THE USE of MAJOR RISK FACTORS for COMPUTER-BASED DISTINCTION of DIABETIC PATIENTS with ISCHEMIC STROKE and WITHOUT STROKE

Sibel Oge Merey¹, Fikret Gurgen², Nurgul Gurgen²
¹Electronics and Electrical Eng. Dept., Technical Univ of Istanbul, Maslak-Istanbul/TURKEY
²Dept of Computer Eng., Boğaziçi University, TR-34342 Bebek-Istanbul/TURKEY
E-mail: gurgen@boun.edu.tr

Abstract—This study proposes a computer-based decision support system to investigate the distinctive factors of diabetes mellitus (DM) with ischemic (non-embolic type) stroke and without stroke. Database consists of a total of 16 features that are collected from 44 diabetic patients. Features include age, gender, duration of diabetes, cholesterol, higher density lipoprotein (HDL), triglycerid levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease (PVD), myocard infarction rate, fasting and random glucose levels, taking medicine, blood pressure. Metric and non-metric features are distinguished. First, the statistics, mean and covariance, of data are estimated and the correlated components are observed. Second, principal component analysis (PCA) is used for major components. Finally, decision making approaches, k-nearest neighbor (k-NN) and MLP, are employed for classification of all the components and major components case. Macrovascular changes were generally ineffective discriminators. Recommendations were made according to the rules of evidence-based medicine. Briefly, this case study supports theories of stroke in DM and also concludes that the use of intelligent data analysis improves personalized prevention.

Key words: Intelligent Data Analysis (IDA), decision support system, diabetes mellitus (DM), non-embolic (ischemic) stroke, principal component analysis (PCA), decision making, k-nearest neighbor (k-NN), multilayer perceptron (MLP)

1 INTRODUCTION

Stroke has been an important health issue and expressing and interpreting risk factors provides vital information [1-13]. This study discusses a computational method for highlighting the major risk factors of diabetic patients with non-embolic stroke and without stroke by performing dependency analysis and decision making. For this purpose, the follow up data of 22 diabetic patients with ischemic stroke (non-embolic type) and 22 diabetic patients without stroke were collected during a few years. Average population age was 66.2±9.9 for the stroke group and 61±6.1 for the control group. It is known that diabetes mellitus (DM) is diagnosed by fasting glucose level is higher than 140 mg/dl and random glucose level is higher than 200 mg/dl in repeated measurements. The study population of 44 patients were chosen with these glucose levels. Then, a set of tests are applied to construct the parameters of each feature vector. The tests include age, gender, duration of diabetes, cholesterol, higher density lipoprotein (HDL), triglycerid levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease, myocard infarction rate, fasting and random glucose levels (FGL and RGL), taking medicine, systolic and diastolic blood pressure. Metric and non-metric features are distinguished. Monitoring DM patients with various diagnostic tests and searching evidences of disease are generally routine and critical also in stroke [4-5]. Measuring major risk factors of ischemic stroke in DM can also provide arguments to the attending physician to initiate preventive measures adapted to patient’s case. Typical microvascular complications are neuropathy, nephropathy, retinopathy and macrovascular ones are coronary artery diseases (CAD) and peripheral vascular diseases (PVD) [6-7]. As known, microvascular and macrovascular complications may occur during DM. The macrovascular complications such as cholesterol, HDL, triglycerid levels, FGL and RGL, systolic and diastolic blood pressure are considered the risk factors of nonembolic-ischemic stroke on DM subjects [5-12]. Various studies have addressed the relationship between DM and stroke, and the probable risk factors [4-12]. Even though ischemic cerebrovascular is generally caused by macrovascular changes [7-9], the cerebral ischemia is claimed by the small artery occlusions in DM [6]. In DM, age, hypertension, MI, PVD are the known risks of stroke [4-5]. Also, fewer studies are available to indicate the relationship between neuropathy, nephropathy, retinopathy and stroke.
# The Use of Major Risk Factors for Computer-Based Distinction of Diabetic Patients with Ischemic Stroke and Without Stroke

- **Contract Number**
- **Grant Number**
- **Program Element Number**

**Author(s)**
- Project Number
- Task Number
- Work Unit Number

**Performing Organization Name(s) and Address(es)**
Electronics and Electrical Eng. Department Technical Univ of Istanbul Maslak-Istanbul/Turkey

**Sponsoring/Monitoring Agency Name(s) and Address(es)**
US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500

**Distribution/Availability Statement**
Approved for public release, distribution unlimited

**Supplementary Notes**
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

**Abstract**

**Subject Terms**

**Report Classification**
unclassified

**Classification of this page**
unclassified

**Classification of Abstract**
unclassified

**Limitation of Abstract**
UU

**Number of Pages**
5
Means, standard deviations and correlations are estimated from the observed population. It is known that small sample sizes reduce the statistical power of the study and generally result in wide confidence interval of estimated parameters. Thus, identical studies on larger samples may identify more important differences that have gone undetected here. Also, the deviation from the normality may cause errors in parameter estimations. As a compensation to hard-collected, small size of study population problem, we first observe these statistics by normal distribution assumption then, attempt to employ nonparametric classification techniques such as k-NN and MLP with jackknife method. Even though there exist ways of checking the fitness: for example, how well the normality of distribution for each component can be evaluated by various tests such as Kolmogorov-Smirnov test. But our aim is to build a decision support mechanism.

Regardless of the limitations imposed by technical artifacts and sample size constraints, our results describe several important findings emerged from this preliminary investigation of the relationship of diabetic patients between non-embolic stroke and without stroke.

Mean and standard deviation of metric components: cholesterol, HDL, triglycerides, FGL and RGL, systolic and diastolic blood pressure were found (Table 1). Then, correlation estimate denoted pairwise relationships of the components. Nonmetric components were counted as the percentage of existence or nonexistence cases (Table 1).

3 INVESTIGATING PRINCIPAL FACTORS OF DATA

By PCA analysis [14-18] of all components, we find the first macrovascular components of Table 1: cholesterol, triglycerides, glucose levels, blood pressures as the principal components. In decision making point of view, we observe that they are the most informative components in the study population. In fact, this confirms the reports of major risk factors in the literature [4-13]. Among the other components, the choice of age and gender are designed during data collection and have less information. Microvascular changes, neuropathy, nephropathy, retinopathy, are also assessed by only two distinctive levels (existence-nonexistence) to show the subgroups of DM subjects. Thus, they have a limited information that can be used for decision. PCA is found to be valuable for summarizing the trends of features of high risk subjects with DM.

4 DECISION MAKING BY PRINCIPAL FACTORS

It is known that there is no standard approach to establish the optimum decision criterion for a generic problem. Nonparametric classification approaches [15-17] overcome the difficulty of accurate estimation of underlying distributions from small population size. They are used without assumption about form of density function and their decision procedures bypass probability estimation and go directly to decision functions. Due to small population, we here employ nonparametric techniques instead of parametric ones such as K-Means, vector quantization (VQ) and gaussian mixture model (GMM) which, otherwise, would be very useful with the large population. Furthermore, fuzzy modelling methods enhance the decision ability [19].

Among various nonparametric techniques, we choose the local k-nearest neighbor (k-NN) and the global multilayer perceptron (MLP) approach. The concept of locality and globality is related to the location of information that is extracted for class decision. k-NN rule employs local information, in contrast, MLP extracts global information.

In k-NN rule, the class decision of unknown sample is based on the majority of the nearest k samples. In other way, a local voting process decides the class of unknown with the highest vote. This, in fact, an approximation of Bayes Decision Theory in a local environment [16]. This process can further be extended to weighted voting and gaussian weighted voting (Parzen window) [16-18]. To recognize pattern v, k minimum distance samples are computed among all the samples. The distance can be computed by various norms: minkowski-norm-based distances, e.g. euclidian as second order, covariance-based distance or mahahonobis, entropy-based distance or kullback-leibler distances. The simplest and the most suitable for small sample size is euclidian metric which is defined as

\[ d_j = |v-v_j|^2 = (v-v_j)^T(v-v_j) \]

where \( d_j \) defines the distance between patterns \( v \) and \( v_j \), k-NN rule classifies \( v \) by assigning it a label that is the most frequently represented among the k nearest samples:

\[ k_i = \max \{ k_1, \ldots, k_L \} \quad \text{subject to} \quad \bigwedge_{j=1}^{L} \omega_j \]

Here, \( k_i \) is the number of neighbors belong to \( \omega_i \) class among the \( k \) nearest neighbors.

The global MLP [15-18] is a parallel, feedforward structure that consists of input, output and hidden layers. Each layer has sigmoidal units interconnected through weight connections. The MLP is trained with supervised back propagation (BP) algorithm to efficiently compute partial derivatives of an approximating function \( F(w,x) \). The network has an adjustable weight vector \( w \) that is computed with respect to all training data for a given value of input vector \( x \) and output vector \( y \). The weights are adjusted to fit a set of surfaces to the input space. The surfaces are constructed by sigmoid by the best linear regression concerning the cluster membership. The mean square (MSE) (5) error function, which is the difference between the
network's output and the supervisor output, is minimized to find the cluster membership:

\[ \text{MSE} = \sum_{q} (y_q - F(w, x_q))^2 \]  

(3)

In k-NN rule, an arbitrary discriminant function is constructed for class decision. The nearest neighbor contributes the half of the classification information. A MLP can generate any nonlinear discriminant function of input by incorporating multiple constraints. Each sigmoidal unit contributes to global discriminant function by a linear constraint.

5 TESTS AND RESULTS

The proposed system is shown in Fig. 1. Figure 2 shows the distribution of RGL-FGL-cholesterol features. The study population contains 44 diabetic patients' collected by neurology specialists. The data has a total of 16 metric and non-metric components. The computer-based system uses two stages: first, PCA method reduces the features to 7 dimensions, then, the best discriminators for class membership are searched by nonparametric classifiers, k-NN and MLP. Multi-parameter classifiers were found to be significantly improve upon the classification performance of single parameter designs. Instead of single parameter based conclusions, we employed the decision produced by major risk factors. This makes the study have stronger arguments on the distinctive factors. The small sample population size has an acknowledgeable potential effect on the statistical power of the study but we use classification techniques like k-NN and MLP to overcome this drawback.

The results are shown in Table 3 and Table 4. An average classification score of 68.18% is obtained for k-NN. The performance of MLP is tested using hold-one-out method (jackknife) as 100%. Both classifiers identify the macrovascular changes as the best discriminators as of this case study on stroke.

The figure of component clusters clearly provides convincing visual information to the physician and patients as the end-users. This easy-to-use graphic offers users information about the limits of the major risk factors.

6 DISCUSSIONS AND CONCLUSION

In summary, this paper presents a computer-based decision support system for physicians to improve prevention measures for ischemic stroke in DM subjects. PCA helps to identify the basic trends that distinguish two cases: ischemic stroke and no stroke in DM. In this study, major distinctive factors are found to be cholesterol, triglyceride level, fasting and random glucose levels (FGL and RGL), systolic and diastolic blood pressure levels. Physicians furthermore can suggest diastery supplies, drugs, exercises, etc. for the prevention of ischemic stroke for diabetic patients with these findings. Visual graphics of the components also support the discussion for clues of prevention.

The aim of this study is to support doctors with intelligent data analysis techniques. The system may be thought as an indirect assistance tool for physicians. We also plan to improve the information of diagnosis with a certain risk assessment factor. A grading for the risk of ischemic stroke for diabetic patients will support the prevention measures according to evidence-based medicine.

The major concern of the study is that the investigators could not have a bigger size of study population. This raises concerns about the meaningfulness of results but the authors believe that the methods used here clearly useful to overcome this handicap. Also the results of the study are supported by the risk factor findings of the other stroke and diabetic studies in the literature [4-13]. When there is a large database, this will provide more convincing statistical power to the study. In this case, the general statistics will confirm the results and parametric decision making techniques such as expectation-maximization (EM) algorithm would be utilized.

The other limitation is the inability to directly account for certain microvascular factors in the decision process due to imprecision of their records: for example, retinopathy has only two identifying labels as exist or non exist. In fact, there are typically four cases that can be distinguished by the physician but the limited data size prevents us to include finer details. This may cause an error. The lack of precision, in general, reduces the effect of the feature in decision making.

Among the various other tests, we also believe that the doppler ultrasound measurements may present more detailed information on this case study. The areas of arteries and veins can be monitored for deciding the stages of the disease, and the condition of occlusions can be important for the preventive follow up.

The other weak point of this investigation may be the quality issues of the recording such as the use of nonstandardized procedures, the use of potentially low grade equipment or the measurement quality. However, as best supported by a group of physicians' observations, this was not the case.

As a conclusion, this case study points at a fruitful line of enquiry in intelligent medical data analysis and the content of the work can further be extended to the other areas of stroke diagnosis and prevention.

7 REFERENCES


Table 1 Statistics of metric components of DM patients with ischemic stroke without stroke

<table>
<thead>
<tr>
<th>Component</th>
<th>First class</th>
<th>Second class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>9.8992</td>
<td>6.1011</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>54.5235</td>
<td>50.9238</td>
</tr>
<tr>
<td>Triglycerin</td>
<td>97.7951</td>
<td>89.8721</td>
</tr>
<tr>
<td>HDL</td>
<td>6.4842</td>
<td>6.8307</td>
</tr>
<tr>
<td>RGL</td>
<td>86.1668</td>
<td>206.9909</td>
</tr>
<tr>
<td>FGL</td>
<td>75.3457</td>
<td>187.7727</td>
</tr>
<tr>
<td>Sistolic</td>
<td>26.2851</td>
<td>14.6015</td>
</tr>
<tr>
<td>Diastolic</td>
<td>12.9267</td>
<td>10.0195</td>
</tr>
<tr>
<td>DM (month)</td>
<td>91.0637</td>
<td>87.6972</td>
</tr>
</tbody>
</table>

Table 2 Statistics of non metric components of DM patients with ischemic stroke without stroke

<table>
<thead>
<tr>
<th>Component</th>
<th>First class</th>
<th>Second class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>6 (%27.3)</td>
<td>8 (%36.4)</td>
</tr>
<tr>
<td>Medicine</td>
<td>11 (%50)</td>
<td>17 (%77.3)</td>
</tr>
<tr>
<td>Nephropathy</td>
<td>7 (%31.8)</td>
<td>9 (%40.9)</td>
</tr>
<tr>
<td>PVD</td>
<td>6 (%27.3)</td>
<td>7 (%31.8)</td>
</tr>
<tr>
<td>Retinopathy</td>
<td>4 (%18.2)</td>
<td>1 (%4.5)</td>
</tr>
</tbody>
</table>

Table 1 Statistics of metric components of DM patients with ischemic stroke without stroke

Table 2 Statistics of non metric components of DM patients with ischemic stroke without stroke

![Figure 1: Distinctive risk factors of stroke in DM and without stroke](image1)

![Figure 2: Appearance of RGL-FGL-cholesterol data](image2)
### Table 3 Results of k-NN method

<table>
<thead>
<tr>
<th>$k$</th>
<th>All Components</th>
<th>Principle Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52.2727</td>
<td>52.2727</td>
</tr>
<tr>
<td>3</td>
<td>65.9091</td>
<td>65.9091</td>
</tr>
<tr>
<td>5</td>
<td>68.1818</td>
<td>68.1818</td>
</tr>
<tr>
<td>7</td>
<td>63.6364</td>
<td>63.6364</td>
</tr>
<tr>
<td>10</td>
<td>68.1818</td>
<td>68.1818</td>
</tr>
<tr>
<td>20</td>
<td>63.6364</td>
<td>63.6364</td>
</tr>
<tr>
<td>40</td>
<td>29.5455</td>
<td>29.5455</td>
</tr>
</tbody>
</table>

**Epsilon=0.1 alpha=0.8 iteration=10000**

<table>
<thead>
<tr>
<th>$h$</th>
<th>first class</th>
<th>second class</th>
<th>whole</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>86.36</td>
<td>86.36</td>
<td>86.36</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4 Results of MLP method