Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER Computers and Mathematics with Applications 51 (2006) 1075–1092
An International Journal
computers & mathematics
with applications
www.elsevier.com/locate/camwa

Data Mining for the Diagnosis of Type II Diabetes from Three-Dimensional Body Surface Anthropometrical Scanning Data

CHAO-TON SU
Department of Industrial Engineering and Engineering Management
National Tsing Hua University, Hsinchu, Taiwan
ctsu@mx.nthu.edu.tw

CHIEN-HSIN YANG
Department of Industrial Engineering and Management
National Chiao Tung University, Hsinchu, Taiwan

KUANG-HUNG HSU
Department of Health Care Management
Chang Gung University, Tao-Yuan, Taiwan

WEN-KO CHIU
Department of Industrial Design
Chang Gung University, Tao-Yuan, Taiwan

(Received February 2005; revised and accepted August 2005)

Abstract—Diabetes mellitus has become a general chronic disease as a result of changes in customary diets. Impaired fasting glucose (IFG) and fasting plasma glucose (FPG) levels are two of the indices which physicians use to diagnose diabetes mellitus. Although this is a fairly accurate approach, the tests are expensive and time consuming. This study attempts to construct a prediction model for Type II diabetes using anthropometrical body surface scanning data. Four data mining approaches, including backpropagation neural network, decision tree, logistic regression, and rough set, were used to select the relevant features from the data to predict diabetes. Accuracy of classification was evaluated for these approaches. The result showed that volume of trunk, left thigh circumference, right thigh circumference, waist circumference, volume of right leg, and subjects' age were associated with the condition of diabetes. The accuracy of the classification of decision tree and rough set was found to be superior to that of logistic regression and backpropagation neural network. Several rules were then extracted based on the anthropometrical data using decision tree. The result of implementing this method is not only useful for the physician as a tool for diagnosing diabetes, but it is sophisticated enough to be used in the practice of preventive medicine.

Keywords—Data mining, Type II diabetes, Backpropagation neural network, Diagnosis.

This research was supported in part by Grant No. 93-2213-E-007-110 from National Science Council (Taiwan). The authors would like to thank the medical staff at the Department of Health Examination at Chang Gung Memorial Hospital, Taiwan. We are grateful to the Chang Gung Biomedical Research Team members who provided the data and shared their experiences of medical research with us.

0898-1221/06/$ - see front matter Typeset by AMS-TEX
doi:10.1016/j.camwa.2005.08.034

1. INTRODUCTION

Diabetes mellitus (DM) is a major chronic disease, affecting up to 3% of the population in industrialized countries. There are approximately 135 million people suffering from DM, and the number will rise to 300 million, or 5.4% of world population by 2025. Consequently, researchers all over the world are now paying more attention to the diagnosing and/or predicting of DM.

According to the definition from the Canadian Diabetes Association [1], diabetes mellitus is a condition in which the body either cannot produce insulin or cannot effectively use the insulin it produces. Diabetes mellitus is divided into two types: Type I diabetes and Type II diabetes. Type I diabetes (or insulin-dependent diabetes, IDDM) occurs when the pancreas no longer produces any or very little insulin. The body needs insulin to use sugar as an energy source. It usually develops in childhood or adolescence and affects 10% of people with diabetes. Different from Type I, Type II diabetes (or non-insulin-dependent diabetes, NIDDM) occurs when the pancreas does not produce enough insulin to meet the body's needs or the insulin is not metabolized effectively. Type II usually occurs later in life and affects 90% of people with diabetes.

In the past a statistical approach, such as analysis of variable (ANOVA), multi variable analysis, and factor analysis, was used to predict DM. For instance, Kim *et al.* [2] investigated the association between microalbuminuria and the insulin resistance syndrome, independent of Type II diabetes, using a multiple regression analysis and multiple logistic regression analysis. The result shows that the body mass index (BMI) and waist hip ratio (WHR) are both important factors for DM. Chen *et al.* [3] studied the association of hypertension and insulin-related metabolic syndrome in nondiabetic Chinese using factor analysis. The result of their study shows a significant association between hypertension and the insulin-related metabolic syndrome. However, a simple statistical approach such as logistic regression cannot clearly explain the relationship among the input variables and DM. On the other hand, artificial intelligence (AI) could be a good candidate to avoid this problem. Since the early 1990s, feedforward artificial neural networks have been used increasingly in various fields, such as backpropagation for clinical diagnosis [4–6] and self-organizing map breast cancer clustering [7]. Other algorithms, like genetic algorithms, genetic programming, evolution strategies, evolutionary programming, classifier systems, and hybrid systems are being reported continuously [8]. In addition, two indices, sensitivity and specificity, are used to evaluate the prediction models when conducting epidemiological studies using the statistical method [9–11]. In practice, we can understand that researchers do not use statistics-based analysis due to the fact it may face some limitations. However, AI studies put emphasis on the accuracy of classification rather than on sensitivity and specificity. Although biochemical examination is a general approach for the diagnosis of diabetes, it has some disadvantages as a diagnostic for DM. Repeated diagnoses can lead to increased inconvenience for both the physician and the patient. On the other hand, some studies [12,13] have shown that there is a relationship between body composition and DM. Based on this we use four data mining approaches to find the relationship of anthropometrical data and diabetes mellitus. It is our intention to provide a new diagnostic approach for physicians to diagnose DM, and thereby reduce government expenditures and enhance the health for all citizens.

The remainder of this paper is organized as follows. In Section 2, we introduce the four data mining techniques and their procedures. Methods and results are presented in Section 3. Technical and medical discussions are provided in Section 4. Finally, we draw our conclusion and make suggestion from this study in Section 5.

2. SELECTED DATA MINING APPROACHES

Diagnosis is the process of selectively gathering information concerning the health status of a patient, and interpreting this information based on previous knowledge, as evidence for or against the presence or absence of a disorder [14]. Feature selection has always been one of the

processes for diagnosis and prognosis [15]. Some tools, like neural network and decision tree, are helpful for analyzing the results of feature selection. Logistic regression is also a traditional tool in many medical researches, including the process of feature selection. As for rough set, it is good at processing large and vague data where the character of the data is consistent with general medical data.

2.1. Neural Network

Neural network is one of the methods of artificial intelligence. It is characterized by

(1) its pattern of connections between the neurons,
(2) its method of determining the weights of these connections, and
(3) its activation function.

A neural net consists of a large number of simple processing elements called neurons. Similar to neural systems, each neuron is connected to other neurons by means of directed communication links, each with an associated weight, with the weights representing the level of information. Each neuron has an internal state, called its activation, which is a function of the inputs it has received. Typically, a neuron sends its activation as a signal to several other neurons. It is important to note that a neuron can send only one signal at a time, although that signal is broadcast to several other neurons. In addition, it is convenient to visualize neurons as being arranged in layers. Typically, neurons in the same layer behave in the same manner. A multilayer net generally is composed of one input layer, hidden layers, and an output layer [16]. Usually, a neural network with signal hidden layer can provide a good performance of classification and prediction.

Neural network systems are divided into two groups: supervised learning and unsupervised learning. Backpropagation neural network is a typical network of supervised learning, and very useful for selecting features. The multilayer network modeling is accomplished via two phases: the training and the testing process. Even though it can basically approximate any function, the neural network method still has a few problems such as time consuming convergence, overfitted training, high complexity in computation and black boxes in the training results [17]. Some developed algorithms [18,19] reported that a suitable pruning of some of the input nodes might be helpful for rule extracting and knowledge acquisition.

Consequently, we refer to Su *et al.* [20] concerning an algorithm of feature selection. As per their research, a neural network is shown in Figure 1. This neural network is a multilayer perceptron. It is composed of a single input layer with n input nodes, a single hidden layer with m hidden

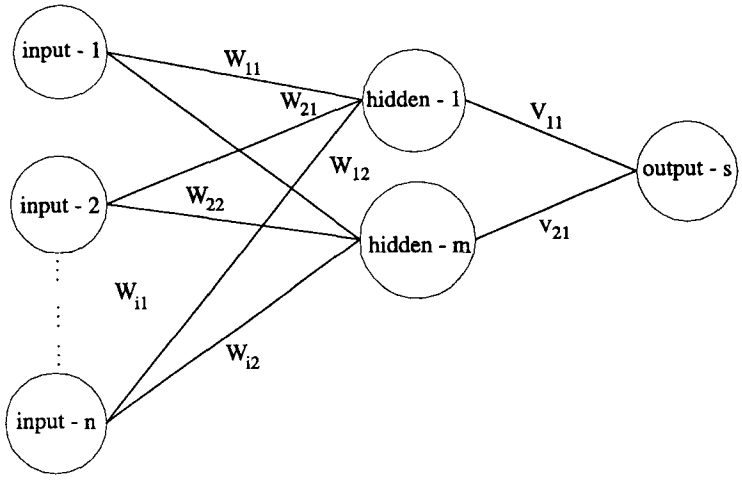

Figure 1. A typical neural network.

nodes, and a single output layer with s output nodes. The connection (weight) of input to hidden is W; another hidden to output is V. The relationship between the two layers is determined by weight. It is important to note that the priorities of these input nodes are according to W and V. As a result of not all of the connections being identical, i.e., sometimes W is greater than V and sometimes it is not. The priority of the input nodes is determined by P_i which is defined as follows,

$$P_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{s} |W_{ij} \times V_{jk}|, \tag{1}$$

where

W_{ij} is the weight between the i^{th} input node and the j^{th} hidden node;
V_{jk} is the weight between the j^{th} hidden node and the k^{th} output node;
P_i is the sum of absolute multiplication values of the weights W_{ij} and V_{jk}.

For the sake of convenience for the user, we calculate the mean of total P_i to determine the important input nodes according to equation (1). Thus, the i^{th} input node is found to be an important node which will be selected if $P_i \geqslant mean$ or else will be removed if $P_i < mean$.

In summary of the above, we illustrate an algorithm for feature selection as follows.

2.1.1. Algorithm

Step 1: Calculate the product (P_i) of the connection of input-hidden and hidden-output for each input node.
Step 2: Sort the products and compute the mean of total P_i.
Step 3: Remove input node if its product (P_i) is less than the value of the mean of total P_i.
Step 4: Go to Step 1 till the number of input nodes that are users is as expected.

2.2. Decision Tree

A decision tree is another feature selection approach. It is a popular classifier in machine learning applications and is also used as a diagnostic model in medicine. Decision tree is connected via nodes and branches. The tree construction process is heuristically guided by choosing the 'most informative' attribute at each step, aimed at minimizing the expected number of tests needed for classification. Let E be the entire initial set of training examples, and $c_1, \ldots, c_N$ be the decision classes. A decision tree is constructed by repeatedly calling a tree construction algorithm in each generated node of the tree. Tree construction stops when all examples in a node are of the same class, or if some other stopping criteria are satisfied. In brief, a decision tree is a flow-chart-like tree structure, in which each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or a class distribution. The topmost node in a tree is the root node [21]. A typical decision tree is shown in Figure 2.

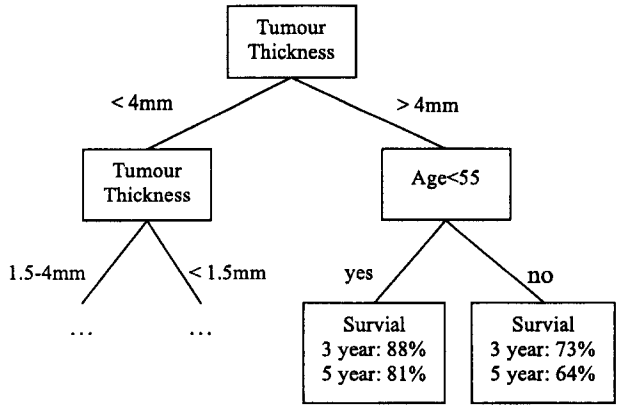

Figure 2. A decision tree for predicting survival.

C4.5 is a well-known decision tree construction software (C5.0 is its recent upgrade), that is widely used, and has been incorporated into medical data mining tools. Quinlan [22] provides ID3 (interactive dichotomizer 3) using a tree representation. It is an entropy-based algorithm used for some of the larger database analysis either consisting of string data or integer data. Also, its interpretability has been maximized. It is important to note that C4.5 is more interpretable than neural networks. C4.5 is easy to use for users with little or no knowledge of statistics and machine learning. Compared with C4.5, the C5 version classifies more accurately, much faster and requires less memory. Although there are all these advantages, overfitting remains an important issue to overcome for the knowledge miner.

CART is another decision tree popular for data mining. Different from C4.5, CART is a binary tree based on the Gini Index (GI) to determine the condition for constructing the tree [23]. Both CART and C4.5 need to prune the initial tree after training and testing. The difference in pruning conditions between these two algorithms is that C4.5 is based on the subtree, while CART is based on the entire tree. In addition, both CART and C4.5 rely on the specific 'cost function' to decrease the probability of misclassification [9].

In this study, entropy-based trees [24] have been chosen to analyze and induct this medical diagnostic tree. Some of the medical information such as the symptom of some diseases, or the classification of body size are vague and difficult to distinguish in a clinical diagnosis. It is worth noting here that CART is not suited for this study because it is an absolute binary classifier. In short, the entropy-based tree using C5 with not only a friendly interface, but also using a flow-chart-like structure makes it more user-friendly. The procedure, i.e., hypothesis and algorithms of entropy-based tree is as follows.

HYPOTHESIS. *Let's take a training set* S. $C_i \in S$, $\forall i = 1, 2, 3, \ldots, n$. *The number of class is* $\text{freq}(C_i, S)$. $|S|$ *is the total number of training sets. Hence, the probability of occurrence of the number of class is* $(\text{freq}(C_i, S))/|S|$.

ALGORITHMS.

Step 1: Measure the information $(-\log_2(\text{freq}(C_i, S)/|S|))$ of each class.

Step 2: Calculate the mean information of training set S.

$$\text{info}(S) = -\sum_{i=1}^{n} \frac{\text{freq}(C_i, S)}{|S|} \log_2 \left(\frac{\text{freq}(C_i, S)}{|S|} \right) \tag{2}$$

Step 3: Partition S into S_i base on attribute A, i.e., let $\{S_1, S_2, S_3, \ldots, S_n\} \in S$.

Step 4: Calculate the information which is partitioned.

$$\text{info}_A(S) = \sum_{i=1}^{n} \frac{|S_i|}{S} \times \text{info}(S_i) \tag{3}$$

Step 5: Compute the information gain.

$$\text{gain}(A) = \text{info}(S) - \text{info}_A(S) \tag{4}$$

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

2.3. Logistic Regression

Logistic regression is a typical statistical approach which is good at binary data analysis. For example, some medical research, as a result of output such as survivals (yes or no) and disease

(positive or negative) are categorical data, where logistic regression is a useful analysis approach. Also accuracy of classification could satisfy some researchers and clinical operators. Different from traditional simple regression, it is a nonlinear approach to analyze the dependent variable that is categorical, such as binary data.

To obtain the logistic model from the logistic function, we write z as the linear sum α plus β_1 times X_1 plus β_2 times X_2, and so on to β_k times X_k, where the X are independent variables of interest and α and β_i are constant terms representing unknown parameters. In essence, then, z is an index that combines the $X's$ (see equation (5)).

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k. \tag{5}$$

We now substitute the linear sum expression for z in the right-hand side of the formula for $f(z)$ to get the expression $f(z)$ equals 1 over 1 plus e to minus the quantity α plus the sum of $\beta_i X_i$ for i ranging from 1 to k (see equation (6)),

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\alpha + \Sigma \beta_i X_i)}}. \tag{6}$$

Actually, to view this expression as mathematical model, we must place it in an epidemiologic context. Suppose we have observed independent variables X_1, X_2, and so on up to X_k on group of subjects, for whom we have also determined disease status, as either 1 if "with disease" or 0 if "without disease".

We with to use this information to describe the probability that the disease will develop, in a disease-free individual with independent variable values X_1, X_2, up to X_k. The probability being modeled can be denoted by the conditional probability statement as follows,

$$P(D = 1 \mid X_1, X_2, \ldots, X_k).$$

The model is defined as logistic if the expression for the probability of developing the disease, given the X is 1 over 1 plus e to minus the quantity α plus the sum from i equals 1 to k of β_i times X_i. The terms α and β_i in this model represent unknown parameters that we nee to estimate based on data obtained on the $X's$ and on D (disease outcome) for a group of subjects.

For notational convenience, we denote the probability statement $P(D = 1 \mid X_1, X_2, \ldots, X_k)$ as simply $\mathbf{P}(\mathbf{X})$ where the bold $\mathbf{X}$ is a shortcut notation for the collection of variables X_1 through X_k,

$$\mathbf{P}(\mathbf{X}) = P(D = 1 \mid X_1, X_2, \ldots, X_k). \tag{7}$$

Thus, the logistic model [25] may be written as equation (8),

$$\mathbf{P}(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \Sigma \beta_i X_i)}}. \tag{8}$$

2.4. Rough Set

The rough set theory was proposed by Pawlak in 1982, and provides a mathematical tool for representing and reasoning about vagueness and uncertainty. The notion of *indiscernibility* plays an important role in this theory. Rough set theory is good at data deduction, i.e., elimination of superfluous data, discovery of data dependencies, estimation of the significance of data, and the discovering of cause-effect relationships. Based on the above, rough set provides a powerful function for physicians and in medical studies as a diagnosis tool for some diseases.

The rough set theory has many important advantages [26] which are as follows,

(1) provides efficient algorithms for finding hidden information in data,
(2) finds minimal sets of data,
(3) evaluates the significance of data,
(4) generates minimal sets of decision rules from data,
(5) easy to understand, and
(6) offers straightforward interpretation of obtained results.

According to the rough set theory, we should find the indiscernibility. Formally, let U be a set of training objects, A be a set of attributes describing the objects, C be a set of classes and V_j be a value domain of an attribute A_j. $v_j^{(i)}$ represents the value of attribute A_j for the i^{th} object $\text{Obj}^{(i)}$. $\text{Obj}^{(i)}$ and $\text{Obj}^{(k)}$ are said to have an indiscernibility relation with attribute A_j while $\text{Obj}^{(i)}$ and $\text{Obj}^{(k)}$ have the same value of attribute A_j. Also, if $\text{Obj}^{(i)}$ and $\text{Obj}^{(k)}$ have the same values for each attribute in subset B of A, $\text{Obj}^{(i)}$ and $\text{Obj}^{(k)}$ are also said to have an indiscernibility relation with attribute set B.

The lower approximation and upper approximation are defined as $B_*(X)$ and $B^*(X)$ respectively, as follows,

$$B_*(X) = \{x \mid x \in U,\ B(x) \subseteq X\}, \tag{9}$$

$$B^*(X) = \{x \mid x \in U,\ \text{and } B(x) \cap X \neq \phi\}. \tag{10}$$

After the lower and the upper approximation have been found, the rough set theory can be used to derive both certain and uncertain information, and induce certain and possible rules from them.

2.5. Accuracy

Three accuracy indices are used to evaluate medical models. For data mining researches, accuracy of classification is often used. The other two indices, 'sensitivity' and 'specificity' are always used in epidemiological studies. For a two class problem the accuracy of classification can be estimated as $P/(P+N)$ or $(P+1)/(P+N+2)$ where P is the number of positive examples and N is the number of negative examples of the selected class. However, it is practical for four subsets to be considered.

True positives (TP): true positive answers of a classifier denote the correct classification of positive cases.

True negatives (TN): true negative answers denote the correct classification of negative cases.

False positives (FP): false positive answers denote the incorrect classification of negative cases into a class positive.

False negatives (FN): false negative answers denote the incorrect classification of positive cases into a class negative.

According to the above definitions, the classification accuracy measures the proportion of correctly classified cases as follows,

$$\text{Accuracy of Classification} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \tag{11}$$

Regarding the other indices, sensitivity measures the fraction of positive cases that are classified as positive, and specificity measures the fraction of negative cases classified as negative. In other words, sensitivity can be viewed as a detection rate that one wants to maximize, while specificity can be seen as a false alarm rate which one wants to maximize.

3. IMPLEMENTATION

3.1. Equipments and Materials

3.1.1. Three-dimension whole body scanner

A three-dimension whole body scanner system (Figure 3) with six 3D sensor heads mounted on three vertical scanning mechanisms that could be motion synchronized, was used in this study. Based on optical triangulation techniques, including a CCD (charge coupled devices) image plane

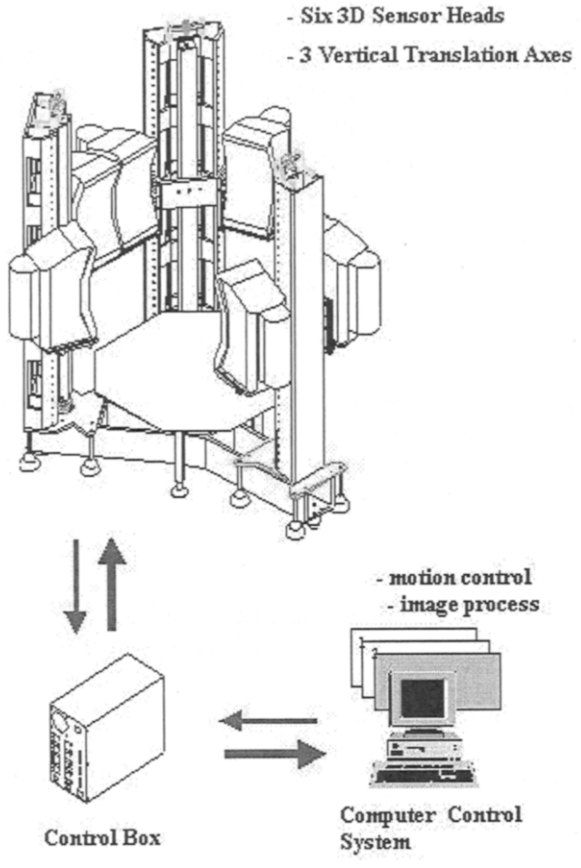

Figure 3. A three-dimension whole study system.

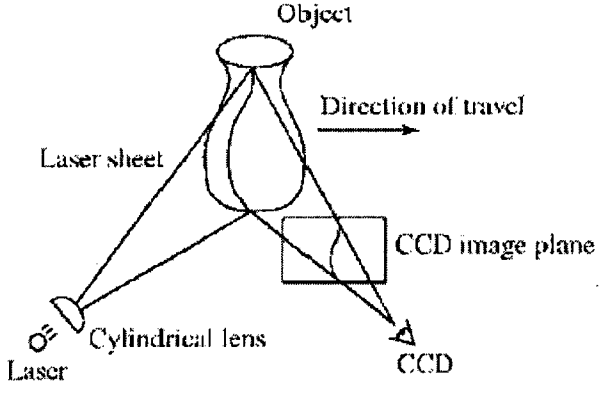

Figure 4. The optical triangulation techniques of whole body scanner.

(a 768×492 pixel array and laser sheet), six lateral laser projections were formed (Figures 4 and 5). Afterwards these six projections needed to be merged. A total of about 280 measurement results were collected from the scan data within 24 seconds. In order to ensure accuracy of measurement, the subjects were asked to extend their arms 30° out from their bodies. If there was any fault in this check, then the subjects were remeasured.

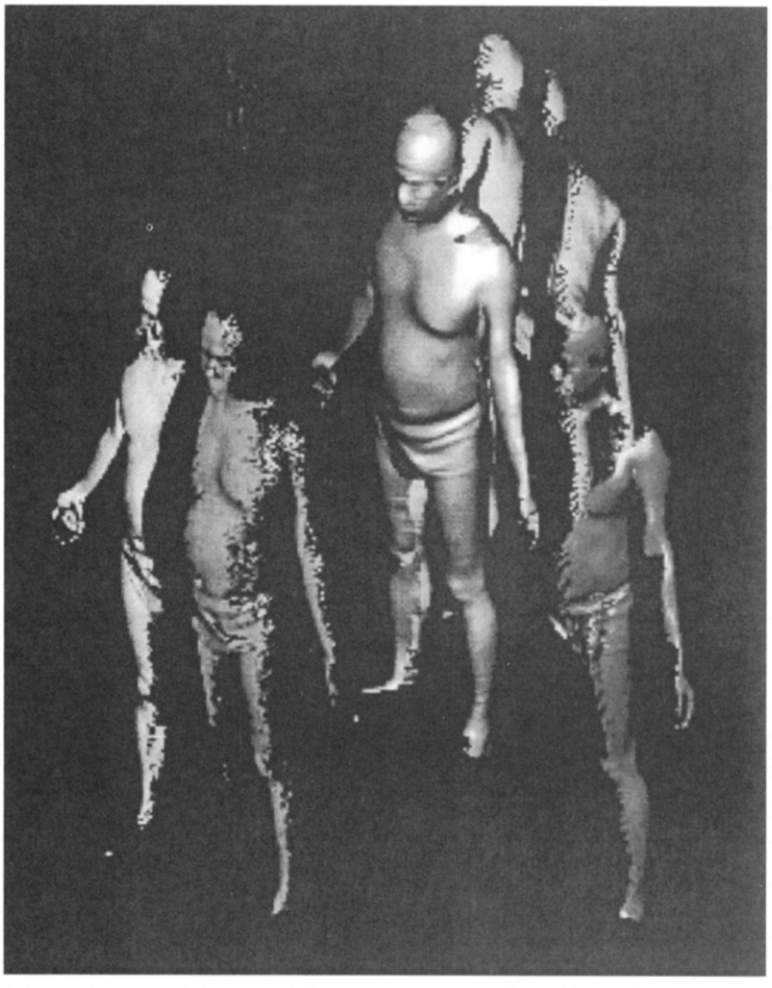

Figure 5. Six laser lateral projectors and a merged whole body.

3.1.2. Subjects

A total of 7020 subjects (3435 men and 3585 women) were recruited via the Department of Health Examination from those seeking an annual physical health check-up at Chang Gung Memorial Hospital in Tao-Yuan, Taiwan. Thirty-two anthropometrical data were measured by the whole body scanner. These data included: height, weight, head circumference, breast circumference, waist circumference, hip circumference, left upper arm circumference, right upper arm circumference, left fore arm circumference, right fore arm circumference, right thigh circumference, left thigh circumference, right leg circumference, left leg circumference, breast width, waist width, hip width, breast profile area, hip profile area, volume of head, surface area of head, volume of trunk, surface area of trunk, volume of left arm, surface area of left arm, volume of right arm, surface area of right arm, volume of left leg, surface of left leg, volume of right leg, surface area of right leg. In addition to these measurements, the subjects' age and gender were collected as well.

3.2. Data Preprocessing

Some of this anthropometrical data tended to be incomplete and inconsistent, so we needed to perform data cleaning prior to implementation. Data cleaning tasks were carried out as follows.

(1) Missing value. We ignored some missing tuples as a result of that occupies a few proportions of all anthropometrical data.

(2) Noisy data. Some repeated data (e.g., repeated key-in by operator) was deleted. The data with key-in error was also deleted.
(3) Inconsistent data. Some tuples were without a record of biochemical examination, but were in the anthropometrical database. We deleted it as well.

After data preprocessing, 6023 subjects (2947 men and 3076 women) were retained. The summary of these subjects is shown in Table 1. Four approaches, neural network, decision tree, logistic regression and rough sets analysis were performed.

Table 1. Comparison of feature of all the included subjects with non-DM.

Variable (Abbreviation)	General Population*	Diabetes*	Non-Diabetes*
Age [years]	53.4 ± 12.1	59.9 ± 10.1	52.6 ± 12.1
Height [cm]	159.4 ± 8.3	159.4 ± 8.2	159.4 ± 8.3
Weight [kg]	63.4 ± 11	66.4 ± 12.0	63.0 ± 10.8
Head Circumference [cm] (HEAD_CIR)	58.8 ± 2	58.5 ± 2.3	58.8 ± 2.0
Breast Circumference [cm] (BRAS_Cir)	97.7 ± 11.4	101.0 ± 11.6	97.3 ± 11.4
Waist Circumference [cm] (WAIST_CI)	86.6 ± 10.9	92.2 ± 10.7	85.9 ± 10.8
Hip Circumference [cm] (HIP_CIRC)	96.9 ± 6.8	98.0 ± 6.9	96.8 ± 6.8
Left Upper Arm Circumference [cm] (LEFT_UAR)	30.1 ± 4.1	30.5 ± 4.4	30.0 ± 4.0
Right Upper Arm Circumference [cm] (RIGHT_UA)	30.3 ± 3.9	30.7 ± 4.3	30.2 ± 3.9
Left Fore Arm Circumference [cm] (LEFT_FAR)	23.9 ± 2.7	24.3 ± 2.8	23.8 ± 2.7
Right Fore Arm Circumference [cm] (RIGHT_FA)	24.3 ± 2.6	24.7 ± 2.7	24.3 ± 2.6
Right Thigh Circumference [cm] (RIGHT_TH)	52.1 ± 4.7	50.6 ± 5.3	52.3 ± 4.6
Left Thigh Circumference [cm] (LEFT_THI)	51.9 ± 4.7	50.5 ± 5.4	52.1 ± 4.6
Right Leg Circumference [cm] (RIGHT_LE)	33.3 ± 3.1	33.2 ± 3.4	33.3 ± 3.0
Left Leg Circumference [cm] (LEFT_LEG)	33.4 ± 3.1	33.4 ± 3.4	33.4 ± 3.0
Breast Width [cm] (BRAS_WID)	31.4 ± 2.5	32.2 ± 2.6	31.2 ± 2.5
Waist Width [cm] (WAIST_WI)	29.8 ± 3.1	31.1 ± 3	29.6 ± 3.1

Table 1. (cont.)

Variable (Abbreviation)	General Population*	Diabetes*	Non-Diabetes*
Hip Width [cm] (HIP_WIDT)	34.6 ± 2.2	34.8 ± 2.3	34.6 ± 2.2
Breast Profile Area [cm^2] (BRAS_PRO)	663.8 ± 114.1	718.5 ± 118.2	656.8 ± 111.8
Waist Profile Area [cm^2] (WAIST_PR)	566.5 ± 141.9	646.3 ± 149.2	556.5 ± 137.8
Hip Profile Area [cm^2] (HIP_PROF)	688.9 ± 98.7	706.1 ± 110.3	686.6 ± 97.1
Volume of Head [cm^3] (HEAD_VOL)	4746.5 ± 470.0	4765.6 ± 505.0	4743.5 ± 465.1
Surface Area of Head [cm^2] (HEAD_SUR)	1256.6 ± 103.1	1257.3 ± 110.0	1256.4 ± 102.1
Volume of Trunk [cm^3] (TRUNK_VO)	39264.6 ± 7788.9	42502.1 ± 8320.5	38853.8 ± 7626.7
Surface Area of Trunk [cm^2] (TRUNK_SU)	6278.2 ± 830.2	6548.7 ± 863.4	6243.5 ± 819.1
Volume of Left Arm [cm^3] (LEFT_ARM)	2200.5 ± 451.3	2262.3 ± 463.1	2193.0 ± 450.0
Surface Area of Left Arm [cm^2] (LT_AR_S)	1312.2 ± 168.7	1338.7 ± 166.8	1309.0 ± 168.7
Volume of Right Arm [cm^3] (RIGHT_AR)	2321.3 ± 460.5	2388.9 ± 478.4	2312.9 ± 457.8
Surface of Right Arm [cm^2] (RT_AR_SR)	1347.5 ± 168.2	1373.3 ± 164.4	1344.1 ± 168.4
Volume of Left Leg [cm^3] (LT_LEG_V)	6030.0 ± 1119.6	5772.6 ± 1173.2	6061.5 ± 1108.3
Surface of Left Leg [cm^2] (LT_LEG_S)	2094.5 ± 250.2	2044.0 ± 252.5	2100.8 ± 249.2
Volume of Right Leg [cm^3] (RT_LEG_V)	5996.7 ± 1100.4	5722.9 ± 1169.6	6030.6 ± 1085.2
Surface of Right Leg [cm^2] (RT_LEG_S)	2087.2 ± 252.6	2033.7 ± 255.7	2093.8 ± 251.3

3.3. Implementation Results

A total of 6000 data sets were selected randomly from the original database via data preprocessing. They were divided into two groups: 80% of the cases were the training sets and the others were the testing sets, i.e., the training sets were 4800 tuples and the testing sets were 1200 tuples.

3.3.1. Neural network

All of the anthropometrical data as well as the subjects' age and gender are the input nodes. One output node represents if the subject suffer from DM. So, the structure of this neural network could be expressed as 34-X-1 where X denotes the number of hidden nodes. In this study, Professional II Plus software [27] was used to perform the computation to obtain the structure with the maximum classification rate. In this multilayer neural network, nodes from the hidden layer, from 1 to 30, were chosen. The other parameters, like momentum were set at

Table 2. Results of neural network training (original model).

Structure	Training Accuracy	Testing Accuracy
34-2-1	81.24%	80.46%
34-3-1	81.23%	80.44%
34-4-1	81.27%	80.40%
34-5-1	81.29%	80.41%
34-6-1	81.30%	80.42%
⋮	⋮	⋮
34-23-1	81.38%	80.47%
⋮	⋮	⋮
34-28-1	81.27%	80.36%
34-29-1	81.26%	80.36%
34-30-1	81.26%	80.35%

Table 3. Results of neural network training (reduced model).

Structure	Training Accuracy	Testing Accuracy
12-23-1	80.68%	80.15%

0.9, 0.8, and 0.7, the learning rate was set at 0.1, 0.2, and 0.3, and the number of iterations was set at 20,000. After trial and error, the accuracy of the classification of the training set and the testing set are shown in Table 2. The result shows that the structure 34-23-1 provides the better performance when the learning rate is 0.1 and the momentum is 0.9. Next, the network is pruned. Based on equation 1, the mean of P_i is 1.86. The input nodes with $P_i < 1.86$ are removed from the network. After that, twelve anthropometrical factors (subjects' age, waist profile area, right thigh circumference, breast profile area, left thigh circumference, volume of trunk, volume of left leg, waist circumference, volume of right leg, waist width, head circumference, breast width) were determined. Based on these twelve factors (see Table 3), we found that the performance of the reduced model was similar to the original neural network model.

3.3.2. Decision tree

A decision tree with 167 branches was inducted from the whole of the anthropometrical data (6000 instances, 34 attributes) by See 5 software. In See 5, we place training sets with 4800 tuples and testing sets with 1200 tuples to construct the decision tree (a medical diagnostic tree). Furthermore, cost files were defined and used to reduce the probability of misclassification from positive to negative. The original medical diagnostic tree shows that the proportion of misclassification is approximate 9.3%. The criterion of feature selection is to collect all of the nodes that form on each layer of the medical decision trees from all the folds. As a result of some nodes having been repeated, a total of thirteen anthropometrical factors (height, weight, breast circumference, waist circumference, left upper arm circumference, right thigh circumference, left circumference, breast profile area, hip profile area, volume of trunk, surface area of left arm, subjects' gender and their age) were found from the medical diagnosis tree. Next, a decision tree was inducted from these thirteen attributes, and the proportion of misclassification was raised to 9.6%.

3.3.3. Logistic regression

A logistic regression model was constructed using the SPSS V10.0 software. Being similar as the R square of a simple linear regression, a likelihood ratio test was always used to test the variance and significance of this model. Next, the Lemeshow Test was used to test the goodness of fit of this model. The result showed that the likelihood ratio was 2816 and that the model chi-square was significant. The goodness of fit was satisfied in our model.

Next, 4800 training sets were used to construct an original logistic regression model. The accuracy of classification from the testing sets with 1200 tuples for this model is 88.50%. Then, a total of thirteen anthropometrical factors were selected through logistic regression analysis. These significant factors were weight, head circumference, right thigh circumference, left thigh circumference, breast width, waist width, hip width, waist profile area, surface area of head, volume of trunk, surface area of trunk, volume of right leg, and subjects' age, respectively. Rather than constructing the complete logistic regression model, the reduced logistic regression model is shown as equation (12). The accuracy of classification of the reduced model is 88.57%.

$$\begin{aligned} y = {} & (0.0159 \times \text{AGE}) + (0.0616 \times \text{WEIGHT}) - (0.0796 \times \text{HEAD\_CIR}) \\ & - (0.2371 \times \text{RIGHT\_TH}) + (0.0027 \times \text{BRAS\_WID}) - (0.0672 \times \text{WAIST\_WI}) \\ & + (0.0325 \times \text{HIP\_WIDT}) + (0.0021 \times \text{WAIST\_PR}) - (0.0015 \times \text{HEAD\_SUR}) \\ & + (0.0001 \times \text{TRUNK\_VO}) - (0.0009 \times \text{TRUNK\_SU}) - (0.0002 \times \text{RT\_LEG\_V}) + 10.7055 \end{aligned} \tag{12}$$

In equation (12), we can find the three significant factors are right thigh circumference, head circumference and waist width. The other factors are not significant.

3.3.4. Rough Set

According to the rough set theory, 34 attributes were reduced by the Rosetta GUI version 1.4.41 software package [28]. Two steps, data discretization and computing the minimal reducts needed to be performed. We used the entropy-based algorithm to process data discretization. However, a genetic algorithm (built-in Rosetta) was used to produces a set of minimal attribute subsets (minimal reducts) that define the functional dependencies. According to the algorithms, a total of 12 anthropometrical factors including height, waist circumference, right thigh circumference, left thigh circumference, right leg circumference, left leg circumference, hip width, volume of head, surface area of head, volume of trunk, volume of right leg, and surface area of right leg were used to construct a reduced model to evaluate DM. The accuracy of the classification of the reduced model is approximate to 89%.

Table 4. Summary of feature selection of four approaches.

	Neural Network	Decision Tree	Logistic Regression	Rough Set	Total
SEX		★			1
AGE	★	★	★		**3**
HEIGHT		★		★	2
WEIGHT		★	★		2
HEAD_CIR	★		★		2
BRAS_CIR		★			2
WAIST_CI	★	★		★	**3**
HIP_CIRC					0

Table 4. (cont.)

LEFT_UAR		★			1
RIGHT_UA					0
LEFT_FAR					0
RIGHT_FA					0
RIGHT_TH	★	★	★	★	**4**
LEFT_THI	★	★	★	★	**4**
RIGHT_LE				★	1
LEFT_LEG				★	1
BRAS_WID	★		★		1
WAIST_WI	★		★		2
	Neural Network	Decision Tree	Logistic Regression	Rough Set	Total
HIP_WIDT			★	★	2
BRAS_PRO	★		★		2
WAIST_PR	★		★		2
HIP_PROF		★			1
HEAD_VOL				★	1
HEAD_SUR			★	★	2
TRUNK_VO	★	★	★	★	**4**
TRUNK_SU			★		1
LEFT_ARM		★			1
LT_AR_S					0
RIGHT_AR					0
RT_AR_SR					0
LT_LEG_V	★				1
LT_LEG_S					0
RT_LEG_V	★		★	★	**3**
RT_LEG_S				★	1

Summary of feature selections from four data mining approaches are shown in Table 4. Each approach has its algorithm or function to provide some factors which are significant. Thus, we can calculate the significant number (frequency) of the four approaches based on their functions in the last column. A total frequency as per these four approaches shows that the volume of trunk, right thigh circumference, and left thigh circumference are greater than the other anthropometrical factors. The ones following after that are waist circumference, volume of right leg, and subjects' age.

3.3.5. Rule induction

We choose six important anthropometrical factors including volume of trunk, right thigh circumference, left thigh circumference, waist circumference, volume of right leg, and subjects' age as per Table 4 to perform the rule induction. Again, See 5 was implemented and rules with an accuracy greater than 80% are shown in Table 5. It is interesting to note that one of the rules was described as "if a subjects' age is under 51 years of age, he cannot suffer from DM".

Table 5. Summary of accuracy of classification for rules.

No.	Rule	Diagnosis	Accuracy of Classification
1	TRUNK_VO <= 46184.23 & RT_LEG_V > 3857.71	Negative	90.9%
2	WAIST_CI <= 103.65 & LEFT_THI > 49.67 & TRUNK_VO <= 50366.44	Negative	93.7%
3	AGE <= 51	Negative	94.8%
4	WAIST_CI <= 72.7755	Negative	97.2%
5	AGE > 60 & WAIST_CI <= 103.65 & RIGHT_TH <= 52.0715 & LEFT_THI > 49.67 & TRUNK_VO > 50366.44 & RT_LEG_V > 6361.76	Positive	85.7%
6	AGE $\in$ (55, 69) & WAIST_CI $\in$ (103.65, 107.25) & RIGHT_TH <= 55.28 & LEFT_THI 48.7 & TRUNK_VO $\in$ (43706, 50183.94)	Positive	83.3%

4. DISCUSSION

We found the different body factors from these four data mining approaches. The accuracy of the classification of the models conducted from these four approaches all exceeded 80%. The result of that is acceptable when compared with other epidemiological research. Many epidemiological researches often use a statistics approach, such as logistic regression, to predict a disease. However, we can't obtain the full meaning of input x to output y even though this relationship is significant via logistic regression approach. Therefore, the other approaches, i.e., neural network, decision tree and rough set were introduced in this study. The order of accuracy of these four models is decision tree, rough set, logistic regression, and with neural network the least accurate. Note that decision tree with a function of adjusting cost is helpful for the classification tasks. So, it is a principal cause that the performance of decision is greater than all the other approaches, perhaps.

In clinical practice, impaired fasting glucose (IFG) and fasting plasma glucose level (FPG) are often two predictors for diabetes mellitus. Nevertheless, there exists no accurate and precise measure for body composition. Although some researches indicate that BMI and WHR are related to metabolic syndrome, hypertension, diabetes and hyperlipidemia, pure height and/or weight measure vary significantly across ethnic groups [29,30]. However, the result of this study is not restricted to BMI and WHR, i.e. height, weight, waist circumference, and hip circumference. Obviously, the artificial intelligence approach brings with it new features which are different from the traditional statistics.

Studies have reported a steady increase in the incidence of chronic diseases such as diabetes with increasing BMI. However, physicians may find some inconsistent conditions in their diagnosis. Such as with some metabolic diagnosis, a patient without obesity, i.e., although BMI in normal, he/she has been suffering from diabetes for a long time. On the other hand, some patients do not suffer from diabetes but their BMI are classed in the abnormal level. Furthermore, ethnic groups create a bias across the BMI [31–33]. Thus, the BMI seems to have limitations in the

interpretation of its association with diabetes. Physicians may risk making a misdiagnosis for diabetes if they base their assessment solely on BMI.

In practice, some research shows that body fat is related to diabetes [12,13,30]. For adults, the fat usually disperses uniformly to viscera and subcutaneous tissue. However, Erwin *et al.* [12] consider that the subcutaneous adipose tissue in people with diabetes, especially in the lower trunk, is greater than in healthy people. Furthermore, some research shows that the visceral fat is a major risk for impaired fasting glucose [29,30]. Therefore, they infer that the visceral fat is a risk factor for diabetes. However, we obtained a crude index when we based on WHR, i.e., waist circumference, to evaluate visceral fat in relation to diabetes. It was relatively easy to do, even though advanced medical techniques, such as computer topography (CT) and magnetic resonance imaging (MRI) are available to evaluate visceral fat. These techniques however are much too expensive for screening all patients, and it could reduce the wish to be examined for diabetes for some patients. Thus, a simple and accurate approach is worthy of performing.

In this study, we find six factors associated with diabetes, and they are in order of importance: volume of trunk, left thigh circumference, right thigh circumference, waist circumference, volume of right leg, and subjects' age. This result provides a new approach for the diagnosis of diabetes. It is not only a simple approach, but the accuracy of classification is satisfied in the diabetes diagnosis when we measure the patient's thighs. Furthermore, we may obtain a better performance of DM diagnostic using thighs because the body weight is almost entirely loaded on the thighs. Some studies show that patients with a metabolic syndrome such as diabetes have dimensions that are greater than that of the healthy group. As anticipated, their thighs also have dimensions larger than that of the healthy group. In addition, this noncontact method of making measurements may decrease some problems, such as bacterium infection. Basically, when using our approach, the evaluation of subcutaneous adipose tissue of the thigh is more accurate than the evaluation of the visceral fat at the waist. Nevertheless, this study shows that the waist circumference is also an important factor, and it is consistent with previous researches [34].

5. CONCLUSION

The aim of this study was to investigate what the risk factors were for anthropometrical data of Type II diabetes using four data mining approaches. Accuracy of classification was used to evaluate the performance of these four models. First, we found six factors including right thigh circumference, left thigh circumference, volume of trunk, waist circumference, volume of right leg and subjects' age to diagnose DM. Compared with the traditional approach for diagnosing DM, in particular the biochemical test, our study provides a new way with regard to anthropometry interventions, for doing that. We also found that the thigh circumference is a good factor (i.e., with high weight or significance) among the anthropometrical data in any one of these four approaches. It is obvious that using the thigh circumference to diagnose DM is a better alternative than using BMI or WHR. Furthermore, measuring the thigh circumference can be done quickly and simply, but the 3D-whole-body-scanning procedure can reduce the discomfort of the subjects.

At the same time, the accuracy of classification of all of the models was greater than 80%. This indicates that all the approaches of either the statistics-based or the AI-based could provide a good performance of classification of the case. Even though each approach is founded in a strong theory, the performance of the decision tree (entropy-based) and the rough set (via indiscernibility) are still greater than the logistic regression and neural network. In addition, the decision tree with a flow-chart-like tree structure is good at interpreting the results, and is a good tool for persons who have no informatics knowledge, such as physicians. Also, the rules from the decision tree induction are helpful in a physician's diagnosis, and are also good for the prevention of DM in clinical medicine. Concerning DM diagnosis, in particular the evaluation of the risk for contracting DM using anthropometrical data such as thigh circumference, is certainly worth investigating in future study.

REFERENCES

1. Canadian Diabetes Association, *Diabetes Dictionary* **http://www.diabetes.ca**, (2003).
2. Y.I. Kim, C.H. Kim, C.S. Choi, Y.E. Chung, M.S. Lee, S.I. Lee, J.Y. Park, S.K. Hong and K.U. Lee, Microalbuminuria is associated with the insulin resistance syndrome independent of hypertension and type 2 diabetes in the Korean population, *Diabetes Research and Clinical Practice* **52**, 145–152, (2001).
3. C.H. Chen, K.C. Lin, S.T. Tsai and P. Chou, Different association of hypertension and insulin-related metabolic syndrome between man and women in 8437 nondiabetic Chinese, *American Journal of Hypertension* **13** (7), 846–853, (2000).
4. G. Dorffner and G. Porenta, On using feedforward neural networks for clinical diagnostic tasks, *Artificial Intelligence in Medicine* **6**, 417–435, (1994).
5. D.B. Fogel, E.C. Wasson III, E.M. Boughton and V.W. Porto, Evolving artificial neural networks for screening features from mammograms, *Artificial Intelligence in Medicine* **14**, 317–326, (1998).
6. R. Folland, E. Hines, R. Dutta, R. Boilot and D. Morgan, Comparison of neural network predictors in the classification of tracheal-bronchial breath sounds by respiratory auscultation, *Artificial Intelligence in Medicine* **31**, 211–220, (2004).
7. M.K. Markey, J.Y. Lo, G.D. Tourassi and C.E. Jr. Floyd, Self-organizing map for cluster analysis of a breast cancer database, *Artificial Intelligence in Medicine* **27**, 113–127, (2003).
8. A. Carlos, R. Pena and S. Moshe, Evolutionary computation in medicine: An overview, *Artificial Intelligence in Medicine* **19**, 1–23, (2000).
9. M. Kukar, I. Kononenko, C. Grošelj, K. Kralj and J. Fettich, Analysing and improving the diagnosis of ischaemic heart disease with machine learning, *Artificial Intelligence in Medicine* **16**, 25–50, (1999).
10. F. Mizoguchi, H. Ohwada, M. Daidoji and S. Shirato, Using inductive logic programming to learn classification rules that identify glaucomatous eyes, In *Intelligent Data Analysis in Medicine and Pharmacology*, (Edited by N. Lavrač, E. Keravnou and B. Zupan), pp. 227–242, Kluwer, (1997).
11. W.R. Shankle, S. Mani, M.J. Pazzani and P. Smyth, Dementia screening with machine learning methods, In *Intelligent Data Analysis in Medicine and Pharmacology*, (Edited by N. Lavrač, E. Keravnou and B. Zupan), pp. 149–166, Kluwer, (1997).
12. T. Erwin, M. Reinhard, S. Karl and R. Gilbert, The determination of three subcutaneous adipose tissue compartments in non-insulin-dependent diabetes mellitus women with artificial neural networks and factor analysis, *Artificial Intelligence in Medicine* **17**, 181–103, (1999).
13. T. Erwin, M. Reinhard, S. Karl and R. Gilbert, Artificial neural networks compared to factor analysis for low-dimensional classification of high-dimensional body fat topography data of healthy and diabetic subjects, *Computers and Biomedical Research* **33**, 365–374, (2000).
14. P.J.F. Lucas, Analysis of notions of diagnosis, *Artificial Intelligence* **105** (12), 295–343, (1998).
15. P. Abdolmaleki, L.D. Buadu and H. Naderimansh, Feature extraction and classification of breast cancer on dynamic magnetic resonance imaging using artificial neural network, *Cancer Letters* **171** (2), 183–191, (2001).
16. L. Fausett, *Fundamentals of Neural Networks*, Prentice-Hall International, (1994).
17. H. Tsukimoto, Extraction rules from trained neural networks, *IEEE Transactions on Neural Networks* **11**, 377–389, (2000).
18. R. Andrews, J. Diederich and A.B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge-Based System* **8**, 373–389, (1995).
19. R. Andrews and J. Diederich, Rules and networks, *Proceedings of Rule Extraction Trained Artificial Neural Networks Workshop, AISB*, (1996).
20. C.T. Su, H.H. Hsu and C.H. Tsai, Knowledge mining from trained neural networks, *Journal of Computer Information Systems* **42** (4), 61–70, (2002).
21. J.W. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, (2001).
22. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, (1993).
23. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Wadsworth Int., Belmont, CA, (1984).
24. J.R. Quinlan, Induction of decision trees, *Machine Learning* **1** (1), 81–106, (1986).
25. D.G. Kleinbaum, *Logistic regression: A self-learning text*, Springer-Verlag, New York, (2002).
26. Z. Pawlak, Why rough sets, *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems* **2**, 738–743, (1996).
27. NeuralWare, *NeuralWorks Professional II / Plus getting started: A tutorial for Microsoft Windows computers, Version 5.50*, Carnegie, PA, (2001).
28. Rosetta GUI Version 1.4.41 **http://idi.ntnu.no/~aleks/rosetta/**, NTNU, Norway, (2001).
29. M.J. NcNeely, E.J. Boyko, J. B. Shofer, L. Newell-Morris, D.L. Leonetti and W.Y. Fujimoto, Standard definitions of overweight and central adiposity for determining diabetes risk in Japanese Americans, *American Journal of Clinical Nutrition* **74**, 101–104, (2001).
30. H. Nagaretani, T. Nakamura, T. Funabashi, K. Kotani, M. Miyanaga, K. Togunaga, M. Takahasi, H. Nishizawa, K. Kishida, H. Kuriyama, K. Hotta, S. Yamashita and Y. Matsuzawa, Visceral fat is major contribution for multiple risk clustering in Japanese men with impaired glucose to tolerance, *Diabetes Care* **24**, 2127–2133, (2001).

31. N.G. Norgan, Population difference in body composition in relation to the body mass index, *European Journal of Clinical Nutrition* **48** (Supplement), S10–S25, (1994).
32. J. Wang, J.C. Thornton, M. Russel, S. Burastero, S.B. Heymsfield and R.N. Pierson, Asians have lower BMI but higher percentage body fat than do white: Comparison of anthropometric measurements, *American Journal Clinical Nutrition* **60**, 23–28, (1994).
33. D. Gallagher, M. Visser, D. Sepulveda, R.N. Pierson, T. Harris and S.B. Heymsfield, How useful is BMI for comparison of body fatness across age, sex and ethic groups, *American Journal of Epidemiology* **143**, 228–239, (1996).
34. A.H. Kissebah, D.S. Freedman and A.N. Peiris, Health risks of obesity, *Medicine Clinics of North America* **73**, 111–138, (1989).