Occupational Ergonomics 4 (2004) 291–305
IOS Press

Classification of jobs with risk of low back disorders by applying data mining techniques

Jozef Zurada[a], Waldemar Karwowski[b,*] and William Marras[c]
[a]*Computer Information Systems, College of Business and Public Administration, University of Louisville, Louisville, KY 40292, USA*
[b]*Center for Industrial Ergonomics, University of Louisville, Lutz Hall, Room 445, Louisville, KY 40292, USA*
Tel.: +1 502 852 7173; Fax: +1 502 852 7397; E-mail: karwowski@louisville.edu
[c]*Industrial, Systems and Welding Engineering Department, Ohio State University, 1971 Neil Avenue, Columbus, OH 43210, USA*

Abstract. Work related low back disorders (LBDs) continue to pose significant occupational health problem that affects the quality of life of the industrial population. The main objective of this study was to explore the application of various data mining techniques, including neural networks, logistic regression, decision trees, memory-based reasoning, and the ensemble model, for classification of industrial jobs with respect to the risk of work-related LBDs. The results from extensive computer simulations using a 10-fold cross validation showed that memory-based reasoning and ensemble models were the best in the overall classification accuracy. The decision tree and memory-based reasoning models were the most accurate in classifying jobs with high risk of LBDs, whereas neural networks and logistic regression were the best in classifying jobs with low risk of LBDs. The decision tree model delivered the most stable results across 10 generations of different data sets randomly chosen for training, validation, and testing. The classification results generated by the decision tree were the easiest to interpret because they were given in the form of simple 'if-then' rules. These results produced by the decision tree method showed that the peak moment had the highest predictive power of LBDs.

Keywords: Low back disorders, assessment of lifting jobs, knowledge discovery, data mining techniques

1. Introduction

Work-related low back disorders (LBDs) continue to pose significant occupational health problem that can affect the quality of life of the industrial population [2,18]. In 1988, overexertion injuries across all industries accounted for 28.2% of all work injuries involving disability, while approximately 25% of all worker compensation claims were related to back injuries [24]. According to the Liberty Mutual Workplace Safety Index of Leading Occupational Injuries [18], the leading cause of workplace injuries in 2001 was an overexertion, accounting for 27.3% of all injuries. The top three injury causes (overexertion, falls on same level and bodily reaction) were the fastest growing of all injury causes, representing 50.1 percent of the total costs, i.e., about $23 billion a year or $450 million a week. The above include a large number of disabling injuries to the lower back due to either cumulative exposure to manual handling of loads over a long period of time or to isolated incidents of overexertion when handling heavy objects.

*Corresponding author.

Epidemiological studies [25,26,34] demonstrated that frequency rates (number of injuries per man-hour on the job) and severity rates (number of hours lost due to injury per man-hour on the job) of back injuries increase significantly when: (i) heavy objects are lifted, (ii) objects are bulky, (iii) objects are lifted from the floor, (iv) objects are frequently lifted, and (v) loads are lifted asymmetrically (by one hand or at the side with the torso twisted).

Marras developed a multiple logistic regression model based on the combination of five trunk motion and workplace factors, including: (1) lifting frequency, (2) trunk twisting velocity, (3) load moment, (4) the trunk sagittal angle and (5) trunk lateral velocity [20]. Load moment and lifting frequency are the workplace factors. Lateral trunk velocity, twisting trunk velocity and satiggal flexion angles are the trunk motion factors. As the magnitude of each of these variables increases, the risk increases. The predictive power of the above described model was more than three times greater than that of the NIOSH Lifting Guide of 1981, and allows for distinguishing between high and low risk of occupationally related low back disorders with the odds ratio of 1:10.7. It should be noted that the 1981 and 1991 NIOSH Guides are conceptually comparable. In the 1991 NIOSH Revised Lifting Equation [25,34,35], the relationship between the load lifted on the job and Recommended Weight Limit (RWL) is defined by the lifting index (LI). The LI value at 1.0 is equivalent to the old action limit (AL). The LI value of 3.0 is equivalent to the concept of Maximum Permissible Limit (MPL). The MPL value was set to be three times the Action Limit (AL) according to the 1981 NIOSH Guide.

1.1. Objectives

Knowledge discovery is defined as the process of identifying valid, novel, and potentially useful patterns, rules, relationships, rare events, correlations, and deviations in data [9]. This process relies on well-established technologies, such as machine learning, pattern recognition, statistics, neural networks, fuzzy logic, evolutionary computing, database theory, artificial intelligence, and high performance computing. These technologies are often used together to find relevant knowledge in data. For example, Karwowski et al. [17] and Zurada et al. [37] used the neural network techniques to explore their effectiveness in detecting high risk jobs with respect to LBDs. The results of the study [37] show that an artificial neural network-based diagnostic system can be used as an expert system which, when properly trained, will allow us to classify lifting jobs into two categories of associated LBDs risk, based on the available job characteristics data. In an earlier paper, Karwowski et al. [16] proposed to use the catastrophe theory in modeling the risk of injury in manual lifting tasks.

For comparison purposes, it should be noted that the 1981 NIOSH Guide correctly classified 91% of low risk jobs, but only 10% of the high risk jobs, whereas the 1991 NIOSH Guide correctly classified 55% of the low risk jobs and 73% of the high risk jobs [22]. The classification system developed by Zurada et al. [37], classified 65 out of 87 cases correctly (74.7%). For 50 low risk jobs, only 14 jobs (28%) were incorrectly classified as high risk jobs. Out of 37 high risk jobs, eight jobs (21.6%) were incorrectly classified into low risk job category. Therefore, the neural network-based system developed in this study appears to be an improvement over the application of NIOSH Guides (1981 and 1991) [25, 34], especially with respect to identification of the high risk jobs (78.4% correct classification rate). The developed neural network-based classification system also showed great promise because it significantly reduces the time consuming job analysis and classification performed by traditional methods.

In the ergonomics literature, there are limited examples of studies that apply the artificial intelligence techniques for classifying the risk of injury due to manual lifting tasks performed in industry. However, such techniques have been successfully used for different business problems. For example,

memory-based reasoning tool (along with neural networks) was used for fraud detection in long-distance telephone services using variables such as frequency of calls, time of day, duration, and geography [28]. Radding [28] suggests, however, that, "No technique solves every problem and even the experts aren't always sure which technique will work best in a given situation. Experimentation with multiple techniques is the rule." Several studies found that ensemble (combined) models are more powerful than single models when used for classification tasks. For example, in predicting foreign exchange rates, ensemble models consisting of different neural networks outperformed those using the single best network [36]. And in a medical diagnostic decision support system, combined models were significantly more accurate than 23 single models for four of five applications studied [19].

The main objective of this study was to explore the application of various data mining techniques for classification of industrial jobs with respect to risk of work-related LBDs. Such investigation was undertaken in order to assess the potential benefits of different knowledge discovery tools for hazard analysis and injury prevention due to manual handling of loads in the occupational environments.

2. Model development and experimental data

2.1. Knowledge discovery tools

The knowledge discovery process is typically composed of the following phases: understanding the overall problem domain; obtaining a data set; cleaning, preprocessing, transforming, and reducing data; applying data mining tools; interpreting mined patterns; and consolidating and implementing discovered knowledge. Data mining, which is an important phase in the knowledge discovery process, uses a number of analytical tools such as logistic regression analysis, discriminant analysis, neural networks, decision trees, fuzzy logic and sets, rough sets, genetic algorithms, association rules, and k-nearest neighbor (memory-based reasoning) which are suitable for the tasks of classification, prediction, clustering, summarization, aggregation, and optimization.

Classification is the task that we deal with in this paper. This is the most common and perhaps the most straightforward data mining task. The binary classification task consists of examining the features of a newly presented object and assigning it to one of a predefined set of two classes or outcomes (high risk job or low risk job). A data mining model employing one of the mentioned data mining tools must typically be trained using pre-classified examples. The goal is to build a model that will be able to accurately classify new data based on the outcomes and the interrelation of several variables such as lift rate, peak twist velocity, peak moment, peak sagittal angle, and peak lateral velocity contained in the training set.

2.2. Experimental data

Similarly to our previous study [37], the experimental data collected by Marras [21] was used for the purpose of data mining. This field data was collected in industry based on the surveillance of the trunk motions and quantification of workplace factors in high and low risk of LBDs repetitive tasks. A total of 403 industrial lifting jobs from 48 manufacturing companies were analyzed. Based on examination of the injury and medical records, all jobs were divided into two groups: 1) high risk of LBDs, and 2) low risk of LBDs. According to Marras [21], whenever possible, company medical records were used to categorize risk. Each job was weighted proportionally to the number of worker-hours from which the injury and turnover rates were derived. The odds ratio for LBDs was defined as the ratio of the

probability that an LBD occurs (probability of being in the high risk LBD group) to the probability that an LBD does not occur (probability of being in the low risk LBD group).

The jobs with *low risk* of LBDs were defined as those jobs with at least three years of records showing no injuries and no turnover. Turnover was defined as the average number of workers who left a job per year. The jobs with *high risk* of LBDs were those jobs associated with at least 12 injuries per 200,000 hours of exposure. The high risk group category incidence rate of LBDs corresponded to the 75th percentile value of the 403 jobs examined. Out of the 403 jobs examined, 124 of the jobs were categorized as being in the low risk group, and 111 were categorized as being in the high risk group. The remainder of the analyzed jobs (168), which were categorized as "medium risk", were not utilized in this paper. The dependent variables in this study [21] consisted of workplace and trunk motion characteristics which were indicative of each job.

3. Modeling methods

This section provides the basic features of the data mining tools we have used in our study. For more formal and thorough description on neural networks, logistic regression, decision trees, memory-based reasoning, and the ensemble model, the reader is referred to [1,7,11–15,23,27], and SAS Enterprise Miner [30].

3.1. Neural networks

Neural networks are mathematical models of the human brain. They try to mimic the way human brain functions and processes information. Neural networks are particularly useful for prediction tasks because they can efficiently detect very complex nonlinear relationships and interrelations between independent and dependent variables, and between dependent variables themselves. A neural network is composed of a set of elementary computational units, called neurons, linked together through weighted connections. Each neuron is typically composed of a summation node and a sigmoidal activation function. These neurons are organized in layers so that in a fully connected network every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. These layers can be of three types: input, hidden, and output. In a network, the information flows through the three layers from an input layer, through a hidden layer, to an output layer. The input layer transmits information to the hidden layer; it does not perform any calculation. The output layer produces the final results. The hidden layer has no direct contact with the external environment. Its function is to take the relationship between the input (independent) variables and the output (dependent) variable and adapt it more closely to the data. Neural networks learn from training examples/patterns in a supervised mode (with a teacher) and an unsupervised mode (without a teacher). In the supervised mode used in this study, the training patterns are shown repeatedly to the network, one a time, until the cumulative error value, measured as the difference between the actual class and the predicted class values for the dependent variable RISK reaches minimum on the validation data set. In this study, we have used a popular feed-forward neural network with back-propagation and the default learning algorithm provided by the SAS Enterprise Miner data mining software. We used the hidden layer with 1 neuron to protect the network from overfitting. For more details, see [11–14,23,30].

3.2. Logistic regression

Linear regression is used to model continuous-value functions. The logistic regression method estimates the probability p that the dependent variable will have a given value.

Logistic regression is used when the output variable of the model is defined as a binary categorical. Suppose that the dependent variable RISK has 2 possible classes coded as 1 and 0 representing "high" and "low" risk of LBDs, respectively. Based on the available data one can compute the probabilities for both values for the given input sample: $p(RISK_j = 0) = 1 - p_j$ and $p(RISK_j = 1) = p_j$, where j is a case/observation number. The model that will fit these probabilities is accommodated linear regression of the form:

$$\log\left(\frac{p_j}{1-p_j}\right) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \ldots + \beta_n X_{nj}.$$

This equation is known as the linear logistic model. The function $\log\left(\frac{p_j}{1-p_j}\right)$ is often written as logit(p). Based on a training data set one can find the values for the parameters $\alpha, \beta_1, \beta_2, \beta_3, \ldots \beta_n$. Let us assume that for a new sample/case j, to be classified as 1 (high risk job) or 0 (low risk job), for which the values of $X_{1j}, X_{2j}, X_{3j}, \ldots X_{nj}$ are known, the value of logit(p) $= 0.6$. Then the probability that the variable RISK takes the value of 1 is $p(RISK = 1) = \frac{e^{0.6}}{1+e^{0.6}} = 0.65$, and the probability that it takes the value of 0 is $1 - 0.65 = 0.35$. In this study we used the stepwise approach for the selection of the variables. For a more formal approach, see [1,7,15].

3.3. Decision trees

Decision trees are efficient in classification tasks as well. They are flow-chart-like tree structures, where nodes denote the tests on attributes, branches represent conditions, and leaf nodes represent outcomes or classes. Like neural networks, they learn from data in a supervised manner. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the prediction class for that sample. Decision trees can easily be converted to if-then rules that represent the relationships among independent variables and a dependent variable. Rules are very important because they provide compact explanation of data and give more insight into the operation of the model.

The most popular ID3 algorithm for decision tree induction, which we used in the study, is a computationally very intensive greedy algorithm that builds decision trees in a top-down recursive divide-and-conquer manner. It splits the original data into finer and finer subsets, or clusters. The attribute that has the most predictive power is always placed at the top node (the root) of the tree. Each rule represents a unique path from the root to each leaf. At each node in the tree, one can measure the number of cases/samples entering the node, the way those cases would be classified if this were a leaf node, and the percentage of cases classified correctly at each node. The goal of the algorithm is to make the clusters at the nodes purer and purer by progressively reducing the impurity/disorder in the original data set. Impurity/disorder is measured by entropy (the concept borrowed from information theory). Entropy measures the information content in a cluster of data. There are two issues the algorithm faces: (1) what variable to place at the top node and what variables to place at subsequent nodes, and where to split the data; and (2) how much splitting is appropriate? The algorithm tests all possible splits on all possible independent variables and it then computes all the resulting gains in purity, and picks up a split that maximizes the gain. The algorithm then finds the minimum number of splits so that the error rate on the validation data set will be minimum. The fewer the splits, the fewer branches are and the more understandable the tree is. For a more formal description, the reader is referred to [8,15,23,27].

Table 1
Illustration of the k-nearest neighbor algorithm for the binary dependent variable

Case id	Dependent variable: RISK (High, Low)	Case ranking based on the distance to the probe: (1 – closest, 5 – farthest)
6	High	1
15	Low	2
45	High	5
79	High	4
230	High	3

k	Case id of nearest neighbors	Target value of nearest neighbors	Posterior probabilities of the probe
1	79	High	Prob (High) = 100.0% Prob (Low) = 0%
3	6, 15, 230	High, Low, High	Prob (High) = 66.7% Prob (Low) = 33.3%
5	6, 15, 45, 79, 230	High, Low, High, High, High	Prob (High) = 80.0% Prob (Low) = 20.0%

3.4. Memory-based reasoning

The Memory-based reasoning is a modeling tool that uses a k-nearest neighbor algorithm to categorize or predict observations. The k-nearest neighbor algorithm takes a data set and a probe (an instance to be classified), where each observation in the data set is composed of a set of variables and the probe has one value for each variable. The distance between an observation and the probe is calculated. The k observations that have the smallest distances to the probe are the k-nearest neighbor to that probe. The k-nearest neighbors are determined by the Euclidean distance between an observation and the probe. Based on the target values of the k-nearest neighbors, each of the k-nearest neighbors votes on the target value for a probe. The votes are the posterior probabilities for the class dependent variable (RISK).

Table 1 shows the voting approach of these neighbors for a binary dependent variable RISK when different values of k are specified. Say, cases 6, 15, 45, 79, and 230 are the five closest cases to the probe, where cases 6 and 45 have the shortest and the longest distances to the probe, respectively. The k-nearest neighbors are first k cases that have the closest distances to the probe. If the value of k is set to 3, the target values of the first three nearest neighbors (6, 15, and 230) are used. The target values for these three neighbors are High (1), Low (0), and High (1). Therefore, the posterior probability for the probe to have the target value High (1) is 2/3 (66.7%). Similarly, if $k = 5$, the target values for the five nearest neighbors (6, 15, 230, 79, and 45) are used. The target values for these five neighbors are High (1), Low (0), High (1), High (1), and High (1), respectively. Therefore, the posterior probability for the probe to have the target value High (1) is 4/5 (80%) [30].

The method requires no model to be fitted, or function to be estimated. Instead it requires all observations to be maintained in memory, and when a prediction is required, the method recalls items from memory and predicts the probability of the RISK variable for a particular case. Two crucial choices in the nearest neighbor-method are the distance function and the cardinality k of the neighborhood. We used the Euclidean distance measure for all the independent variables to calculate the similarity between the cases. After performing several experiments, we chose $k = 14$ because this value of k seemed to give the lowest misclassification rates. This means that 14 most similar cases (neighbors) had the influence on the classification of variable RISK. For more details, refer to [11,13,23].

3.5. Ensemble model

The Ensemble model creates a new model by averaging the posterior probabilities for class targets from multiple models, i.e., the neural network, logistic regression, and decision tree models. (SAS Enterprise Miner, which was used for computer simulation, does not allow one to include the memory-based reasoning model in the ensemble model.) The new model is then used to score new data. The ensemble model can only be more accurate than the individual models if the individual models disagree with one another [30]. This approach can yield a potentially stronger solution.

4. Data set and computer simulation

The data set used in the computer simulation contained 6 variables and 235 cases/observations. The five independent variables describing occupational risk factors for development of LBDs were (1) lift rate in number of lifts per hour (LIFTR), (2) peak twist velocity average (PTVAVG), (3) peak moment (PMOMENT), (4) peak sagittal angle (PSUP), and (5) peak lateral velocity maximum (PLVMAX). The dependent or target variable was risk of LBD (RISK), which took two values 0 (124 cases −52.8%) and 1 (111 cases −47.2%) that represented low and high risk jobs, respectively.

We used stratified sampling to divide the data set into three non-overlapping parts and maintain the proportion of 47.2% of high risk jobs and 52.8% of low risk jobs in each of the three parts. Consequently, we allocated 40% (94 cases), 30% (71 cases), and 30% (70 cases) to the training, validation, and test sets, respectively. The training set was used to build the models. The validation set, although not directly applied in training, was used to fine tune the models. It may also be utilized to test the performance of the models. The test set was applied to obtain an unbiased performance of the models in the real world environment. Splitting this small data set containing 235 cases into three parts implies a loss of information as the number of observations used in building and testing the models is reduced. Also, the extra sampling process may introduce a new source of variability and decrease the stability of the results. To counter this, however, we applied 10-cross validation, i.e., ran the computer simulation for 10 different random generations of the three sets and averaged the classification results. This approach provided an unbiased estimate of the future performance of the models. If we had run computer simulation just on training and validation sets only, it is likely that the prediction results would have been too optimistic [11].

5. Discussion

Simulation results summarized in Table 2 show that the memory-based reasoning method and the ensemble model produce the best average overall correct classification rates equal to 75.6% and 74.3%, respectively. The logistic regression and neural network models appear to be the best in correctly classifying low risk jobs. Their actual performance rates are 80.0% and 78.4%, respectively. Decision tree and memory-based reasoning models yield the correct classification rates for high risk jobs equal to 79.1% and 78.2%, respectively. The results appear to be an improvement over the results reported in the 1991 NIOSH Guide (Table 2) [34]. In addition, out of the five methods, one can note that the decision tree generates the most stable classification results with standard deviation equal to 2.5, 2.3, and 2.7 for overall, high, and low risk jobs, respectively. The memory-based reasoning comes close second with standard deviations equal to 3.4, 3.2, and 2.8, respectively (Table 2).

Table 2
The average, best, and worst counts and percentage of correct classification rates for the five models and 1991 NIOSH Guide

		Neural Network	Logistic regression	Decision tree	Memory-based reasoning	Ensemble model	1991 NIOSH Guide
Counts							
Average	Overall	51.7	49.4	51.1	52.9	52.0	
	High	22.7	19.8	26.1	25.8	23.9	
	Low	29.0	29.6	25.0	27.1	27.3	
Best	Overall	56	57	55	59	59	
	High	28	29	31	32	30	
	Low	33	33	30	30	30	
Worst	Overall	47	46	47	46	46	
	High	17	16	23	22	20	
	Low	23	27	21	22	21	
Percentages							
Average	Overall	73.9%	70.6%	73.0%	**75.6%**	**74.3%**	
	High	68.8%	60.0%	**79.1%**	**78.2%**	72.4%	**73%**
	Low	**78.4%**	**80.0%**	67.6%	73.2%	73.8%	**55%**
Best	Overall	80.0%	81.4%	78.6%	84.3%	84.3%	
	High	84.8%	87.9%	93.9%	97.0%	90.9%	
	Low	89.2%	89.2%	81.1%	81.1%	81.1%	
Worst	Overall	67.1%	65.7%	67.1%	65.7%	65.7%	
	High	51.5%	48.5%	69.7%	66.7%	60.6%	
	Low	62.2%	73.0%	56.8%	59.5%	56.8%	
Standard	Overall	3.5	3.7	**2.5**	**3.4**	4.0	
Deviation	High	3.6	3.9	**2.3**	**3.2**	3.2	
	Low	2.9	**1.8**	2.7	**2.8**	2.9	

Out of the five methods, however, the results produced by the decision tree are the easiest to interpret since the outcomes are presented in the form of the if-then rules. Table 3 shows that for all 10 different generations, the decision tree algorithm chose the peak moment variable (PMOMENT) as the variable which has the most predictive power (Importance = 1). For 9 out of 10 generations, the classification of jobs in to low and high risk is just based on a single rule which uses the PMOMENT variable only. This rule built on the training data set for generation 1 is:

IF PMOMENT < 9.3
  THEN
    N = 37, HIGH = 10.8%, LOW = 89.2%
  ELSE
    N = 57, HIGH = 70.2%, LOW = 29.8%

For generation 1, the same rule is used to test the performance of the decision tree on the test set yielding 93.9% (31/33 cases) and 62.2% (23/37 cases) correct classification rates for high risk and low risk jobs, respectively. This rule is also depicted in Fig. 1. Figure 2 shows the distribution of the PMOMENT variable and Table 4 presents the simple descriptive statistics for all five independent variables, including PMOMENT. The results from computer simulation show that relatively small values of PMOMENT from within the range [9.3, 29.5] determine the classification boundary between the low and high risk jobs (Table 3). Only in 1 out of 10 generations, i.e., generation 9, the decision tree uses three variables (PMOMENT, peak twist velocity average – PTVAVG, and peak sagittal angle – PSUP) and several rules to classify jobs into high and low risks with respect to LBDs (Table 3).

In many applications, it is not adequate to characterize the performance of a model by three numbers that measure the correct classification rates of high and low risk jobs and overall rates. Cumulative percent

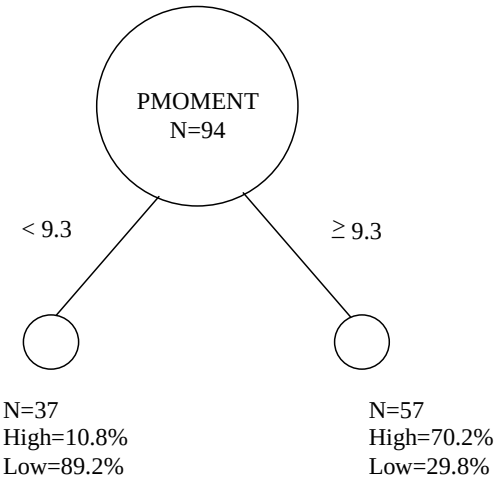

Fig. 1. Decision tree built on the training set for generation 1.

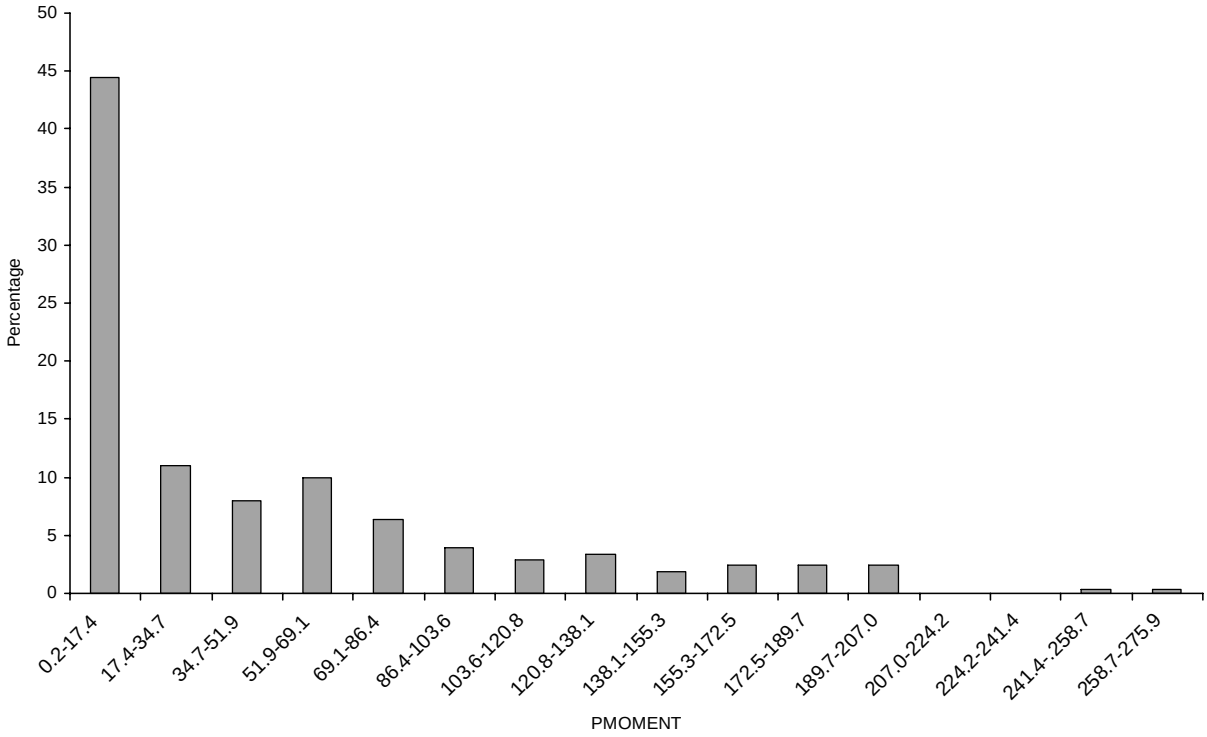

Fig. 2. Distribution of the peak moment (PMOMENT) variable.

response, threshold-based, and receiver operating characteristic (ROC) charts, presented in Fig. 3 through 5 allow one for a more thorough and subtle interpretation of the obtained results.

To properly interpret the cumulative percent response charts, however, let us consider first how these charts are constructed. If the target event is detecting high risk jobs denoted by 1, the response chart sorts the observations in the test set on the basis of their score from the highest probability of response to

Table 3
If-Then rules generated by the decision trees for the training set for 10 different generations of the training, validation, and test data sets. (PMOMENT, PTVAVG, and PSUP stand for peak moment, peak twist velocity average, and peak sagittal angle, respectively.)

Generation #	Number (N) and percentage of cases classified as HIGH/LOW for each generation
1	IF PMOMENT $<$ 9.3 THEN N = 37, HIGH = 10.8%, LOW = 89.2% ELSE N = 57, HIGH = 70.2%, LOW = 29.8%
2	IF PMOMENT $<$ 22.45 THEN N = 46, HIGH = 13.0%, LOW = 87.0% ELSE N = 48, HIGH = 79.2%, LOW = 20.8%
3	IF PMOMENT $<$ 18.5 THEN N = 45, HIGH = 20.0%, LOW = 80.0% ELSE N = 49, HIGH = 71.4%, LOW = 28.6%
4	IF PMOMENT $<$ 17.6 THEN N = 40, HIGH = 17.5%, LOW = 82.5% ELSE N = 54, HIGH = 68.5%, LOW = 31.5%
5	IF PMOMENT $<$ 22.45 THEN N = 48, HIGH = 16.7%, LOW = 83.3% ELSE N = 46, HIGH = 78.3%, LOW = 21.7%
6	IF PMOMENT $<$ 22.85 THEN N = 47, HIGH = 14.9%, LOW = 85.1% ELSE N = 47, HIGH = 78.7%, LOW = 21.3%
7	IF PMOMENT $<$ 10.1 THEN N = 36, HIGH = 8.3%, LOW = 91.7% ELSE N = 58, HIGH = 70.7%, LOW = 29.3%
8	IF PMOMENT $<$ 22.15 THEN N = 48, HIGH = 18.8%, LOW = 81.2% ELSE N = 46, HIGH = 76.1%, LOW = 23.9%
9	Importance of the variables: PMOMENT = 1, PTVAVG = 0.78, PSUP = 0.6 IF PMOMENT $<$ 9.55 THEN N = 37, HIGH = 16.2%, LOW = 83.8% IF PTAVG $>=$ 8.05 AND PMOMENT $>=$ 9.55 THEN N = 24, HIGH = 95.8%, LOW = 4.2% IF PSUP $>=$ 8.6 AND PTVAVG $<$ 8.05 AND PMOMENT $>=$ 9.55 THEN N = 21, HIGH = 66.7%, LOW = 33.3% IF PTVAVG $<$ 1.95 AND PSUP $<$ 8.6 AND PMOMENT $>=$ 9.55 THEN N = 1, HIGH = 100.0%, LOW = 0.0% IF 1.95 $<=$ PTVAVG $<$ 8.05 AND PSUP $<$ 8.6 AND PMOMENT $>=$ 9.55 THEN N = 11, HIGH = 0.0%, LOW = 100.0%
10	IF PMOMENT $<$ 29.5 THEN N = 54, HIGH = 24.1%, LOW = 75.9% ELSE N = 40, HIGH = 77.5%, LOW = 22.5%

the lowest probability of response. The observations are then grouped into ordered bins (deciles), each containing approximately 10% of the data. Then using the dependent variable RISK, the percentage of actual successes in each bin is counted. (A success is defined as detecting a high risk job. The high risk jobs are those whose probability is greater than or equal to a cut-off probability = 0.5.) The more upward and to the left the curve is, the better the strength of the model. To improve interpretation, models' response charts are typically compared with a horizontal line that represents the baseline rate (approximately 47.2%), for which the probability estimates are drawn in the absence of models. In other words, the 47.2% baseline represents the worthless model.

It appears that for the best set of models (generation 1) the neural network appears to be the best if one is to target 10%, 20%, 30% or 40% of the highest risk jobs (Fig. 3). It detects 100% of the high risk jobs

Table 4
Descriptive statistics for the independent variables

Name of variable	Min	Max	Mean	Std Dev.	Skewness	Kurtosis
Lift rate in number of lifts per hour (LIFTR)	5.4	1500.0	140.9	174.0	3.9	20.1
Peak twist velocity average (PTVAVG)	0.7	34.8	7.0	5.3	2.1	6.4
Peak moment (PMOMENT)	0.2	275.9	47.6	55.8	1.5	2.0
Peak sagittal angle (PSUP)	−41.5	99.1	14.3	18.1	0.6	1.2
Peak lateral velocity maximum (PLVMAX)	12	119.9	40.4	17.1	1.3	2.9

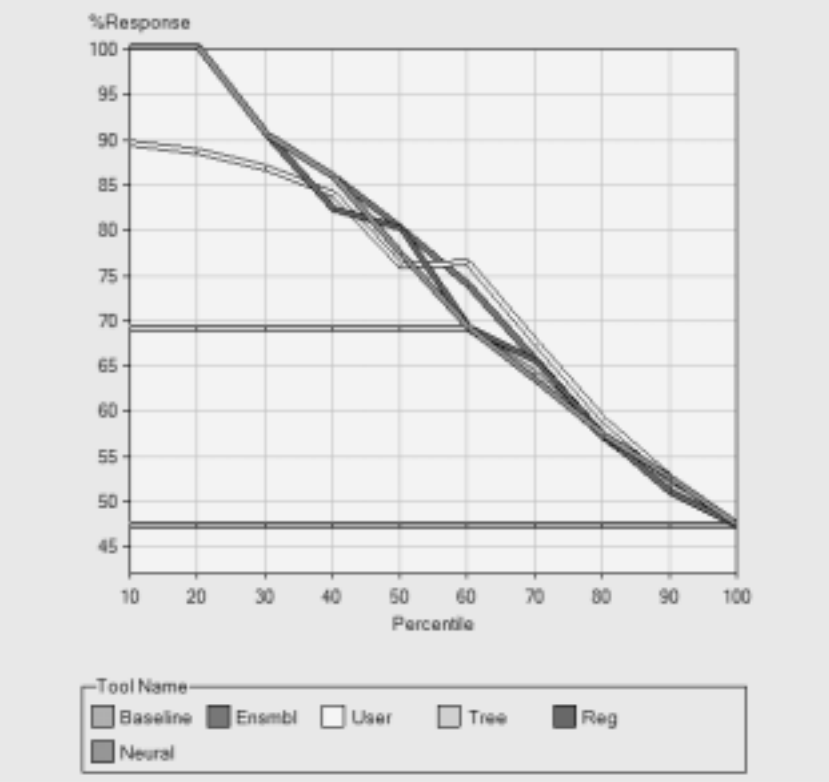

Fig. 3. Percent cumulative response charts for the best set of the five modeling methods. (Legend: Tool name stands for model name. The abbreviations Ensmbl, User, Tree, Reg, and Neural stand for the ensemble model, memory-based reasoning, decision tree, logistic regression, and neural network methods, respectively.).

in the 1st and 2nd deciles. And in the 3rd and 4th deciles, the neural network correctly classifies 90% and 86% of high risk jobs with respect to LBDs, respectively. However, the memory-based reasoning method (denoted in Figs 3–5 as User) with 77% detection ratio seems to be slightly outperforming other models if one is to target 60% of the highest risk jobs. Although the decision tree's results can be easily interpreted and its average detection ratio of high risk jobs is the best (Table 3), out of the 5 models it performs the worst in the 1st through 6th deciles.

Threshold-based charts (Fig. 4) enable one to display the agreement between the predicted and actual target values across a range of threshold levels. The threshold level is the cutoff that is used to classify an observation that is based on the event level posterior probabilities. The default threshold level is 0.50. Figure 4 shows the performance of the decision tree and the other four models for different cut-off points. It confirms the earlier observation that even though the decision tree's average correct classification rate

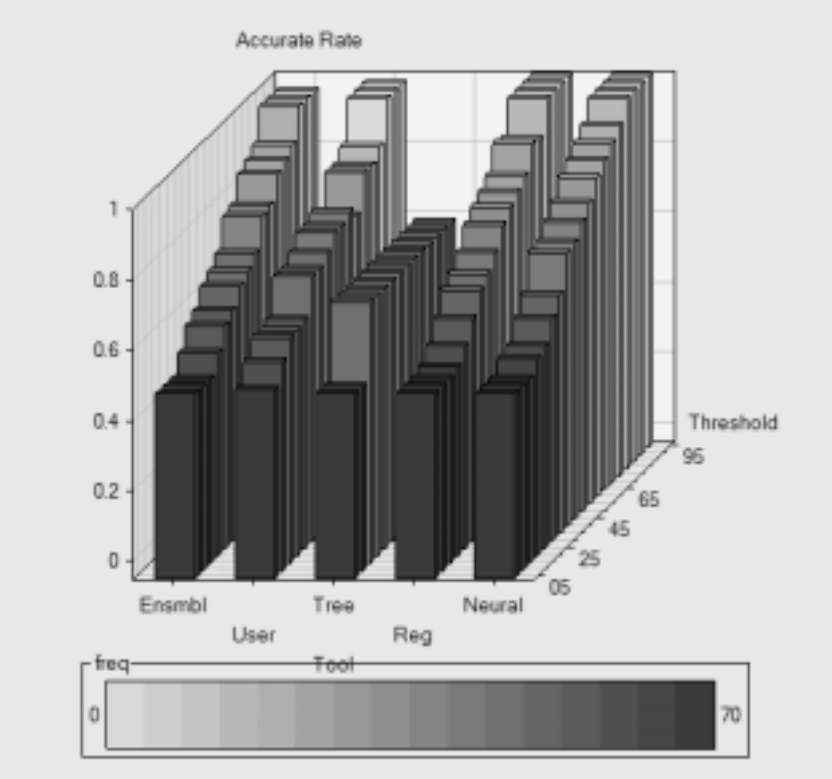

Fig. 4. Threshold-based charts for the best set of the five models. (Legend: Tool stands for model name. The abbreviations Ensmbl, User, Tree, Reg, and Neural stand for the ensemble model, memory-based reasoning, decision tree, logistic regression, and neural network methods, respectively.).

is the best among the five models for the high risk jobs, the decision tree assigns relatively low probability value of about 0.7 to these high risk jobs. In other words, the decision tree determines that the probability that these jobs are indeed high risk jobs is about 0.7. The other 4 models appear to be more reliable, i.e., they determine that some jobs are high risk jobs with the probability of 1 (Fig. 4).

The ROC charts show the trade-off between sensitivity and 1-specificity values (Fig. 5). In classification tasks, two types of error rates are of interest. The false acceptance rate (FAR) called sensitivity and false reject rate (FRR) called specificity. The FAR is the proportion of test cases correctly predicted as high risk jobs to the total number of high risk jobs in the test set. The FRR is the proportion of the test cases representing low risk jobs that are correctly classified as low risk jobs to the total number of low risk jobs in the test set. It is clear that in this study the objective would be to minimize the misclassification of high risk jobs into low risk jobs, i.e., maximize the FAR ratio. A classification model can be tuned by setting an appropriate threshold value to operate at the desired value of FAR. If one tries to decrease the FAR parameter of the model, it would increase the FRR and vice versa. The ROC curve enables one to analyze both errors at the same time. This curve permits one to assess the performance of the model at various operating points (thresholds in a decision process using the available model) and the performance of the model as a whole (using as a parameter the area between the ROC curve and the 45° line). Points in the upper right and lower left parts of the graphs correspond to lower and higher probability cut-offs, respectively. The performance quality of the models is demonstrated by the degree to which the ROC curves push upward and to the left. The curves will always lie above the 45° line.

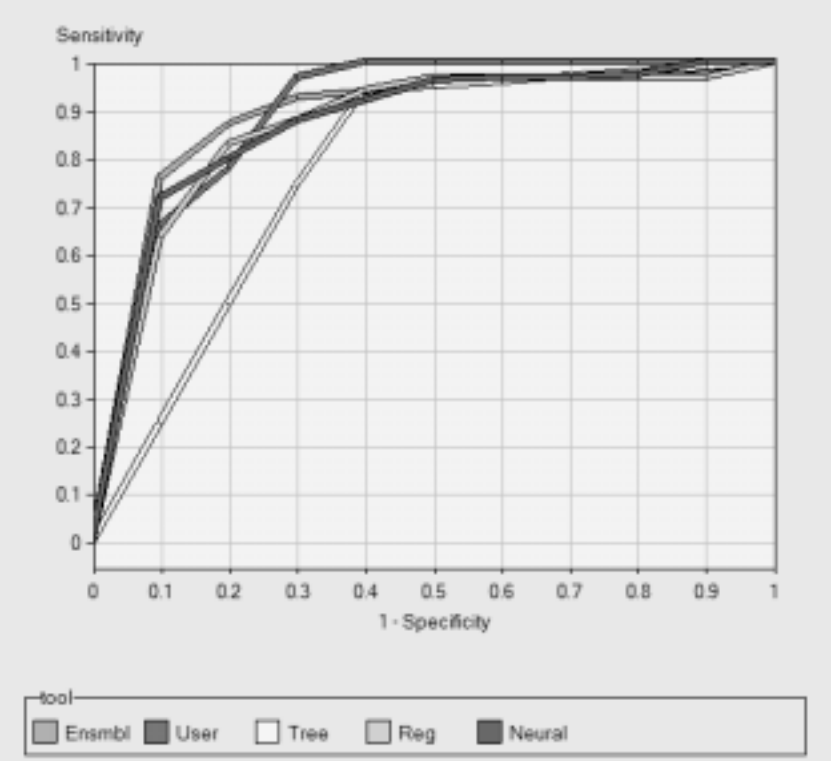

Fig. 5. Receiver operating characteristic (ROC) charts for the best set of the five models. (Legend: Tool stands for model name. The abbreviations Ensmbl, User, Tree, Reg, and Neural stand for the ensemble model, memory-based reasoning, decision tree, logistic regression, and neural network methods, respectively.).

The performance of the models depends on the selection of the operating point. The decision tree performs the worst for the values of 1-specificity $\leqslant 0.4$. For values of 1-specificity < 0.1, there is no clear distinction between the performances of the four remaining models. Overall, the ensemble and memory-based reasoning models appear to be the best because the curves representing these models push upwards and to the left.

6. Other modeling factors

As in our previous study [37], the present project focused only on those workplace risk factors that can be controlled through the engineering design (as recommended by the NIOSH 1981 and 1991 Lifting Guides) [25,34]. However, these modeling efforts do not account for a variety of psychological and psychosocial risk factors of LBDs that are also related to work environment [2,6]. These factors include jobs satisfaction, time pressure, managerial responsibility, or the extent of social support from colleagues and supervisors. For example, workers with LBDs showed a higher frequency of psychological symptoms than those without such disorders [10], and such psychological symptoms were predictive of the future incidence of LBDs [3,4]. However, as pointed out by Riihimaki [29] most of these studies [4,31–33] have been retrospective and it is difficult to determine whether these factors are antecedents or consequences of back problems.

7. Conclusions

The results of this study showed that data mining tools such as neural networks, logistic regression, decision trees, memory-based reasoning, and ensemble model allow for successful classification of lifting jobs into two categories of associated of LBDs risk, based on the available job characteristics data. Computer simulation showed that memory-based reasoning and ensemble models were the best in the overall classification accuracy, the decision tree and memory-based reasoning models were most accurate in classifying high risk jobs, whereas neural networks and logistic regression were the best in classifying low risk jobs. The decision tree model delivered the most stable results across all 10 generations of different data sets randomly chosen for training, validation, and tests. The classification results generated by the decision tree were the easiest to interpret because they were given in the form of simple 'if-then' rules. The results also showed that peak moment had the most predictive power for assessing the risk of LBDs, and in 9 out of 10 generations was the only variable on which such classification was based. The classification results were superior to those published by NIOSH 1991 Guide. The models developed in this study significantly reduce the time consuming job analysis and classification performed by traditional methods. Future work will consider fine-tuning the models as well as collecting a larger sample of data.

Acknowledgement

This study was supported in part by a grant from the Department of Health and Human Services (2002–2006) entitled "Development of a Neuro-Fuzzy System to Predict Spinal Loading as a Function of Multiple Dimensions of Risk" (W.S. Marras, PI).

References

[1] A.A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*, Van Nostrand Reinhold Co., New York, 1990.
[2] M.M. Ayoub, W. Karwowski and P.G. Dempsey, Manual materials handling, in: *Handbook of Human Factors and Ergonomics*, (2nd edition), G. Salvendy, ed., John Wiley, New York, 1996.
[3] F. Biering-Sorensen and C. Thornsen, Medical, social and occupational history as risk indicators for low-back trouble in general, *Spine* **11** (1986), 720–725.
[4] S.J. Bigos, D.M. Suenpler and N.A. Martin, Back injuries in industry: -a retrospective study: II. Injury factors, *Spine* **11** (1986), 246–251.
[5] S.J. Bigos, M.C. Battik, D.M. Spengler, L.D. Fisher, W.E. Fordyce, T.H. Hansson, A.L. Nachemson and M.D. Wortley, A prospective study of work perceptions and psychosocial factors affecting the report of back injury, *Spine* **16** (1991), 1–6.
[6] P.M. Bongers, C.R. deWinter, M.A.J. Kompier and V.H. Hildbrandt, Psychosocial factors at work and musculoskeletal disease, *Scandinavian Journal of Work, Environment and Health* **19** (1993), 297–312.
[7] R. Christensen, *Log-Linear Models and Logistic Regression*, Springer, New York, 1997.
[8] V. Dhar and R. Stein, *Seven Methods for Transforming Corporate Data into Business Intelligence*, Prentice Hall, 1997.
[9] U.S. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, Menlo Park, California, USA, 1996.
[10] J.W. Frymoyer, K.H. Pope, J.H. Clements, D.G. Wilder, B. MacPherson and T. Ashikaga, Risk factors in low back pain: an epidemiological survey, *J. Bone Joint Surg.* **65-A** (1983), 213–218.
[11] P. Giudici, *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley and Sons, Chichester, West Sussex, England, 2003.
[12] M.T. Hagan, H.B. Demuth and M. Beale, *Neural Network Design*, PWS Publishing Company, 1996.
[13] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers: San Francisco, CA, 2001.

[14] D. Jones, *Neural networks for medical diagnosis, Handbook of Neural Computing Applications*, Academic Press, New York, 1991.
[15] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press/Wiley, 2003.
[16] W. Karwowski, K. Ostaszewski and J. Zurada, Applications of catastrophe theory in modeling the risk of low back injury in manual lifting tasks, *Le Travail Humain* **55** (1992), 259–275.
[17] W. Karwowski, J. Zurada, W.S. Marras and P. Gaddie, A prototype of the artificial neural network-based system for classification of industrial jobs with respect to risk of low back disorders, in: *Proceedings of the Industrial Ergonomics and Safety Conference*, F. Aghazadeh, ed., Taylor and Francis, London, 1994, pp. 19–22.
[18] Liberty Mutual Workplace Safety Index of Leading Occupational Injuries, 2004.
[19] P. Mangiameli, D. West and R. Rampal, Model selection for medical diagnosis decision support systems, *Decision Support Systems* **36**(3) (2004), 247–259.
[20] W.S. Marras, Toward an understanding of dynamic variables in ergonomics, *Occupational Medicine: State of the Art Reviews I* (1992), 655–677.
[21] W.S. Marras, S.A. Lavender, S. Leurgans, L.R. Sudhakar, W.G. Allread, F. Fathallah and S. Ferguson, The role of dynamic three-dimensional trunk motion in occupationally-related low back disorders, *Spine* **18** (1993), 617–628.
[22] W.S. Marras, L. Fine, S. Ferguson and T. Waters, The effectiveness of two types of common lifting measures for the control of low back disorders in industries, *Ergonomics* **42**(1) (1999), 229–245.
[23] T.M. Mitchell, *Machine Learning*, WCB/McGraw-Hill, Boston, Massachusetts, 1997.
[24] National Safety Council, *Accident Statistics Chicago*, Illinois, 1990.
[25] NIOSH, *Work practices guide for manual lifting*, DHHS (NIOSH) Pub. No. 81–122. Cincinnati, OH, US Department of Health and Human Services, 1981.
[26] National Institute for Occupational Safety and Health, *Musculoskeletal Disorders and Workplace Factors*, Cincinnati, Ohio: US Department of Health and Human Services, DHHS Publication No. 97–141, 1997.
[27] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, San Mateo, California, 1993.
[28] A. Radding, *Unpacking the mystery of the black box*, Software Magazine, December, 1997, pp. S8–S9.
[29] H. Riihimaki, Low-back pain, its origin and risk indicators, *Scandinavian Journal of Work, Environment and Health* **11** (1991), 81–90.
[30] SAS Institute, *Data Mining Using Enterprise Miner Software: A Case Study Approach*, 1st edition, http://www.sas.com.
[31] D.M.J. Spengler, S.J. Bigos, N.A. Martin, J. Zeh, L. Fisher and A. Nachemson, Back injuries in industry: a retrospective study, *Spine* **11** (1986), 241–256.
[32] H.-O. Svensson and G.B.J. Andersson, Low-back pain in forty to forty-seven year old men: work history and work environment, *Stand. J. Rehabil. Med.* **8** (1983), 272–276.
[33] H.-O. Svensson and G.B.J. Andersson, The relationship of low-back pain, work history, work environment and stress: a retrospective cross-sectional study of 38-64-year old women, *Spine* **14** (1989), 517–522.
[34] T.R. Waters, V. Putz-Anderson, A. Garg and L.J. Fine, Revised NIOSH equation for the design and evaluation of manual lifting tasks, *Ergonomics* **36** (1993), 749–776.
[35] T.R. Waters, V. Putz-Anderson and A. Garg, *Application Manual for the Revised NIOSH Lifting Equation US Department of Health and Human Services*, Cincinnati, OH, 1994.
[36] G. Zhang and V. Berardi, Time series forecasting with neural network ensembles: An application for exchange rate prediction, *The Journal of the Operational Research Society* **52**(6) (2001), 652–664.
[37] J. Zurada, W. Karwowski and W.S. Marras, A neural network-based system for classification of industrial jobs with respect to low back disorders due to workplace design, *Applied Ergonomics* **28**(1) (1997), 49–58.