Mathematical and Informational Tools for Classifying Blood Glucose Signals - A Pilot Study

Ariel Amadio[a], Andrea Rey[b,*], Walter Legnani[b], Manuel García Blesa[b], Cristian Bonini[c], Dino Otero[a]

[a] *Vehicle Research, Development, and Innovation Center, Universidad Tecnológica Nacional Facultad Regional General Pacheco, Av. Hipólito Yrigoyen 288, B1617, General Pacheco, Argentina*

[b] *Signal and Image Processing Center, Universidad Tecnológica Nacional Facultad Regional Buenos Aires, Av. Medrano 951, C1179, Buenos Aires, Argentina*

[c] *Research, Development, and Innovation in Electric Energy Center, Universidad Tecnológica Nacional Facultad Regional General Pacheco, Av. Hipólito Yrigoyen 288, B1617, General Pacheco, Argentina*

Abstract

A survey campaign was carried out on the dynamics of blood glucose measured through interstitial sensors of relative recent diffusion in the market. These sensors generated time series that were labeled according to medical diagnosis in diabetics and non-diabetics, and that constituted the data core of the classification models. Based on the calculation of the distribution of ordinal patterns of the time series, the corresponding points in the entropy-complexity causal plane were located. Moreover, the transition matrices of these ordinal patterns (OPTMs) were calculated in order to find the proximity using the Manhattan distance of every OPTM with respect to the mean of each group, associating the corresponding signal to each class. On the other hand, the Frobenius norm of every OPTM and the norm of its stationary vector were computed given different values for the considered classes. The effect of repeated values in a signal was also analyzed. Notable differences were obtained in the properties of the OPTMs of each class. In another sense, it is shown that diabetes is a disease that reduces the entropy of the temporal evolution of blood glucose in well-defined time periods, and presents values of complexity significantly higher than those obtained in subjects without

*Corresponding author

Email address: arey@frba.utn.edu.ar (Andrea Rey)

diabetes. The selected alternatives coincide in detecting patients positively diagnosed with type II diabetes mellitus. The calculations on the OPTMs show the correlation among patterns of the signals. At the same time, in the entropy-complexity plane, the considered groups were located in well-defined regions showing the differentiating power of these information measures, and indicating variations in the dynamics of the biological system when diabetes is present. With the four mathematical tools selected and the dynamical characterization given by the causal plane, it was possible to define an index that clearly differentiates the classes under study.

Keywords: glucose, ordinal patterns, entropy-complexity plane, transition matrices, tied data

---

1. Introduction

The fluctuations of glucose in the blood are governed by numerous factors and feedback loops that normally keep a relatively low level that is limited regardless of large variations imposed during the day [9]. Studies in animal species were laying the foundations for the oscillation of the Glucose-Insulin-Glucagon triad. Goodner [15] studied the behavior of this triad in the monkey species Macaca mulatta fasting overnight and discovered regular oscillations with periods of 9 minutes on average. Another antecedent can be found in the year 2011 when Jin Shi informed us about stable and regular oscillatory glucose uptake through the use of micro biosensors based on nanomaterials in the cells of the pancreas [27]. From another point of view, the values of glucose in the blood can be characterized as a dynamic system -chaotic deterministic- [4]. Thus, it is able to reconstruct its attractor and present at least one positive Lyapunov exponent.

Among well-known standard tests to detect Type II Mellitus Diabetes (DMT2), it can be mentioned the Oral Glucose Tolerance Test (OGTT) [11] and the Intravenous Glucose Tolerance Test (IVGTT) [23]. These tools are related to models based on glucose and insulin, which were proposed by Bergman [6] and Akerman [1], respectively. On the other hand, there exists a linear relationship between the daily glucose average and the glycosylated hemoglobin (HbA1c) [28], which can be used to make a diagnosis and to assess the evolution of the patient. By means of the value of the HbA1c, a case of study can be identified in the following classes: without DMT2, with DMT2, and with prediabetes. Data science and machine learning techniques,

such as neural networks, have been extensively applied in the classification of this disease. For instance, k nearest neighbors [12], logistic regression [31], decision trees [24], support vector machine [19], random forest [5, 13], and convolutional neural networks [16].

The aim of this work is to detect individuals with DMT2, using four mathematical tools to analyze the evolution of glucose measured with interstitial sensors. With the measurements of these sensors, a pilot signal database was constructed labeling samples with the disease under study medically diagnosed as positive (P), and subjects who do not suffer from diabetes labeled negative (N). With these signals, it is possible to introduce the ordinal patterns (OPs) method for the analysis of a time series structure, as those that want to quantify the characteristics of a set of data points by characterizing the sequenced distribution of a subset of values of the same signal or time series [2]. The cornerstone of this method lies in the study of the relative position of the measured data in many subsegments of fixed-length, and then collect all the information. In this way, they differ from the traditional methods of nonlinear time series analysis which, in general terms, compute parameters in the reconstructed phase space [8] or those obtained based on symbolic dynamics. The seminal work of Bandt and Pompe [3] opened the road to several types of research which have focused on the investigation of measured signals in complex biological systems, economy, and in a variety of applications (cf. [2, 29, 25, 18, 22], among many others). The wide diversity of applications of the OP method is based on the fact that it is not necessary to assume any characteristics of the system that produces the time series (technically it is said to be domain agnostic [10]). Once the OP were obtained, the next step was the computation of the transition probabilities between these patterns. These probabilities were approximated by the relative frequency of transition between neighboring points of every signal sub-segment defined by the embedding dimension and then, used to construct the OPTM for every signal. The informational measures implemented consisted of the normalized permutation entropy H [3], and the statistical complexity C [21], used to define the so-called $H \times C$ plane. This approach was incorporated to give support from the system dynamics framework for the detection of P records.

The selected mathematical tools can be split into two branches: the Manhattan distance from each OPTM corresponding to the individual signal to mean values of each class, and the matrix metrics taken on the own OPTM (the Frobenius norm and the stationary vector norm). In the first case, it

is possible to quantify the intensity of the transitions relative to a reference state -the mean state of each class-, and in the second case, the intensity of the transitions within signals can be pondered. The norm of a matrix indicates its ability to modify the norm of a vector that is multiplied by it. The Frobenius norm is an intermediate value among other norms that can be chosen, such as the infinity norm or the 2-norm. Besides, inducing an inner product between matrices allows quantifying the distance between them or the operators associated with them. In the present work, the norms of the stationary vector of the OPTMs were used to group the sensor records according to their numerical similarity in two suitably differentiated sets, and then associate them with the corresponding classes under analysis. This way of analysis contributed to explaining the results following the metrics of the OPTM associated with the intensity of the OP transitions.

These four tools have given coincident results, which make it possible to positively detect subjects suffering from DMT2, differentiating them from those who do not suffer from the disease.

This work is structured as follows: motivation and aims are contained in this introduction. A succinct approximation of the ordinal pattern is presented in Section 2. In Section 3, the entropy-complexity plane is described. The presentation of the concept of the ordinal pattern transition matrices is developed in Section 4. In Section 5, a very concise revision of the matrix tools applied in this work is presented. Section 6 is devoted to the introduction of the methodologies applied in the proposed approach to the treatment of tied data. The description of the blood glucose and thresholds for classification classes are contained in Section 7. The results of this pilot research are presented in Section 8. Finally, the closure of the work is done by discussion and conclusions developed in Section 9.

2. Ordinal Patterns

Bandt and Pompe [3] proposed a well-known methodology to describe the ordinal dynamic of the elements in a time series. This approach requires two parameters: the embedding dimension $D > 1$ and the embedding time delay $\tau \geq 1$. Then, for a given time series $X(t)$ of length T, $T - D + 1$ overlapping partitions of length D are defined as:

$$\boldsymbol{s}(t) = \{X(t - \tau(D-1)), X(t - \tau(D-2)), \ldots, X(t-\tau), X(t)\}, \quad (1)$$

where $t = D, D+1, \ldots, T$. It is necessary that $T \gg D!$. Let π be one of the $D!$ permutations of the elements $1, \ldots, D$. It is said that $\boldsymbol{s}(t)$ is of type π if $\boldsymbol{s}(t)_{\pi(i)} < \boldsymbol{s}(t)_{\pi(i+1)}$ for all $i = 1, \ldots, D-1$. Thus, for each partition, the assigned ordinal pattern is given by its permutation type $\pi(1)\pi(2)\cdots\pi(D)$.

3. Entropy-Complexity Plane

The permutation probability of a T-length time series is defined as $\mathscr{P} = \{p_1, p_2, \ldots, p_{D!}\}$ where, for $j = 1, 2, \ldots, D!$, the probability of the permutation π_j of $D!$ elements is given by:

$$p_j(\pi_j) = \frac{\#\{\boldsymbol{s}(t) \text{ of type } \pi_j\}}{T - D + 1}. \tag{2}$$

The well-known Shannon entropy is defined as:

$$S[\mathscr{P}] = -\sum_{j=1}^{D!} p_j \ln(p_j), \tag{3}$$

whose normalized version is $H[\mathscr{P}] = S[\mathscr{P}]/\ln(D!)$. This entropy is also named permutation entropy.

Let $\mathscr{P}_e = \{1/D!, \ldots, 1/D!\}$ be the equiprobable distribution. In [21], the authors defined a way to measure the complexity of a system using the Jensen divergence $\mathscr{D}$ as an assessment of disequilibrium. Precisely, the complexity of the time series can be calculated by:

$$C[\mathscr{P}] = Q_0 \mathscr{D}[\mathscr{P}, \mathscr{P}_e] H[\mathscr{P}], \tag{4}$$

where

$$\mathscr{D}[\mathscr{P}, \mathscr{P}_e] = S\left[\frac{\mathscr{P} + \mathscr{P}_e}{2}\right] - \frac{1}{2}S[\mathscr{P}] - \frac{1}{2}\ln(D!), \tag{5}$$

with normalization constant given by:

$$Q_0 = -2\left[\frac{D!+1}{D!}\ln(D!+1) - 2\ln(2D!) + \ln(D!)\right]^{-1}. \tag{6}$$

In order to study the evolution of the complexity over time, the relation between C versus the time t can be represented. This context is equivalent to the analysis of the curve defined by C versus H, due to the second thermodynamic law that establishes a monotonous increase of the entropy as a

function of time. Then, the causal plane $H \times C$ is formed by the horizontal axis that represents the normalized entropy, and by the vertical axis that represents the complexity. It is known that for a given entropy, there exist several ways to measure the statistical complexity, all of them ranging between two bound curves denoted by $C_{\min}$ and $C_{\max}$ [20]. In [21], the authors proposed an algorithm to compute both curves based on the geometric concept of simplices.

4. Ordinal Pattern Transition Matrices

Starting from the calculation of the OP for signal analysis or time series, two possible methodologies are available. One way is to compute the frequency of the appearance along the signal, and an alternative way is a method of transitional nature, which is precisely the calculation of the frequency of passage from an OP to the next OP present in the data sequence. In this second sense, each OP transition contains information on the short-range temporal structure of the adjacent observations and their linkage with the following segments [2], whose lengths are determined by the embedding dimension. Thus, studying how one OP is followed by the next one in a given data series reveals structural characteristics of the geometrical shape representing the data. Every OP can be labeled by an ordering number, generating a sequence of the form $1, 2, \ldots, D!$. Then, the transition frequencies between OP can be arranged in a matrix format, so that the references to columns and rows represent each OP label respectively, in such a way that constitutes a probability matrix $M = (M_{ij})_{1 \leq i,j \leq D!}$, where $M_{ij} \geq 0$ is the transition probability between patterns i and j, that satisfies $\sum_{i=1}^{D!} M_{ij} = 1$, for each $j = 1, 2, \ldots, D!$. This determines a column probability matrix, with all the appertaining properties of such mathematical entity [26].

5. Matrix Tools

This section is devoted to a collection of some well-known concepts and properties related to matrices that will be helpful in the description of the methodology proposed in this work.

Matrix Distance. Given $M, N \in \mathbb{R}^{n\times n}$, the Manhattan distance between M and N is defined by:

$$\operatorname{dist}(M, N) = \sum_{i,j}^{n} |M_{ij} - N_{ij}|. \tag{7}$$

Notice that this is the Minkowski distance of order 1.

Frobenius norm. Given $M \in \mathbb{R}^{n\times n}$, its Frobenius norm [17] is defined by:

$$\|M\|_F = \sqrt{\operatorname{tr}(MM^t)} = \sqrt{\operatorname{tr}(M^tM)}, \tag{8}$$

where M^t denotes the transpose matrix of M and tr the trace of a matrix in the classical sense.

In the case of stochastic matrices, the Frobenius norm can be used as an index of the degree of randomness of the matrix. In this sense, consider a stochastic matrix $M = \begin{pmatrix} C_1 & C_2 & \cdots & C_n \end{pmatrix} \in \mathbb{R}^{n\times n}$, where C_j is the jth-column of M. Then, its transpose is $M^t = \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{pmatrix} \in \mathbb{R}^{n\times n}$. Thus, $(M^tM)_{ij} = C_i \cdot C_j$ for $i, j = 1, 2, \ldots, n$.

Suppose M is a stochastic matrix that characterizes a deterministic process. It means that each column of M, equivalent each row of M^t, is a vector belonging to the canonical basis of $\mathbb{R}^n$, $E = \{e_1, e_2, \ldots, e_n\}$. It is straightforward that $(M^tM)_{ii} = C_i \cdot C_i = 1$. Then, $\sqrt{\operatorname{tr}(M^tM)} = \sqrt{\sum_{i=1}^{n}(M^tM)_{ii}} = \sqrt{n}$, which is the maximum value of the Frobenius norm.

On the other hand, given M a stochastic matrix that characterizes a random process, each column is of the form $(1/n, 1/n, \ldots, 1/n)$. In this case, $(M^tM)_{ii} = C_i \cdot C_i = \sum_{i=1}^{n} 1/n^2 = 1/n$. Hence, the minimum value of the Frobenius norm is $\sqrt{\operatorname{tr}(M^tM)} = \sqrt{\sum_{i=1}^{n}(M^tM)_{ii}} = \sqrt{1} = 1$.

Stationary vector. Given a right stochastic matrix M of order n (*i.e.* each row sums 1), it is irreducible if and only if $(I + M)^{n-1}$ has all positive entries [30, Appendix A], where I is the identity matrix of order n. In this case, the Perron-Frobenius theorem [26, Section 1.1] states that there exists a unique unitary vector $v \in \mathbb{R}^{1\times n}$ such that $vM = v$. This left-eigenvector of eigenvalue 1 is referred to as the stationary vector of M.

6. Tied Data

A natural question is how to deal with the case in which the time series presents repeated values, named tied data. The following three different methodologies are considered in the treatment of tied data throughout this work.

Omission. In this alternative, tied data are simply eliminated. For instance, consider the time series $X = \{5, 4, 7, 7, 2, 5\}$. With $D = 3$, the partitions are: $\boldsymbol{s}(1) = \{5, 4, 7\}$, $\boldsymbol{s}(2) = \{4, 7, 7\}$, $\boldsymbol{s}(3) = \{7, 7, 2\}$, and $\boldsymbol{s}(4) = \{7, 2, 5\}$. Since $\boldsymbol{s}(2)$ and $\boldsymbol{s}(3)$ have tied data, only the first and last partitions are considered. Thus, the ordinal patterns associated with X are 213 and 312.

Sequential order. Suppose that $\boldsymbol{s}(t)$ is a partition that presents the tied data $\boldsymbol{s}(t)_h = \boldsymbol{s}(t)_k$, with $h < k$. Then, it is defined $\pi(k) = \pi(h)+1$. Table 1 shows the associated patterns of tied data when $D = 3$ and $\tau = 1$.

Table 1: Ordinal pattern assignment when tied data are treated using sequential order in case $D = 3$.

Partition	**Ordinal Pattern**
$\{b, a, a\}$ with $a > b$	123
$\{a, a, b\}$ with $a < b$	123
$\{a, b, a\}$ with $a > b$	213
$\{b, a, a\}$ with $a < b$	231
$\{a, a, b\}$ with $a > b$	312
$\{a, b, a\}$ with $a < b$	132
$\{a, a, a\}$	123

Assignation of weights. In [7] it was proved that, after a linear transformation of the time series 3-length partitions, the regions of the ordinal patterns are well-defined by three lines where tied data are located. Then, tied data lay on the boundary of two neighbor ordinal patterns. Table 2 shows the ordinal pattern weights for each tied data.

It is worth noticing that the OPTM depends on the selected treatment of tied data. From now on the following notation will be used: M^O is the OPTM computed by the omission of tied data, M^S is the OPTM computed using the sequential order for tied data, and M^W is the OPTM computed by weighted tied data.

Table 2: Weights assigned to ordinal pattern for tied data in case $D = 3$, where empty places indicate null weights.

	Ordinal patterns					
Tied data type	123	132	213	231	312	321
(b, a, a) with $a > b$	0.5	0.5				
(a, a, b) with $a < b$	0.5		0.5			
(a, b, a) with $a > b$			0.5		0.5	
(b, a, a) with $a < b$					0.5	0.5
(a, a, b) with $a > b$				0.5		0.5
(a, b, a) with $a < b$		0.5		0.5		
(a, a, a)						

7. Classification of Blood Glucose Evolution Signals

As mentioned before, the purpose of the present work is to establish a technique able to identify patients suffering from diabetes. The definition of the proposed classification model is based on the time series of the blood glucose evolution. The sensor provided by ™FreeStyle Libre (`https://www.freestyle.abbott/`), is applied to the back of the arm, and automatically takes glucose readings every 15 minutes that can be scanned by an app that reads the information. The experiment consists of eight cases of study, divided into two groups of four volunteers each. One group is formed by people who have neither presented symptoms nor have been diagnosed with DMT2, and another group is integrated by patients who have been diagnosed with this disease for at least five years. The age and sex of these people are indicated in Table 3. The positive state Pi is defined when the diagnostic in case i indicates DMT2, being negative state Ni otherwise. The signals obtained by glucose in the blood are shown in Figure 1.

The OPs of these time series are computed using embedding dimension $D = 3$ and embedding time delay $\tau = 1$. In this case, the OPTM for each time series is of order 6.

Four methods are proposed to distinguish persons with and without diabetes based on the glucose time series $\mathcal{G}$:

1. Location of points obtained from $\mathcal{G}$ in the Entropy-Complexity plane, computing the minimum distance to the centroids $c_P = (0.7901, 0.1619)$ of positive cases, and $c_N = (0.9060, 0.0816)$ of negative cases.

Table 3: General characteristics of the cases of study in the preliminary database.

Case	**Sex**	**Age**
N1	Male	39
N2	Female	43
N3	Male	27
N4	Female	48
P1	Male	33
P2	Male	33
P3	Male	59
P4	Male	61

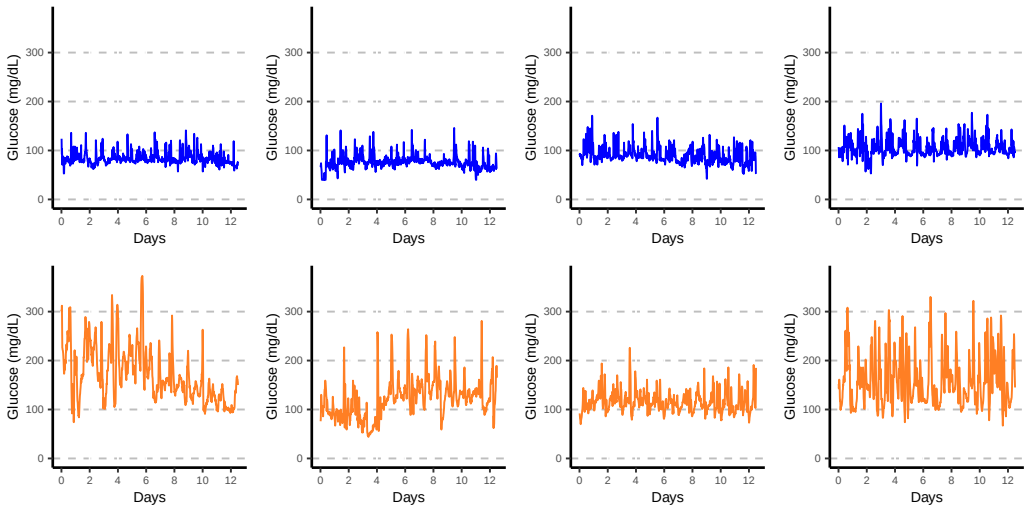

Figure 1: Time series of glucose in the blood for the cases of study: Negative (first row) and Positive (second row).

2. Manhattan distance between the OPTM of $\mathcal{G}$ and the average OPTM μ_P of positive cases, and μ_N of negative cases.
3. Frobenius norm of OPTM of $\mathcal{G}$, that ranges from 1 to $\sqrt{6} \cong 2.4495$.
4. Norm of the stationary vector of the OPTM of $\mathcal{G}$.

Three classes are defined: N (negative case), P (positive case), and D (doubtful case). In a doubtful case, a prediabetes diagnosis could be possible, and medical advice is required. The class labeled by D is defined in terms of thresholds that are selected from the behavior observed by the reference cases. Given a glucose time series $\mathcal{G}$, the proposed methodology assigns a vector of labels ℓ in L^{10}, where $L = \{\mathrm{N}, \mathrm{P}, \mathrm{D}\}$, in the following way.

1. Let $HC_{\mathcal{G}} = (H_{\mathcal{G}}, C_{\mathcal{G}}) \in H \times C$ the corresponding point to $\mathcal{G}$. For d denoting the euclidean distance and Γ the set of all $\mathcal{G}$ under study, let $S^1 = \{(d(HC_{\mathcal{G}}, c_N), d(HC_{\mathcal{G}}, c_P)) : \mathcal{G} \in \Gamma\}$. Consider $S^1_{\min} = \{\min(a,b) : (a,b) \in S^1\}$ and $S^1_{\max} = \{\max(a,b) : (a,b) \in S^1\}$. Then, the threshold η is fixed such that $\max(S^1_{\min}) \leq \eta \leq \min(S^1_{\max})$. In the present work, $\eta = 0.06$ (see Table 5). Thus,

$$\ell_1 = \begin{cases} \mathrm{N} & \text{if } d(HC_{\mathcal{G}}, c_N) < \min\{d(HC_{\mathcal{G}}, c_P), \eta\}, \\ \mathrm{P} & \text{if } d(HC_{\mathcal{G}}, c_P) < \min\{d(HC_{\mathcal{G}}, c_N), \eta\}, \\ \mathrm{D} & \text{otherwise.} \end{cases} \tag{9}$$

2. Let M^T be the OPTM of $\mathcal{G}$, and μ_P^T and μ_N^T be the average OPTMs of positive and negative cases, respectively, for the tied data treatment $T \in \{O, S, W\}$. Then, for $i = 2, 3, 4$ and a threshold η^T,

$$\ell_i = \begin{cases} \mathrm{N} & \text{if } \mathrm{dist}(M^T, \mu_N^T) < \min\{\mathrm{dist}(M^T, \mu_P^T), \eta^T\}, \\ \mathrm{P} & \text{if } \mathrm{dist}(M^T, \mu_P^T) < \min\{\mathrm{dist}(M^T, \mu_N^T), \eta^T\}, \\ \mathrm{D} & \text{otherwise,} \end{cases} \tag{10}$$

where $\eta^O = 0.9$, $\eta^S = \eta^W = 1$. The criterion for threshold selection is similar to the previous case. Let $S^2 = \{\big(\mathrm{dist}(M^T, \mu_N^T), \mathrm{dist}(M^T, \mu_P^T)\big) : \text{for all } M^T \text{ under study}\}$. Consider $S^2_{\min} = \{\min(a,b) : (a,b) \in S^2\}$ and $S^2_{\max} = \{\max(a,b) : (a,b) \in S^2\}$. Then, the threshold η^T must verify $\max(S^2_{\min}) \leq \eta \leq \min(S^2_{\max})$. The values used in the present work were obtained based on Table 6.

3. Let M^T be the OPTM of $\mathcal{G}$ for the tied data treatment $T \in \{O, S, W\}$. Then,

$$\ell_5 = \begin{cases} \text{N} & \text{if } \|M^O\|_F \leq 1.84, \\ \text{P} & \text{if } \|M^O\|_F > 1.85, \\ \text{D} & \text{otherwise,} \end{cases} \tag{11}$$

$$\ell_6 = \begin{cases} \text{N} & \text{if } \|M^S\|_F \leq 1.87, \\ \text{P} & \text{if } \|M^S\|_F > 1.89, \\ \text{D} & \text{otherwise,} \end{cases} \tag{12}$$

$$\ell_7 = \begin{cases} \text{N} & \text{if } \|M^W\|_F \leq 1.63, \\ \text{P} & \text{if } \|M^W\|_F > 1.64, \\ \text{D} & \text{otherwise,} \end{cases} \tag{13}$$

4. Let v^T be the stationary vector of the OPTM M^T of $\mathcal{G}$ for the tied data treatment $T \in \{O, S, W\}$. Then,

$$\ell_8 = \begin{cases} \text{N} & \text{if } \|v^O\| \leq 0.52, \\ \text{P} & \text{if } \|v^O\| > 0.53, \\ \text{D} & \text{otherwise,} \end{cases} \tag{14}$$

$$\ell_9 = \begin{cases} \text{N} & \text{if } \|v^S\| \leq 0.49, \\ \text{P} & \text{if } \|v^S\| > 0.50, \\ \text{D} & \text{otherwise,} \end{cases} \tag{15}$$

$$\ell_{10} = \begin{cases} \text{N} & \text{if } \|v^W\| \leq 0.48, \\ \text{P} & \text{if } \|v^W\| > 0.50, \\ \text{D} & \text{otherwise,} \end{cases} \tag{16}$$

Analogous to the threshold selection procedure, the limits in the last two cases described above, were obtained from Figure 3 in such a way both classes were properly separated.

In this preliminary stage of the work, the size of the data set used to define the models is small. However, the difference between the mean of the groups P and N is significant as can be shown using ANOVA (ANalysis Of VAriance). This statistical test can be applied under three hypotheses: independence of the data that is satisfied since the glucose is measured in different people, an

approximately normal distribution of the residuals verified by the Shapiro-Wilk test, and homoscedasticity or homogeneity of the variances proved by the Levene's test.

8. Results

The p-values of normality and homogeneity tests, shown in Table 4 are larger than 0.05, except for the Manhattan distance to μ_P^S. Thus, the null hypothesis is not rejected, in other words, the assumptions of ANOVA test are satisfied. The same table contains the p-values of the ANOVA test. It can be noticed that there is significant statistical evidence to reject the null hypothesis of ANOVA test saying that the means of both groups under the variables used to define the classifiers are considerably different. The confidence level of significance in all the statistical tests is 95%.

Table 4: p-values of statistical tests.

Variables			Shapiro-Wilk	Levene	ANOVA
Distance to centroids	cN		0.1179	0.1568	7.39×10^{-4}
	cP		0.7647	0.5801	1.69×10^{-4}
Distance to OPTM	O	N	0.3262	0.2471	5.78×10^{-5}
		P	0.7157	0.6897	3.13×10^{-6}
	S	N	0.5580	0.3927	3.90×10^{-5}
		P	0.0243	0.7342	3.07×10^{-5}
	W	N	0.6513	0.3827	1.43×10^{-5}
		P	0.4917	0.0706	2.04×10^{-5}
Frobenius norm of OPTM	O		0.8087	0.2798	5.49×10^{-2}
	S		0.8778	0.6941	4.62×10^{-2}
	W		0.3160	0.8434	2.08×10^{-2}
Norm of stationary vector	O		0.2456	0.2682	1.82×10^{-3}
	S		0.5055	0.2448	2.92×10^{-3}
	W		0.0730	0.1952	2.94×10^{-3}

Figure 2 shows the locations of the points in the $H \times C$ plane, for the considered glucose times series. It can be appreciated that all the points correctly lay in the region defined by the curves $C_{\min}$ and $C_{\max}$, for details about the computation of these limit curves see [21]. The glucose time series corresponding to negative cases are located in the bottom right corner. Meanwhile, for the positive cases, entropy decreases, and complexity increases. This behavior is observed in other disease processes reported in the literature (cf. [14].) The use of the causal plane has resulted in a grouping of the cases labeled as N and P that are in areas of this plane adequately differentiated as shown in Figure 2, so that a centroid can be established for each of the regions formed in order to associate the cases to distinguish them according to their proximity to these centroids. This last fact can be also seen in Table 5. All the N cases have a distance between 0.28% and 11.76%, from the centroid of the cases labeled as negative of the respective centroid of those labeled as P, which indicates that they would be closer to the N or non-diabetic cases. The opposite occurs with the values of the $H \times C$ plane corresponding to the measurements labeled P, which are between 0.35% and 29.19% closer to the centroid of the cases labeled as P than the remaining class.

Table 5: Distances to centroids in the $H \times C$ plane.

Case	$\boldsymbol{c_N}$	$\boldsymbol{c_P}$
N1	0.0075	0.1335
N2	0.0188	0.1598
N3	0.0112	0.1299
N4	0.0004	0.1409
P1	0.1128	0.0284
P2	0.1113	0.0298
P3	0.1411	0.0005
P4	0.1990	0.0581

Table 6 exhibits the Manhattan distances between every case of study and the average OPTM of each group, considering the three tied data treatments introduced in Section 6. It is worth noting that in all the cases under analysis, the distance is considerably smaller for the average matrix corresponding to the group of belongingness, independent of how tied data is dealt with.

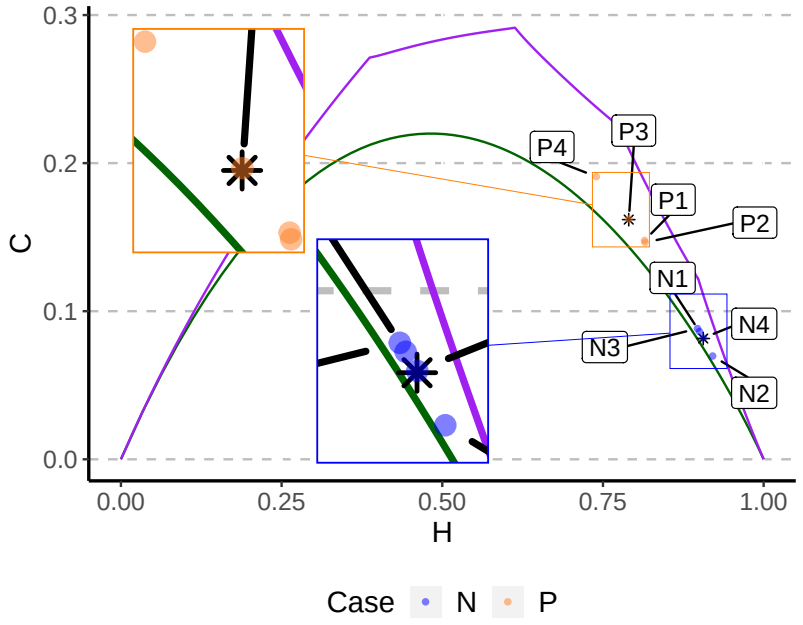

Figure 2: Location of glucose time series in the $H \times C$ plane together with the centroids of each group (asterisks).

Table 6: Manhattan distances to mean transition matrices for the three tied data treatment.

Case	$\boldsymbol{\mu_N^O}$	$\boldsymbol{\mu_P^O}$	$\boldsymbol{\mu_N^S}$	$\boldsymbol{\mu_P^S}$	$\boldsymbol{\mu_N^W}$	$\boldsymbol{\mu_P^W}$
N1	0.6221	1.9196	0.4294	1.6634	0.6097	1.8255
N2	0.5674	1.7989	0.4882	1.4966	0.5367	1.7838
N3	0.5794	1.7793	0.4806	1.4470	0.5574	1.7488
N4	0.2909	1.7407	0.2964	1.5386	0.3902	1.7349
P1	1.9708	0.6683	1.5459	0.3513	1.7810	0.4735
P2	1.6215	0.6620	1.2970	0.3807	1.5532	0.6708
P3	1.7265	0.4728	1.5886	0.3526	1.9683	0.5034
P4	2.1053	0.7542	1.7567	0.7038	1.8499	0.8784

The Frobenius norms of the OPTMs together with the norms of the associated stationary vectors for the eight cases under study are shown in Figure 3. Both groups are well-separated in all tied data treatments when the norm of the stationary vector of the OPTM is considered. This separation is more evident when weight is assigned to tied data. When the Frobenius norm of the OPTM is used, it can be noticed that there exists an overlapping in both groups. It can be observed that the norm of the stationary vector obtained from the dynamic evolution of glucose is a good indicator of diabetic disease.

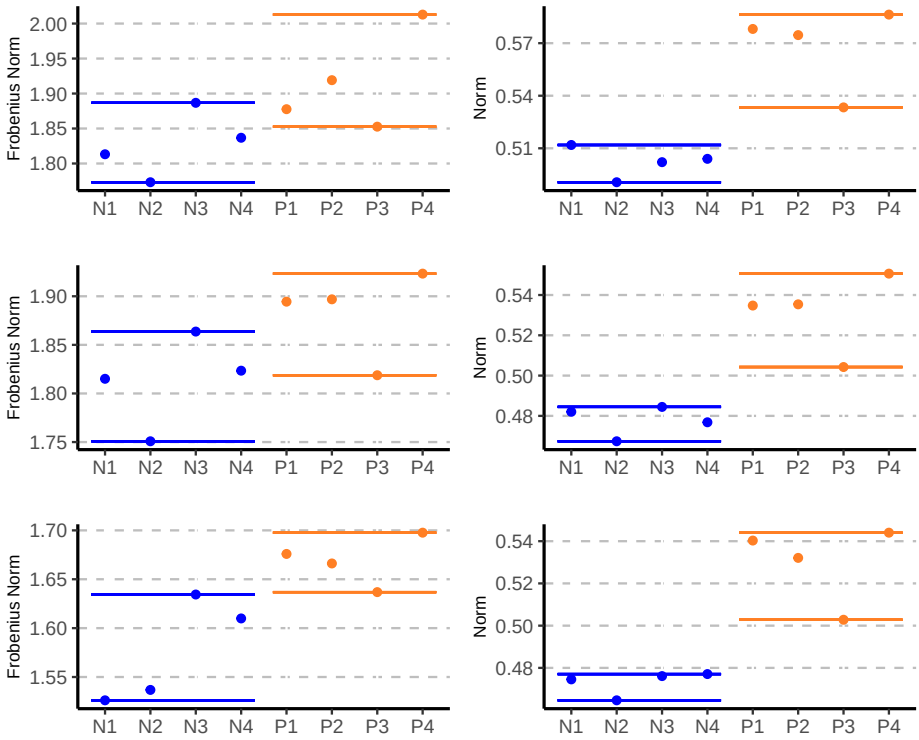

Figure 3: Frobenius norms of the OPTMs (left) and norms of the associated stationary vector (right), for the following tied data treatments: omission (top), sequential order (middle), and assigned weights (bottom).

All the label vector ℓ entries are N for the declared negative cases, except for N3 for which $\ell_7 = \mathrm{D}$. On the other hand, a similar behavior holds for the cases with a positive diagnosis, for which the components of ℓ are P other than $\ell_2 = \mathrm{D}$ for P4.

Finally, the methodology presented in this work is validated in two ways. On one hand, six glucose time series obtained by the same type of sensor are classified: (i) X corresponds to a female person 43 years old with no evidence of any wealth problem; (ii) Y corresponds to a male subject 49 years old with diagnostic diabetes that is controlled, both with diet and physical activity. (iii) Z shows the glucose values of a male person 59 years old obese for a long period of time; (iv) W belongs to the same male individual Z but after a period under a strict diet and physical activity practice. (v) V corresponds to a male person 59 years old with diagnosed hypertension without any hint about his diabetic state. (vi) U corresponds to a male subject 61 years old with diagnostic diabetes. The results are presented in Table 7. The first row of this Table exhibits the case of a young woman with no diabetes diagnosis, as can see all the components of the vector L^{10} result in a classification in the N class, in such a way that agrees with the clinical evidence. The last row or the same table, that corresponds to a diabetic patient, shows all the components of the L^{10} vector classifying in the group of P, except for the $\boldsymbol{\ell_2}$ which results in doubtful. Cases X and U, corresponding to the top and bottom row of Table 7, represent the classification most clearly differentiated in new samples, evidencing that this preliminary study works fine. The cases Z and W under strict control and physical exercise show that the diabetic condition becomes improved since the components of the L^{10} vector reflect this change. Explicitly, $\boldsymbol{\ell_3}$ and $\boldsymbol{\ell_7}$ changes from P to N, $\boldsymbol{\ell_5}$ and $\boldsymbol{\ell_{10}}$ changes from P to D, $\boldsymbol{\ell_6}$ turns on from D to N, and the rest of the components remain unchanged.

Table 7: Results of the proposed methodology applied to new observations.

Case	$\boldsymbol{\ell_1}$	$\boldsymbol{\ell_2}$	$\boldsymbol{\ell_3}$	$\boldsymbol{\ell_4}$	$\boldsymbol{\ell_5}$	$\boldsymbol{\ell_6}$	$\boldsymbol{\ell_7}$	$\boldsymbol{\ell_8}$	$\boldsymbol{\ell_9}$	$\boldsymbol{\ell_{10}}$
X	N	N	N	N	N	N	N	N	N	N
Y	N	P	P	P	P	N	D	P	P	P
Z	N	D	P	D	P	D	P	P	P	P
W	N	D	N	D	D	N	N	P	P	D
V	N	D	N	N	N	N	N	D	N	D
U	P	D	P	P	P	P	P	P	P	P

Finally, an index ν to quantify the possibility of DMT2 risk was constructed as follows: in the first place is the stage of the selection for the ℓ best components. To accomplish this selection was picked ℓ_1 because it reflects

the dynamic behavior originating the signal, then ℓ_3 using the sequential tied treatment of the data, due to the minimum values of the Manhattan distance to the means of each class (see the center columns in Table 6). The combination of Frobenius norm and weighted tied approach was used to select ℓ_7 using the results shown in Figure 3, bottom left, and finally to the choice of ℓ_{10}, the best tied data treatment for the computation of the stationary vector was selected from the results in Figure 3 bottom right. Then, for the construction of the index is necessary to quantify the states indicated in Table 7. The criterium adopted is based on the assigning function $\mathcal{A} : \{N, D, P\} \to \mathbb{R}$ defined by $\mathcal{A}(N) = 0$, $\mathcal{A}(D) = 0.25$, and $\mathcal{A}(P) = 0.50$. Under these considerations, the index is mathematically formulated by computing the euclidean norm of the vector $v_{\mathcal{A}} = (\mathcal{A}(\ell_1), \mathcal{A}(\ell_3), \mathcal{A}(\ell_7), \mathcal{A}(\ell_{10}))$. In other words,

$$\nu = \|v_{\mathcal{A}}\|_2 = \sqrt{[\mathcal{A}(\ell_1)]^2 + [\mathcal{A}(\ell_3)]^2 + [\mathcal{A}(\ell_7)]^2 + [\mathcal{A}(\ell_{10})]^2}. \tag{17}$$

It is worth noticing that ν ranges from 0 (negative case detected by the four classifiers) to 1 (positive case detected by the four classifiers). Moreover, when the four classifiers detect a doubtful case, $\nu = 0.50$.

The results of the indices obtained when the proposal is applied to the new observations are presented in Table 8.

Table 8: Index proposed to classify new observations.

Case	$\mathcal{A}(\ell_1)$	$\mathcal{A}(\ell_3)$	$\mathcal{A}(\ell_7)$	$\mathcal{A}(\ell_{10})$	ν
X	0	0	0	0	0
Y	0	0.50	0.25	0.25	0.61
Z	0	0.50	0.50	0.50	0.87
W	0	0	0	0.25	0.25
V	0	0	0	0.25	0.25
U	0.50	0.50	0.50	0.50	1.00

9. Discussion and Conclusions

The present pilot study has fulfilled its fundamental aim, which is to show the possibility of detecting people suffering from DMT2 in terms of interstitial blood glucose records by means of the original use of dynamical systems and information theory tools. The advantage of having a database

of our own, although small in size, provided the traceability of the records as well as the quality of the data. In addition, it was accompanied by the corresponding medical diagnosis.

Ordinal pattern distributions were calculated, using the Bandt and Pompe approach, and then used to find the corresponding coordinates in the entropy-complexity plane for each record. This allowed the discrimination of the presence or absence of DMT2 since two well-differentiated groups in this plane were formed. The essential differences that can be seen in Figure 1, are that in subjects with DMT2, the mechanism of blood glucose regulation after food intake is characterized by oscillations of greater amplitude compared with the cases of subjects without DMT2. Since entropy and complexity are two variables considered from a global point of view of the signal, the probabilities of transitions between ordinal patterns of the signals were also calculated. This was done to study a possible improvement in the characterization of the dynamic change that occurs in the stabilization of blood glucose in subjects without DMT2. The use of matrix metrics applied to the OPTMs provides different features of these matrices. Thus, the Frobenius norm was used to individually evaluate the intensity of the transitions for each time series obtained from the interstitial sensor measurements. On the other hand, the signals were characterized by calculating the distance between the different OPTMs and their mean values with respect to each class of interest. In this sense, the aim was to evaluate the intensity of the probabilities of the transitions of the different time series that constitute the database with respect to a reference value.

Finally, the norms of the stationarity vectors of the OPTMs were calculated in order to quantify the way in which the OPTMs characterize the time series.

In all cases, both between and within measures from OPTMs have shown to be a channel of differentiation between records from subjects with or without DMT2. The metrics calculated from the OPTMs were sufficiently different to provide strong differentiation thresholds. This fact motivated the definition of a unique value to characterize the records. Then, the introduction of the index ν in the classification of the new subjects results in a clear classification, in such a way that people without DMT2 have $\nu = 0$, the subjects suffering from the disease have an index near to 1, and the doubtful cases have an index rounded 0.5. Following the values in Table 8, the subject X has a null index agreeing with the absence of DMT2. In the other extreme, subject U with DMT2 has a unitary index. Meanwhile, for subject Z, who is

diagnosticated with DMT2, and subject W, who is the same subject after a medium period of food control and physical exercise, the values of ν present a possible improvement of the disease under consideration. Subject Y is a patient with DMT2 for which $\nu = 0.61$, a value very far from the nondiabetic cases. Finally, $\nu = 0.25$ corresponds to subject V who has a hypertensive diagnosis but not DMT2. In this sense, the possibility of having an index is particularly useful in medical practice since professionals usually work with this type of indicator.

A fact that is well-known in the literature is that in the use of ordinal patterns, the treatment of repeated data is not indistinct. In the present work, the results clearly showed the effect that various approaches produce on the subsequent classification.

Regarding the selected components of the ℓ vector, the simplicity and computational efficiency of all the calculations necessary to obtain it can be highlighted.

As expected in pilot studies, the present work establishes a possible methodology to assist in the analysis of a large survey. For the posterior data analysis, any of the conceptual tools from the dynamical systems and mathematics applied in this work were exhibited as a potential way to use.

Specifically, the proposal of this work is the construction of macroscopic variables, based on blood glucose measurement, that reflects the microscopic processes occurring in the human organism. This is a treatment similar to the thermodynamic formulation of kinetic theory and statistical mechanics. In physics, the Clausius-Clapeyron equation and the Gibbs equation for chemical reactions are good examples of information that can be extracted from macroscopic formalisms. In the treatment presented, the application of medical recovery standards is immediately detected by the proposed algorithms. As a conceptual summary, this preliminary research shows a macroscale analysis, so that in the future it would be possible to derive useful information about the dynamics of the system.

Fundings

This work was partially financed by the Universidad Tecnológica Nacional, GRANT PID 8120.

Conflics of Interest

The authors declare no conflicts of interest.

References

[1] Eugene Ackerman, John W. Rosevear, and Warren F McGuckin. A mathematical model of the glucose-tolerance test. *Physics in Medicine & Biology*, 9(2):203, 1964.

[2] José M. Amigó, Karsten Keller, and Valentina A. Unakafova. Ordinal symbolic analysis and its application to biomedical recordings. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2034):20140091, 2015.

[3] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical Review Letters*, 88(17):174102, 2002.

[4] Luís Barreira. Lyapunov exponents and regularity. In *Lyapunov Exponents*, pages 31–41. Springer, 2017.

[5] Sofia Benbelkacem and Baghdad Atmani. Random forests for diabetes diagnosis. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–4. IEEE, 2019.

[6] Richard N. Bergman. Minimal model: perspective from 2005. *Hormone Research in Paediatrics*, 64(Suppl. 3):8–15, 2005.

[7] Cristian Bonini, Andrea Rey, Dino Otero, Ariel Amadio, Manuel Blesa, and Walter Legnani. An alternative computation of the entropy of 1D signals based on geometric properties. *Statistics, Optimization & Information Computing*, 10(4):998–1020, 2022.

[8] Elizabeth Bradley and Holger Kantz. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610, 2015.

[9] George F. Cahill Jr, Donnell D. Etzwiler, and Norbert Freinkel. "Control" and diabetes. *New England Journal of Medicine*, 294(18):1004–1005, 1976.

[10] Isadora Cardoso-Pereira, João B. Borges, Pedro H. Barros, Antonio F. Loureiro, Osvaldo A. Rosso, and Heitor S. Ramos. Leveraging the self-transition probability of ordinal patterns transition network for trans-

portation mode identification based on GPS data. *Nonlinear Dynamics*, 107(1):889–908, 2022.

[11] Enrico Carmina, Frank Z. Stanczyk, and Rogerio A. Lobo. Evaluation of hormonal status. In *Yen and Jaffe's reproductive endocrinology*, pages 887–915. Elsevier, 2019.

[12] Rafael Garcia-Carretero, Luis Vigil-Medina, Inmaculada Mora-Jimenez, Cristina Soguero-Ruiz, Oscar Barquero-Perez, and Javier Ramos-Lopez. Use of a k-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. *Medical & Biological Engineering & Computing*, 58(5):991–1002, 2020.

[13] Omid Ghorbanzadeh, Thomas Blaschke, Khalil Gholamnia, Sansar Raj Meena, Dirk Tiede, and Jagannath Aryal. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2):196, 2019.

[14] Ary L. Goldberger, Luis A.N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen C. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[15] Charles J. Goodner, Barbara C. Walike, Donna J. Koerker, John W. Ensinck, Arthur C. Brown, Elliott W. Chideckel, Jerry Palmer, and Lynne Kalnasy. Insulin, glucagon, and glucose exhibit synchronous, sustained oscillations in fasting monkeys. *Science*, 195(4274):177–179, 1977.

[16] Manu Goyal, Neil D. Reeves, Adrian K. Davison, Satyan Rajbhandari, Jennifer Spragg, and Moi Hoon Yap. Dfunet: Convolutional neural networks for diabetic foot ulcer classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(5):728–739, 2018.

[17] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University Press, 1991.

[18] Inmaculada Leyva, Johann H. Martínez, Cristina Masoller, Osvaldo A. Rosso, and Massimiliano Zanin. 20 years of ordinal patterns: Perspectives and challenges. *Europhysics Letters*, 138(3):31001, 2022.

[19] Rian Budi Lukmanto, Ariadi Nugroho, Habibullah Akbar, et al. Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. *Procedia Computer Science*, 157:46–54, 2019.

[20] Ricardo López-Ruiz, Héctor Mancini, and Xavier Calbet. A statistical measure of complexity. *Physics Letters A*, 209(5-6):321–326, 1995.

[21] María Martin, Ángel Plastino, and Osvaldo A. Rosso. Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A: Statistical Mechanics and its Applications*, 369(2):439–462, 2006.

[22] Diego M. Mateos, Jaime Gómez-Ramírez, and Osvaldo A. Rosso. Using time causal quantifiers to characterize sleep stages. *Chaos, Solitons & Fractals*, 146:110798, 2021.

[23] Shlomo Melmed, Ronald Koenig, Clifford Rosen, Richard Auchus, and Allison Goldfine. *Williams textbook of endocrinology: South Asia edition, 2 vol set-E-book*. Elsevier India, 2020.

[24] A. Mary Posonia, S. Vigneshwari, and D. Jamuna Rani. Machine learning based diabetes prediction using decision tree J48. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 498–502. IEEE, 2020.

[25] Osvaldo A. Rosso, Felipe Olivares, Luciano Zunino, Luciana De Micco, André L.L. Aquino, Angelo Plastino, and Hilda A. Larrondo. Characterization of chaotic maps using the permutation Bandt-Pompe probability distribution. *The European Physical Journal B*, 86(4):1–13, 2013.

[26] Eugene Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[27] Jin Shi, Eric S. McLamore, David Jaroch, Jonathan C. Claussen, Raghavendra G. Mirmira, Jenna L. Rickus, and D. Marshall Porterfield. Oscillatory glucose flux in ins 1 pancreatic β cells: A self-referencing microbiosensor study. *Analytical Biochemistry*, 411(2):185–193, 2011.

[28] Cas Weykamp. HbA1c: a review of analytical and clinical aspects. *Annals of Laboratory Medicine*, 33(6):393–400, 2013.

[29] Massimiliano Zanin, Luciano Zunino, Osvaldo A. Rosso, and David Papo. Permutation entropy and its main biomedical and econophysics applications: a review. *Entropy*, 14(8):1553–1577, 2012.

[30] Dongmei Zhao. *Power Distribution and Performance Analysis for Wireless Communication Networks*. Springer Science & Business Media, 2012.

[31] Changsheng Zhu, Christian Uwa Idemudia, and Wenfang Feng. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17:100179, 2019.