SHAPING DATA: A SELF ORGANIZING MAP APPROACH FOR DATA MINING OF ORAL GLUCOSE TOLERANCE TEST CURVES IN WOMEN WITH PREVIOUS GESTATIONAL DIABETES

L. Gaetano[A], G. Di Benedetto[A], A. Tura[B], G. Balestra[C], F.M. Montevecchi[A], A. Kautzky-Willer[D], G. Pacini[B], U. Morbiducci[A]

[a] *Dipartimento di Meccanica, Politecnico di Torino, Torino, Italy*
[b] *Institute of Biomedical Engineering, CNR, Padova, Italy*
[c] *Dipartimento di Elettronica, Politecnico di Torino, Torino, Italy*
[d] *Department of Internal Medicine III, University of Wien, Wien, Austria*

INTRODUCTION

Gestational diabetes mellitus (GDM) is defined as the diabetic condition with onset during pregnancy and it affects from 1% to 14% of all pregnancies depending on the population studied [1]. In general, shortly after delivery, glucose homoeostasis is restored to the antepartum condition, but women with a history of GDM show at least a seven-fold increased risk of developing type 2 diabetes compared with those who had a normoglycaemic pregnancy [2]. Hence, since these women represent a high-risk population, there is the need to develop appropriate preventive strategies and to identify reliable prognostic factors. In the last years, different antepartum and postpartum independent predictors of later abnormal glucose tolerance have been identified, but anyone seems so reliable.

In this study, we hypothesize that the future evolution to a condition of normal glucose tolerance or type 2 diabetes is predictable from the morphology of the OGTT curves at baseline. In order to evaluate this potential predictor capacity, the first step is to evaluate if additional and useful information is contained into curves shape besides the two specific values used for current diagnosis of normal/diabetic condition. For this reason, we used a particular neural network, i.e. Self-organizing map (SOM), in order to cluster OGTT curves basing on their shape. The SOM-based analysis was compared with the clinical, shape-independent classification of the glucose tolerance condition (i.e. the gold standard), in order to assess whether the morphology of the OGTT curves (i) are correlated to such condition, and (ii) maintain memory of the previous disease (GDM condition). Moreover, having a small number of curves at disposal, the other aim of this study is to find out an efficient method able to guide data mining in presence of small datasets.

METHODS

A group of 92 Caucasian women with GDM was investigated together with a control group (CNT) of 40 women. Gestational diabetes was diagnosed according to American Diabetes Association (ADA) criteria [3]. They were studied for a maximum of 5 years after delivery. All women underwent a standard 75-g OGTT every year. In this preliminary analysis, however, only OGTT data at the baseline condition were used. According to the criteria proposed by ADA in 1997 [4], the population was divided into a normotolerant group (NGT), a group with impaired glucose tolerance (IGT), and a group with type 2 diabetes (T2DM).

The morphological analysis was performed over all the glucose, insulin and C-peptide curves made available by the OGTT test. The analysis was conducted both on measured curves, and on curves obtained by the measured ones by removing their mean value for investigating whether curves can be classified exclusively in terms of their morphology, or the classification is biased by the exact value of each sample of the curve. The analysis was conducted using Self-organizing map (SOM), a subtype of artificial neural networks which uses a competitive learning technique to train itself in an unsupervised manner [5]. Using SOM, we could obtain a map that is topologically ordered: this means that *n* topologically close input data vectors map to *n* adjacent map neurons or even to the same single neuron, underlining shape input similarities and dissimilarities. For the SOM design, a hexagonal lattice map, a linear initialization of prototype vectors and a batch training algorithm was chosen, while the dimension of the grid depended on the size of the training sets used. The input sets were different:

- curves belonging to CNT (n = 40) and NGT groups (n = 43) were used with the aim of discovering whether a sort of memory of the previous disease (the GDM condition) is maintained in the morphology of the NGT curves (memory-of-disease-driven shape);
- combination of curves belonging to CNT, NGT, IGT (n = 24) and T2DM (n = 21) groups were used for training SOMs, aiming at evaluating whether there are substantial morphological differences among curves of different groups.

After training, the natural clustering tendency of curves was evaluated. Moreover, the available a priori knowledge about the input dataset was then used: each neuron was, in fact, afterwards labelled with class of the most numerous group of curves represented by that node. In this way, it was possible to understand how the current classification is represented by curves morphology and if SOM makes additional information come out.

The entire analysis was performed inside Matlab environment (The Mathworks, Inc., Natick, USA).

RESULTS AND DISCUSSION

First of all, there was any important difference between results obtained with measured curves and those obtained with curves subtracted of their mean value (data not shown). A visual depiction of the analysis performed over glucose curves belonging to CNT and NGT groups is shown in Figure 1a: notably, different colours associated to different groups did not identify well separated regions in the map. This means that SOMs were not able to assign CNT and NGT curves to distinct regions of the map. On the contrary, concerning the morphological differences between NGT and T2DM groups, the SOM based analysis put in evidence a clear division in the morphology of the OGTT measured waveforms (Figure 1b), in particular when only glucose curves were considered. When the IGT group was included in the analysis, SOM analysis identified specific regions in the map referable to NGT, IGT and T2DM curves (Figure 1c, using glucose and insulin curves combined).

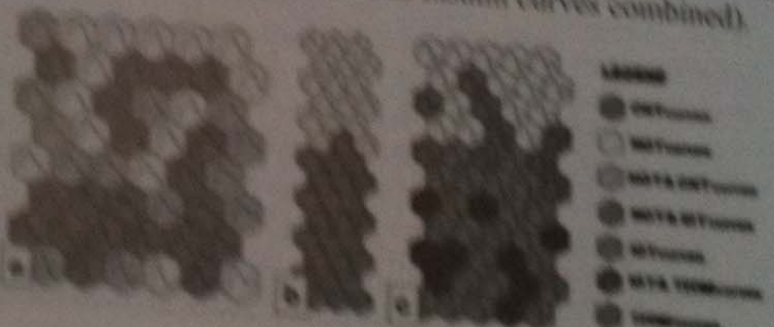

Fig. 1 SOMs obtained by comparison between CNT and NGT (a), NGT and T2DM (b), and among NGT, IGT and T2DM groups.

In conclusion, we succeeded in mining novel knowledge from our dataset even if it is relatively small. having not a large number of curves, through SOM, we have however extracted shape information that could be used for pattern recognition and feature selection in the next step, in which a relation between morphology characteristics and follow-up will be sought. Our results show that the whole morphology of the OGTT measured curves contain information about the current status of the patient with a history of GDM, because the SOM-based clustering clearly allows to discriminate subjects belonging to healthy or diabetic group even when the mean values is removed from the measured curves. Moreover, there are additional information that lead SOM to map nearer or not curves that currently belong to different groups. Exactly this topographic arrangement could be predictive of future evolution of patients.

REFERENCES

[1] ACOG Practice Bulletin, Obstet Gynecol, 2001, 98: 525-538.

[2] Bellamy L. et al., Lancet, 2009, 373: 1773-9.

[3] American Diabetes Association, Diabetes Care, 2003, 26: 103-5.

[4] The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, Diabetes Care 1997, 20 (7): 1183-97.

[5] Kohonen T., Springer-Verlag (New York), 2001.