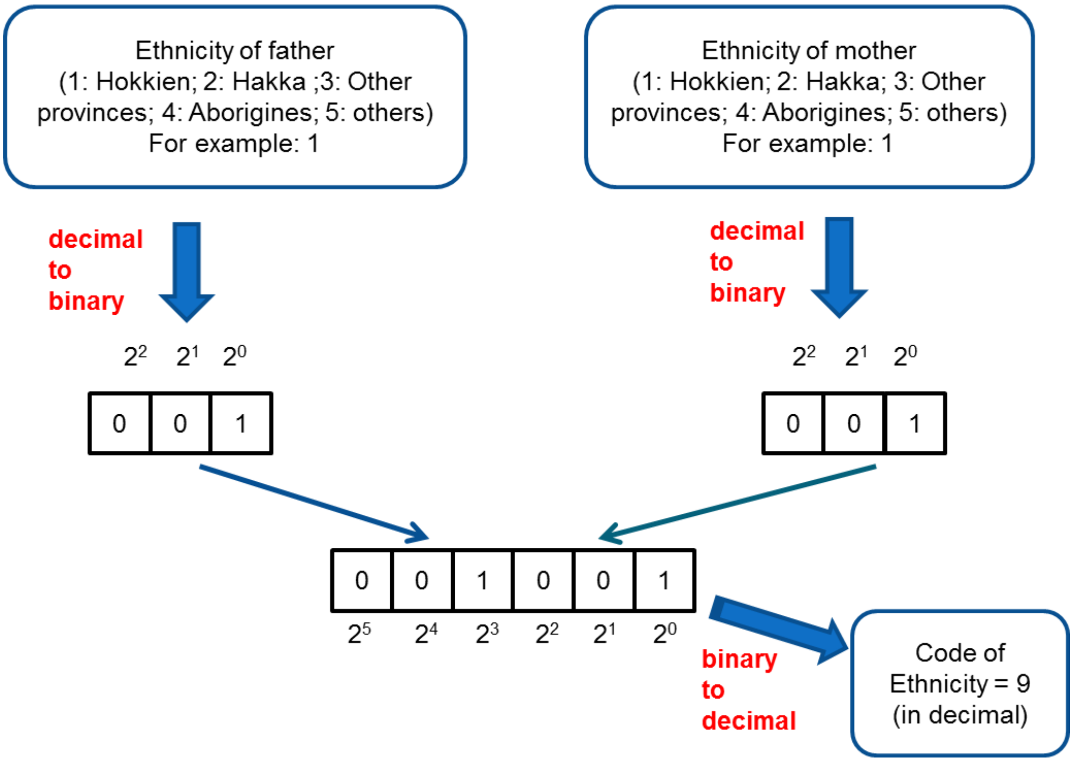


Fig. 2. An example for coding the variable of ethnicity, the raw data are the ethnicities of father and mother (e.g., both are 1) and coded to a value of 9 for applying into the ensemble BPNN models.

2.5 Evaluation of the predictive performance

The ultimate goal of the prediction model is to find a set of neural network weights and bias values so that the input data generates output values that best match the target values. However, weights and bias values may match the data extremely well, but when presented with a new, previously unseen set of input data, the neural network would likely predict very poorly. This phenomenon is called over-fitting. To avoid over-fitting, the process of cross-validation is used to estimate the quality of the ANNs prediction model (Lutz Prechelt, 1998).

The idea behind k-fold cross-validation is to divide all the available data items into roughly equal-sized sets. Each set is used exactly once as the test set while the remaining data is used as the training set (Rodríguez et al., 2010). Based on the fivefold cross-validation method to examine the prediction models produced by the above algorithms, the whole database was randomly divided into ten distinct parts because of the amount of test data is 10% of the whole database (Lin et al., 2006). One part was used as testing data, and the remaining nine parts was used in the ensemble method as the 90% part. This procedure was repeated five times such that five datasets (dataset1, dataset2… dataset5) with different testing data can be generated.

The receiver operating characteristic (ROC) curve analysis has been widely applied as a useful tool to evaluate performances of classifiers (Lin et al., 2006; Ottenbacher et al., 2001; Ottenbacher et al., 2004; Yeh et al., 2009). In this study, the ROC curve analysis was also applied to estimate the discrimination power of the prediction models. The area under the ROC curve (AUC) is the main performance index of a classification. An AUC of 1.0 implies perfect discrimination, whereas an AUC of 0.5 is equivalent to a random model. The sensitivity (SEN), specificity (SPE), positive prediction value (PPV), and negative prediction value (NPV) was also calculated for each model at the best threshold to classify true or false. Furthermore, the accuracy of each model was calculated at the best threshold, which was defined by the proportion of true predictions of all predictions, to be an additional performance index to enhance the reliability of the result.

2.6 Predicting important variables via sensitivity analysis and connection weights

It is interesting and necessary to consider which of these input variables are most influential. Some of these variables are extremely expensive or cumbersome to gather, requiring an expert assessment of the patients. Since there are 74 input variables for females and 67 input variables for males, the data recording is a significant cost, including financial, material and human resources. Therefore, there is a great deal of motivation to predict important input variables and reduce the number of variables gathered.

Sensitivity analysis is a critical step in network modeling process. It provides an idea for the model dynamics responses to a variation in the values of some input variables. The purpose of sensitivity analysis is to study the behavior of a model, and to assess the importance of each independent variable on the values of the dependent variable of the model.

A novel sensitivity analysis method was used for the models (Gevrey et al., 2003; Olden et al., 2004), which is simple and effective in identifying key variables. The method is introduced as follows:

(1) Sequentially set each input variable to their minimum value (it is ‘-1’ after normalization).

(2) Assess the change in the root mean square error (RMSE) of the network. When each input variable was set to '-1', a changed RMSE was obtained. Then calculate the ratio, which is changed RMSE over initial RMSE.

(3) Rank the input variables according to the value of ratio. The more important variable is combining with the bigger ratio.

(4) Sequentially set each input variable to their maximum value (it is '1' after normalization), then repeat (2) and (3).

(5) Since it was based on the fivefold cross-validation method, two ranking results for each dataset were obtained. In order to get the final result for the whole data, the 'vote method' was applied. It is got by adding together the ratio of changed RMSE and initial RMSE of each ranked variable in five datasets, and the more important variable has a bigger vote ratio number. The contribution of each variable was also concluded according to vote ratio.

In addition, the method of 'leaving one out' was also tried, which was sequentially removing each input variable from the neural network. But it is required to rebuild and retrain the neural network at each step, since there are 5 datasets, and each dataset has 74 input variables for females and 67 for males, it is time-consuming. Therefore, it is not suitable for the developed model.

In the neural network, the connection weights between neurons are the links between the inputs and the outputs. The relative contributions of the independent variables to the predictive output of the neural network depend primarily on the magnitude and direction of the connection weights. Input variables with larger connection weights represent greater intensities of signal transfer, and therefore are more important in the prediction process compared to variables with smaller weights. The approach is described as calculating the product of the raw input-hidden and hidden-output connection weights between each input neuron and output neuron and sums the products, and then calculating the contribution of each input variable (Cui et al., 2011; Olden et al., 2004).

3. Results

3.1 Prediction ability of the ensemble model

To estimate the generalized prediction abilities of the models, testing data were applied to the models, and the outputs were used to do the ROC curve analysis. Table 2 summarizes the results of ROC curve analysis for the female prediction models with different cross-validation datasets, and Table 3 summarizes the results for the male prediction models. In the results tables, SENs, SPEs, PPVs, NPVs were listed and accuracies at the best thresholds. The best thresholds were obtained in the ROC curve analyses of training data, because in practical applications, the thresholds has to be established before using the models to predict unknown and untrained data (i.e., actual risks are unknown and the best threshold cannot be decided).

Table 2. The summary of ROC curve analyses by testing data for female prediction models.

Dataset	Architecture	AUC	SEN	SPE	PPV	NPV	Accuracy
			at best threshold				
1	**N_{HN}=140, best**	**0.96**	**0.94**	**0.86**	**0.76**	**0.97**	**0.88**
	N_{HN}=37, best	0.91	0.59	0.86	0.67	0.81	0.77
	N_{HN}=140, median	0.77	0.94	0.71	0.62	0.96	0.79
2	**N_{HN}=140, best**	**0.92**	**0.82**	**0.91**	**0.92**	**0.91**	**0.88**
	N_{HN}=37, best	0.79	0.59	0.63	0.44	0.76	0.62
	N_{HN}=140, median	0.65	0.71	0.69	0.52	0.83	0.69
3	**N_{HN}=140, best**	**0.88**	**0.82**	**0.80**	**0.67**	**0.90**	**0.81**
	N_{HN}=37, best	0.79	0.71	0.74	0.57	0.84	0.73
	N_{HN}=140, median	0.72	0.59	0.86	0.67	0.81	0.77
4	**N_{HN}=140, best**	**0.89**	**0.76**	**0.86**	**0.72**	**0.88**	**0.83**
	N_{HN}=37, best	0.82	0.71	0.71	0.55	0.83	0.69
	N_{HN}=140, median	0.70	0.76	0.71	0.57	0.86	0.73
5	**N_{HN}=140, best**	**0.89**	**0.88**	**0.80**	**0.68**	**0.93**	**0.83**
	N_{HN}=37, best	0.85	0.71	0.74	0.57	0.84	0.73
	N_{HN}=140, median	0.76	0.82	0.80	0.67	0.90	0.79
Average	**N_{HN}=140, best**	**0.91±0.028**	**0.84±0.061**	**0.85±0.042**	**0.75±0.091**	**0.92±0.031**	**0.85±0.028**
	N_{HN}=37, best	0.83±0.046	0.66±0.059	0.74±0.074	0.56±0.073	0.82±0.030	0.71±0.051
	N_{HN}=140, median	0.72±0.043	0.76±0.116	0.75±0.065	0.61±0.058	0.87±0.053	0.75±0.039

Note: AUC: area under ROC curve, SEN: sensitivity, SPE: specificity, PPV: positive prediction value, NPV: negative prediction value. “best” and “median” separately represent choosing the best networks and median networks in the ensemble method.

Fig. 3 shows the ROC curves of the analyses for the female models, and Fig. 4 shows the ROC curves of the analyses for the male models. By comparing the AUCs and accuracies in Tables 2 and 3, the best architecture is the structure of setting N_{HN} = 140 and choosing the best networks in the ensemble method (marked in bold in Tables 2 and 3), which achieves a good performance on predicting the risk of hip fracture. The average AUC of the female prediction model is 0.91 ± 0.028 (mean ± SD) and the average accuracy at the best threshold is 0.85 ± 0.029. The average AUC of the male prediction model is 0.99 ± 0.015 and the average accuracy at the best threshold is 0.93 ± 0.020. The low SD values confirm that the results are uniform and reliable. For the dataset 1, 3, and 4 in the males, the prediction models had AUC = 1.

Table 3. The summary of ROC curve analyses by test data for males prediction models

Dataset	Architecture	AUC	SEN	SPE	PPV	NPV	Accuracy
			at best threshold				
	N_{HN} =140, best	**1.00**	**1.00**	**0.87**	**0.78**	**1.00**	**0.95**
1	N_{HN} =34, best	0.93	0.43	0.67	0.38	0.71	0.59
	N_{HN} =140, median	0.58	0.86	1.00	1.00	0.94	0.91
	N_{HN} =140, best	**0.96**	**0.86**	**1.00**	**1.00**	**0.94**	**0.95**
2	N_{HN} =34, best	0.92	0.86	0.93	0.86	0.93	0.86
	N_{HN} =140, median	0.87	0.86	0.87	0.75	0.93	0.86
	N_{HN} =140, best	**1.00**	**1.00**	**0.87**	**0.78**	**1.00**	**0.91**
3	N_{HN} =34, best	0.95	0.57	0.67	0.44	0.77	0.64
	N_{HN} =140, median	0.67	1.00	0.80	0.70	1.00	0.86
	N_{HN} =140, best	**1.00**	**1.00**	**0.93**	**0.88**	**1.00**	**0.95**
4	N_{HN} =34, best	0.94	0.57	0.73	0.50	0.79	0.68
	N_{HN} =140, median	0.68	0.86	0.87	0.75	0.93	0.86
	N_{HN} =140, best	**0.99**	**1.00**	**0.87**	**0.78**	**1.00**	**0.91**
5	N_{HN} =34, best	0.92	0.71	0.80	0.63	0.86	0.77
	N_{HN} =140, median	0.85	1.00	0.87	0.78	1.00	0.86
	N_{HN} =140, best	**0.99±0.015**	**0.97±0.056**	**0.91±0.052**	**0.84±0.087**	**0.99±0.024**	**0.93±0.020**
Average	N_{HN} =34, best	0.94±0.011	0.63±0.146	0.76±0.098	0.56±0.171	0.81±0.076	0.71±0.096
	N_{HN} =140, median	0.73±0.111	0.92±0.069	0.88±0.065	0.80±0.105	0.96±0.033	0.87±0.020

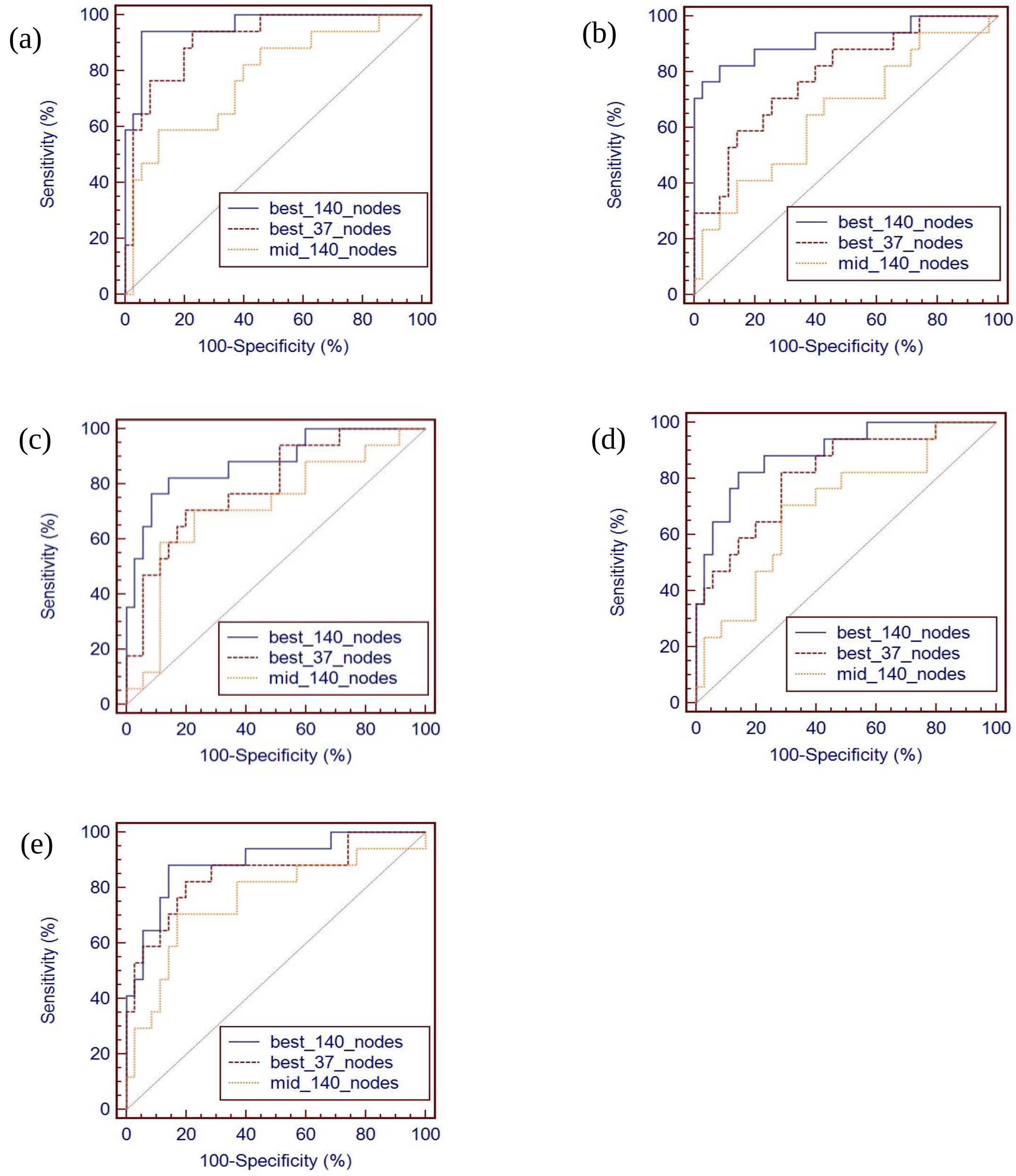


Fig. 3. The ROC curves of the females models: (a) is for dataset 1, (b) is for dataset 2, (c) is for dataset 3, (d) is for dataset 4, and (e) is for dataset 5. "best_140_nodes" means 140 hidden nodes and choosing the best networks in the ensemble method, "mid_140_nodes" means 140 hidden nodes and choosing the median networks in the ensemble method.

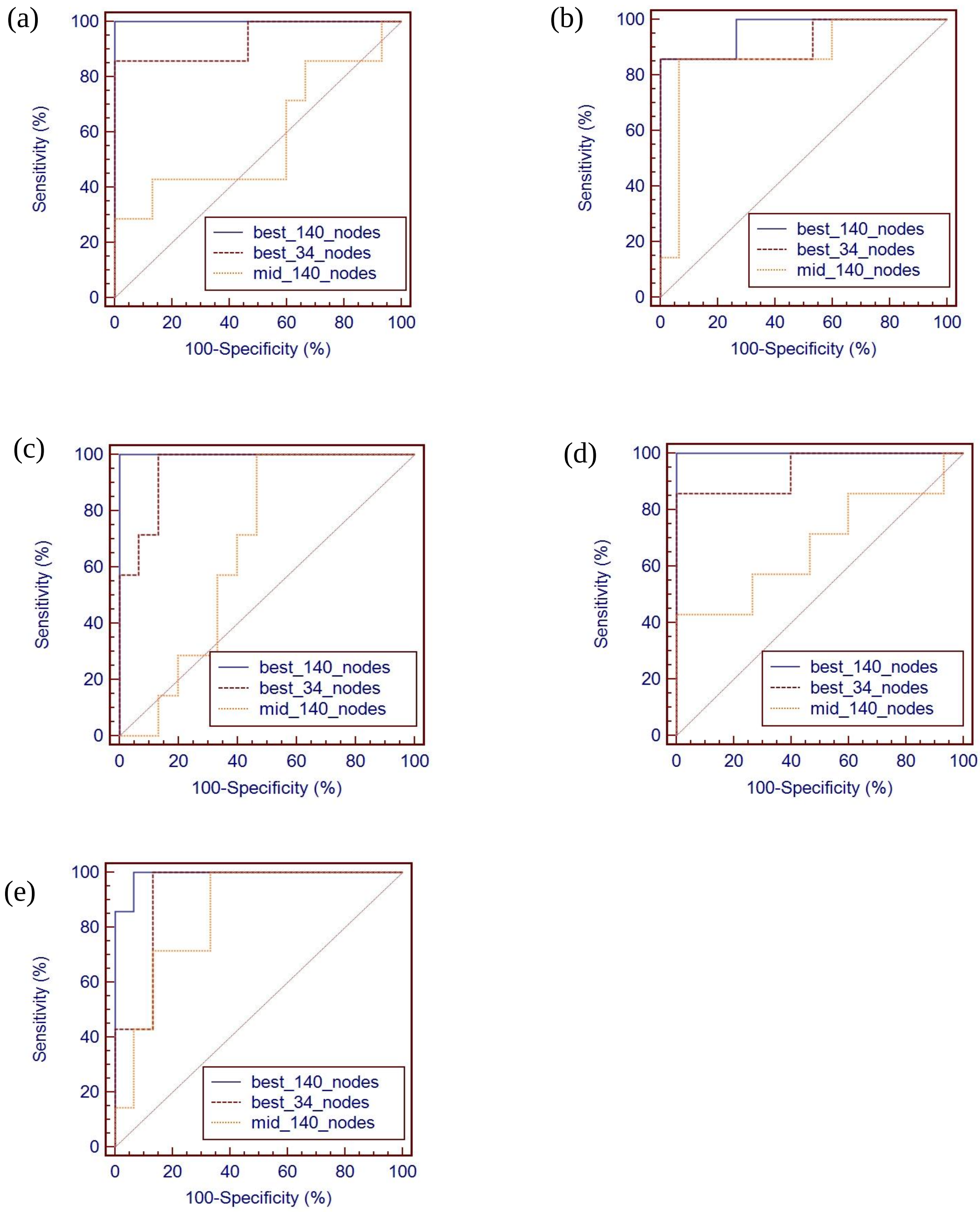


Fig. 4. The ROC curves of the males models: (a) is for dataset 1, (b) is for dataset 2, (c) is for dataset 3, (d) is for dataset 4, and (e) is for dataset 5.

3.2 Results of sensitivity analysis and connection weights approach

The results of sensitivity analysis (i.e., leave one ‘-1’ and leave one ‘1’) and connection weights are based on the best architecture of prediction model (i.e., N_{HN} = 140 and choosing the best networks). For sensitivity analysis, the ratio of changed

RMSE and initial RMSE for females of 5 datasets are shown in Fig. 5, and Fig. 6 is for males. Both Figs. 5 and 6 show that the area under the curve of the top 10 important variables account for the bigger and fast slope changed ratio, which means the top 10 variables have great proportion of contribution.

Based on sensitivity analysis and connection weights, the ranking results of top 10 important variables separately for females and males are shown in Table 4. ‘Ranked variables’ shows the name of selected important variables via ‘vote’ from 5 datasets. ‘Contribution’ gives a clear expression of importance for each factor, which is the average of the specific variable contribution in each dataset. For connection weights, negative contributions represent inhibitory effects on neurons and decrease the value of the predicted response, whereas positive contributions represent excitatory effects on neurons and increase the value of the predicted response. ‘Rank’ presents the index number of variables after ranking.

To compare with our previous study (Lan et al., 2010; Tseng et al., 2013), the variables in red are the same with the factors causing hip fracture shown by Tseng et al. (2013), the variables with underline are exactly the same with the results got by Lan et al.(2010). The 10 variables in bold are the most important factors got from both connections weights approach and sensitivity analysis, including 'Total BMD value', 'Self-assessed health comparison with 1 year ago', 'Self-assessed health-Current', 'history of fall at home', 'Height', 'BMI', 'Hypertension', 'MMSE score', 'fecal incontinence', and 'Education', of which, 7 variables are the same factors with our previous study (Lan et al., 2010; Tseng et al., 2013), and other important variables, such as health status, are key factors causing falls for elders from a previous research (Stalenhoef et al., 2002).

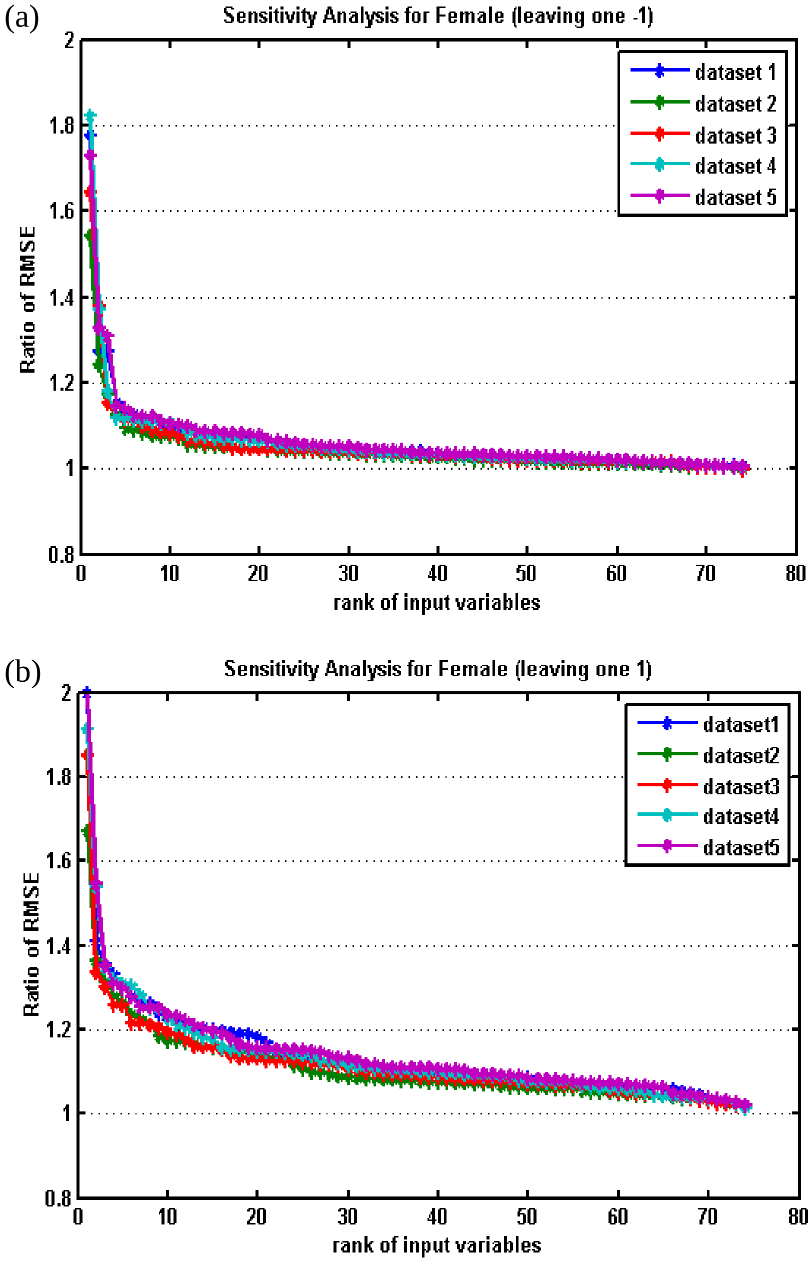


Fig. 5. Ratio of RMSE for females. (a) is got when setting each input variable ‘-1’, (b) is got when setting each input variable ‘1’.

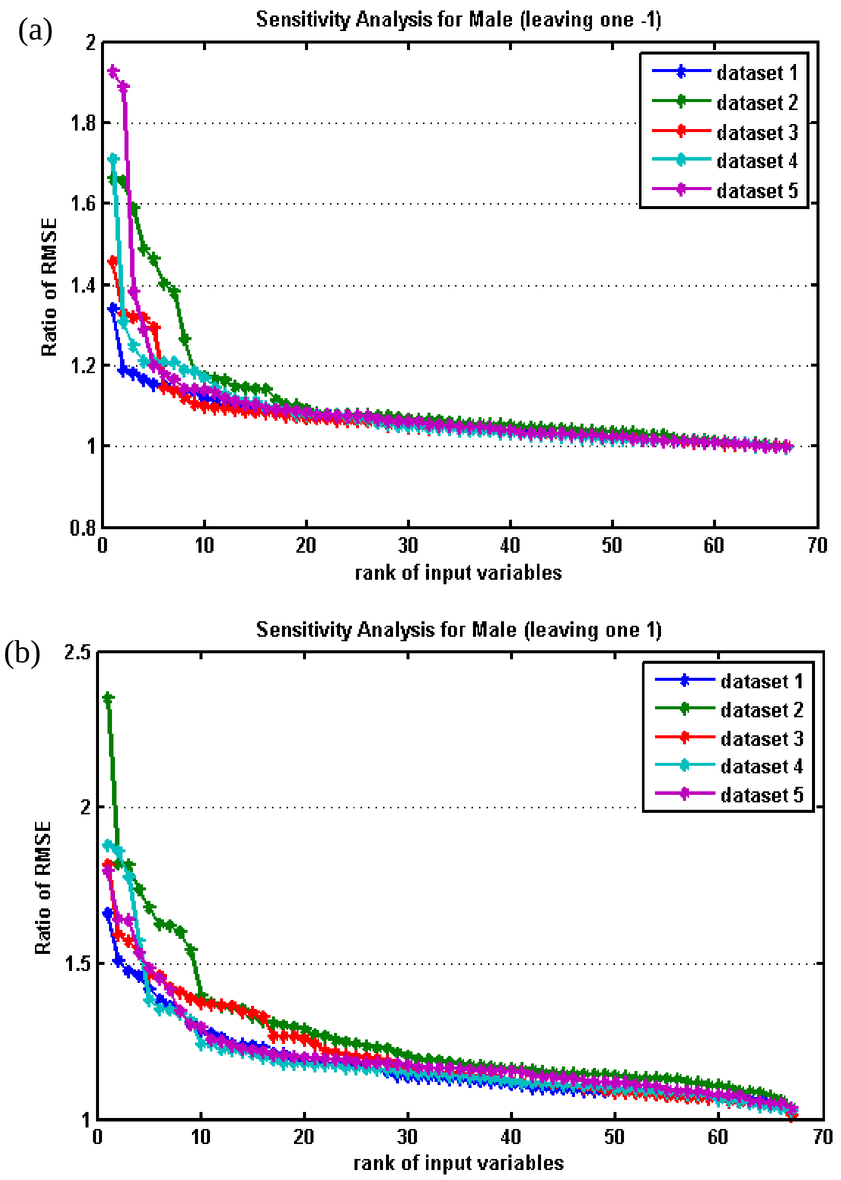


Fig.6. Ratio of RMSE for males. (a) is got when setting each input variable '-1', (b) is got when setting each input variable '1'.

4. Conclusions and discussions

In this study, ensemble BPNN models applied to predict the risk for hip fracture in elder adults are presented. Several architectures of BPNN were tested and ensemble model to establish the prediction models. By ROC curve analysis, the performances of these architectures were compared to each other. The architecture of setting $N_{HN} = 140$ and choosing the best networks in the ensemble models turned out to be the best structure for predicting the risk of hip fracture in elder adults. Also, the SENs, SPEs, PPVs, and NPVs of this architecture are better than the others. That means this architecture is suitable for any situations no matter true cases or false cases are more necessary to be detected. The male model had better performances than the female model, because the male case has a lower complexity than female case. The successful results confirm that ANNs are useful to be applied to approximate a highly dimensional and nonlinear function even such a complex system as a biomedical model with 74 inputs (independent variables). Furthermore, with the help of sensitivity analysis and connection weights methodology, the top 10 important variables are predicted, which reduces all kinds of troubles due to too many independent variables.

In this study, the models that correspond to different activation functions were established, similar results were obtained, and only the best results were presented, which is the activation function of hidden neurons is a tangent sigmoid transfer function, and the activation function of output neurons is a simple linear transfer function. For the male prediction models, the results of ROC curve analysis showed “AUC = 1”. Hip fracture prediction is very complex and cannot be predict so easily. “AUC = 1” was obtained when it automatically chose specific threshold for each dataset. Besides, the testing sample size became much smaller after splitting the whole database into 5 datasets. If a huge database is selected with the same threshold for 5 datasets, it is impossible to get “AUC = 1”.

It is argued that choosing the best networks in the ensemble model to get perfect result is by chance, because it may be local minima. Logically, other testing data independent from the testing data can be used in the training process. However, as shown in the results summarized in Tables 2 and 3, the architecture choosing the median networks has a bad performance on the generalized prediction ability. Therefore, it is not necessary to examine more detailed performances for this consideration. Alternatively, in order to solve this problem, GA can be used to pre-process the initial weights to optimize our ANN model. To some extent, ensemble method is another kind of mutation. With the theory of survival of the fittest, the combination of GA and ANN can get global minima (Chang et al., 2012).

Besides, conditional neural networks should be studied in future. In our study, the sample included a total of 725 cases and controls data, of which, 228 cases were the patients admitted to the National Taiwan University Hospital with first low-trauma hip fracture, and 215 hospital controls (patients in the same hospital but without hip fracture) and 282 community controls (randomly selected dwellers). With the method of fivefold cross-validation and ensemble model, although the training and validation data were randomly selected, it can only make sure test data have equal proportion in the three part of sample data. However, it will be much better and more sensible to select equal proportion of training and validation data in the three part of sample data, which can be regard as conditional neural networks.

In the method of coding variables, binary coding genetic algorithm (GA) was used to code the raw data. Using this method, the multiple data can be easily transferred into one value. Furthermore, this binary coding way can be applied to other analyses in future works, such as testing the robustness of the models. Noise can be easily added to the input signals by randomizing those binary data.

Table 4 presents the rank and contributions of 10 selected important factors from three methods (i.e., leave one '-1', leave one '1' and connection weights) to predict hip fracture. It is hard to say which method is better. Lan et al. (2010) summarized that Low milk intake, peak flow rate, hand grip strength, and bone mineral density in women and low mini-mental state examination (MMSE) score and bone mineral density (BMD) in men were further identified to be independently associated with elevated hip fracture risk. All these six factors are within the results of leave one '-1' in Table 4 (i.e., variables with underline). Leave one '1' and connection weights obtained more similar variables comparing with Tseng et al. (2013). Combining the three methods, this study identified 10 significant factors (i.e. variables in bold in Table), including 'Total BMD value', 'Self-assessed health comparison with 1 year ago', 'Self-assessed health-Current', 'history of fall at home', 'Height', 'BMI', 'Hypertension', 'MMSE score', 'fecal incontinence', and 'Education'.

Table 4 Ranking results of top 10 important variables separately using connection weights and sensitivity analysis

	Female		Male		Rank
	Ranked variables	Contribution (%)	Ranked variables	Contribution (%)	
Leave One "-1"	**'Total BMD value'**	14.2	**'MMSE score'**	12.7	1
	'Self-assessed health comparison with 1 year ago'	11.0	**'Self-assessed health-Current'**	11.4	2
	'Height'	10.2	**'Total BMD value'**	10.3	3
	'mobility difficulty'	9.30	**'Hypertension'**	10.1	4
	'Self-assessed health-Current'	9.30	**'Height'**	9.74	5
	'intake of milk'	9.28	**'Self-assessed health comparison with 1 year ago'**	9.50	6
	'BMI'	9.22	'use of anti-hypertensive'	9.28	7
	'use of anti-hypertensive'	9.21	'urinary incontinence'	9.13	8
	'average hand grip strength'	9.17	'intake of calcium'	9.05	9
	'peak expiratory flow rate'	9.07	'cataract'	8.80	10
Leave One "1"	**'Total BMD value'**	14.4	**'Self-assessed health-Current'**	11.5	1
	'history of fall at home'	10.0	**'fecal incontinence'**	10.8	2
	'average hand grip strength'	9.85	'urinary incontinence'	10.6	3
	'number of stairs in a flight'	9.66	'number of stairs in a flight'	10.1	4
	'use of walking aids'	9.53	**'history of fall at home'**	9.96	5
	'BMI'	9.51	**'ADL difficulty'**	9.71	6
	'Headache or migraine'	9.31	**'Education'**	9.62	7
	'use of analgesic'	9.31	**'Total BMD value'**	9.50	8
	'bone fracture location'	9.26	'intake of calcium'	9.24	9
	'Weakness'	9.18	'Parkinson's disease'	8.98	10
Connection Weights	**'Total BMD value'**	-7.44	**'Total BMD value'**	-4.11	1
	'Self-assessed health comparison with 1 year ago'	3.44	**'Self-assessed health-Current'**	-3.94	2
	'number of stairs in a flight'	3.29	'ADL difficulty'	-3.90	3
	'Height'	3.17	**'fecal incontinence'**	3.62	4
	'Self-assessed health-Current'	-3.00	**'Education'**	-3.25	5
	'Headache or migraine'	2.50	**'Height'**	2.86	6
	'bone fracture location'	2.46	'Living arrangement'	-2.68	7
	'BMI'	-2.37	**'MMSE score'**	-2.56	8
	'heart disease'	2.24	**'Hypertension'**	2.51	9
	'use of analgesic'	2.20	**'history of fall at home'**	2.26	10

5. Acknowledgement

This research was financially supported by the Ministry of Science and Technology (Grant Number: NSC102-2221-E-155-028-MY3), and sponsored by China Scholarship

Council, China (CSC, File No. 2010695013). This research was also supported by the Center for Dynamical Biomarkers and Translational Medicine, National Central University, Taiwan, which is sponsored by the Ministry of Science and Technology (Grant Number: NSC 102-2911-I-008-001).

6. References

Basheer IA, Hajmeer M, 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods 43(1), 3-31.

Baxt WG, 1995. Application of artificial neural networks to clinical medicine. The Lancet 346 (8983), 1135-1138.

Benetos IS, Babis GC, Zoubos AB, Benetou V, Soucacos PN, 2007. Factors affecting the risk of hip fractures. Injury 38(7), 735-744.

Chang YT, Lin J, Shieh JS, Abbod MF, 2012. Optimization the Initial Weights of Artificial Neural Networks via Genetic Algorithm Applied to Hip Bone Fracture Prediction. Advances in Fuzzy Systems, Volume 2012, Article ID 951247, 9 pages.

Cross S, Harrison R, Kennedy R, 1995. Introduction to neural networks. Lancet 346(8982), 1075–1079.

Cui X, Lin C-W, Abbod MF, Liu Q, Shieh J-S, 2012. Diffuse Large B-cell Lymphoma Classification Using Linguistic Analysis and Ensembled Artificial Neural Networks. Journal of the Taiwan Institute of Chemical Engineers, 43(1), 15-23.

Cui X, Abbod MF, Liu Q, ShieH J-S, Chao TY, Hsieh CY, Yang YC, 2011. Ensembled artificial neural networks to predict the fitness score for body composition analysis. The Journal of Nutrition, Health and Aging, 15 (5), 341-348.

Dubey A, Koval KJ, Zuckerman JD, 1999. Hip fracture epidemiology: a review. Am J Orthop 28(9), 497-506.

Eller-Vainicher C, Chiodini I, Santi I, Massarotti M, Pietrogrande L, Cairoli E, Beck-Peccoz P, Longhi M, Galmarini V, Gandolini G, Bevilacqua M, Grossi E, 2011. Recognition of morphometric vertebral fractures by artificial neural networks: analysis from GISMO Lombardia Database. PLoS One 6(11), e27277.

Gevrey M, Dimopoulos I, Lek S, 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Modell 160, 249-264.

Haentjens P, Lamraski G, Boonen S, 2005. Costs and consequences of hip fracture occurrence in old age: an economic perspective. Disabil Rehabil 27(18-19), 1129-1141.

Hansen LK, Salamon P, 1990. Neural network ensembles. IEEE transactions on Pattern Analysis and Machine Intelligence 12(10), 993-1001.

Jette AM, Harris BA, Cleary PD, Campion EW, 1987. Functional recovery after hip fracture. Arch Phys Med Rehabil 68(10), 735-740.

Johnell O, Kanis JA, Oden A, Sernbo I, Redlund-Johnell I, Petterson C et al., 2004. Mortality after osteoporotic fractures. Osteoporos Int 15(1), 38-42.

Kung S-Y, Hwang J-N, 1998. Neural networks for intelligent multimedia processing. Proceedings of the IEEE 86(6), 1244-1272.

Lan T-Y, Hou S-M, Chen C-Y, Chang W-C, Lin J, Lin C-C, Liu W-J, Shih T-F, Tai T-Y, 2010. Risk Factors for Hip Fracture in Older Adults: Case-Control Study in Taiwan.

Osteoporosis International 21(5), 773-784.

Lek, S, Guegan, JF, 2000. Application to Ecology and Evolution. Springer, Berlin. Artificial Neuronal Networks, (Eds.).

Lin CC, Ou YK, Chen SH, Liu YC, Lin J, 2010. Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. Injury 41(8), 869–873.

Lin E, Hwang Y, Wang SC, Gu ZJ, Chen EY, 2006. An artificial neural network approach to the drug efficacy of interferon treatments. Pharmacogenomics 7(7), 1017-1024.

Liu Q, Cui X, Abbod MF, Huang S-J, Han Y-Y, Shieh J-S, 2011. Brain Death Prediction Based on Ensembled Artificial Neural Networks in Neurosurgical Intensive Care Unit. Journal of the Taiwan Institute of Chemical Engineers, 42(1), 97-107.

Lutz Prechelt, 1998. Automatic early stopping using cross validation: quantifying the criteria. Neural Networks 11, 761–767.

Magaziner J, Simonsick EM, Kashner TM, Hebel JR, Kenzora JE, 1989. Survival experience of aged hip fracture patients. Am J Public Health 79(3), 274-278.

Magaziner J, Fredman L, Hawkes W, Hebel JR, Zimmerman S, Orwig DL et al., 2003. Changes in functional status attributable to hip fracture: a comparison of hip fracture patients to community-dwelling aged. Am J Epidemiol 157(11), 1023-1031.

Moller MF, 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural networks 6(4), 525-533.

Olden JD, Joy MK, Death RG, 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modeling, 178, 389-397.

Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV, 2001. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. Journal of clinical epidemiology 54(11), 1159-1165.

Ottenbacher KJ, Linn Richard T, Smith Pamela M, Illig Sandra B., Mancuso Melodee, Granger Carl V, 2004. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Annals of Epidemiology 14(8), 551-559.

Pulido M, Melin P, Castillo O, 2013. Optimization of type-2 fuzzy integration in ensemble neural networks for predicting the US Dolar/MX pesos time series. IFSA/NAFIPS: 1508-1512.

Pulido M, Melin P, Castillo O, 2014. Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange. Inf. Sci. 280: 188-204.

Recknagel, F. 2003. Understanding Ecology by Biologically-Inspired Computation. Springer, Berlin. Ecological Informatics.

Rodríguez JD, Pérez A, Lozano JA, 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Transactions on Pattern Analysis and Machine

Intelligence, 32(3), 569-575.

Sarle, WS, 1995. Stopped training and other remedies for overfitting. Proceedings of the 27th symposium on the interface of computing science and statistics, 352-360.

Sarle, WS, 2002. Neural Network FAQ, part 2 of 7: Learning, periodic posting to the Usenet newsgroup comp. ai. neural-nets. URL: ftp://ftp.sas.com/pub/neural/FAQ.html.

Sarle, WS, 2002. Neural Network FAQ, part 3 of 7: Generalization, periodic posting to the Usenet newsgroup comp. ai. neural-nets. URL: ftp://ftp.sas.com/pub/neural/FAQ.html.

Stalenhoef PA, Diederiks JPM, Knottnerus JA, Kester ADM, Crebolder HFJM, 2002. A risk model for the prediction of recurrent falls in community-dwelling elderly: A prospective cohort study. Journal of Clinical Epidemiology 55,1088-1094.

Tseng WJ, Hung LW, Shieh JS, Abbod MF, Lin J, 2013. Hip fracture risk assessment: artificial neural network outperforms conditional logistic regression in age- and sex-matched case control study. BMC Musculoskelet Disord 14, 207.

Wehren LE, Magaziner J, 2003. Hip fracture: risk factors and outcomes. Curr Osteoporos Rep 1(2), 78-85.

Yeh JR, Fan SZ, Shieh JS, 2009. Human heart beat analysis using a modified algorithm of detrended fluctuation analysis based on empirical mode decomposition. Medical Engineering & Physics 31(1), 92-100.

Zhou ZH, Wu J, Tang W, 2002. Ensembling neural networks: Many could be better than all. Artificial Intelligence 137, 239-263.