Classifying Healthy and Preeclamptic Patients from Recurrence-Based Cardiovascular Time Series Using Complex Networks Methods

G.M. Ramírez Ávila, A. Gapelyuk, N. Marwan, H. Stepan, J. Kurths, Th. Walther and N. Wessel

***Abstract*—We analyze cardiovascular time series with the aim of identifying patients suffering of preeclampsia, a pregnancy-specific disorder causing maternal and fetal morbidity and mortality. For that, we use a novel approach, namely the ε-recurrence networks applied to a phase space constructed by means of the time series of the variabilities of the heart rate, and the blood pressure (systolic and diastolic). Four network measures are considered as parameters for our analysis: average path length, mean coreness, global clustering coefficient, and scale-local transitivity dimension. With these quantities, we perform a quadratic discriminant analysis. This allows us to classify healthy and preeclamptic patients with a sensitivity of 91.7% and a specificity of 68.1%, thus validating the use of this method.**

***Index Terms*—blood flow in cardiovascular system, cardiac dynamics, hemodynamics, networks, time series analysis.**

I. Introduction

Preeclampsia (PE) is a major hypertensive disorder in pregnant women also characterized by proteinuria for which the pathophysiology remains unclear and constitutes a serious risk for both the mother and the fetus. PE affects healthy nulliparous women in a range between 2% and 7% worldwide [1]. Several strategies are used in order to predict PE, among them we can mention some biochemical markers, such as fms-like tyrosine kinase 1 (sFlt-1), placental growth factor (PlGF), soluble endoglin [2], [3], maternal autoantibody, the angiotensin II type I receptor agonistic autoantibody (AT1-AA) [4], the urinary biomarkers [5], noninvasive CV markers [6] or the combination of some of those [7].

Detection of cardiovascular (CV) disorders has been considerably improved due to both technological advances and new methods of time series analysis. Nevertheless, there are still difficulties that cannot be explained by standard data analysis. Nonlinear data analysis and modeling methods of CV physics allow to improve clinical diagnostics and also a better understanding of CV regulation. One of the most important aspects of these methods is that they focus on noninvasive measured biosignals. Among the biosignals that CV physics deals with are the heart rate variability (HRV), the variabilities of systolic blood pressure (SBPV) and diastolic blood pressure (DBPV), and the baroreflex sensitivity (BRS).

Recurrence methods have recently become an useful tool in order to study time series and acquired importance because they do not need long time series to identify transitions in dynamical systems and their use may be applied to a wide diversity of systems and phenomena. Recently, the recurrence concept has been extended to networks and novel time series analysis methods arose [8]. In this work we apply the approach of ε-recurrence networks to analyze noninvasive CV markers with the aim of developing a classification method to identify healthy subjects (control) from patients who develop PE.

N. Wessel, J. Kurths, A. Gapelyuk, and G.M. Ramírez Ávila are with Humboldt-Universität zu Berlin, Department Of Physics, Germany e-mail: wessel@physik.hu-berlin.de

N. Marwan is with Potsdam Institute for Climate Impact Research, Germany

H. Stepan is with Department of Obstetrics and Gynecology, University of Leipzig, Germany

Th. Walther is with Centre for Cardiovascular and Metabolic Research, University of Hull, United Kingdom

II. Methods

A. Clinical aspects

We consider for this study 96 patients with abnormal uterine perfusion (AUP), followed by means of Doppler sonography in the second trimester, between the 18th and the 26th week of gestation (WOG) of pregnancy, at the Department of Obstetrics and Gynecology of the University of Leipzig. Immediately after the Doppler examination, noninvasive continuous blood pressure monitoring was conducted via finger cuff during 30 minutes. The continuous blood pressure curves were used to extract the time series of beat-to-beat intervals, systolic and diastolic blood pressures allowing us to obtain the CV markers (HRV, SBPV, and DBPV). The length of the dataset per variable is roughly 1600. At the time of examination, the women were healthy, normotensive, without clinical signs of cervical incompetence, and on no medication. After the 30th WOG, 24 patients developed PE. Further details on the methodology can be found in [6].

B. Recurrence networks

The basic idea of time series analysis based on complex network techniques lies on the fact that a time series might be transformed into a complex network from which we can extract the adjacency matrix allowing us to obtain local and global network properties. The concept of recurrence applied to a single trajectory of the dynamical system allows us to obtain the recurrence matrix whose elements are given by $R_{i,j} = \Theta(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|)$, where $\Theta(\cdot)$ represents the Heaviside function, $\|\cdot\|$ is a suitable norm, and ε is a threshold distance that should be chosen adequately according to the characteristics of the embedded attractor into the phase space. We interpret the recurrence matrix $\mathbf{R}$ as the adjacency matrix

of an unweighted and undirected complex network, commonly called the ε-recurrence network which is associated with a given time series. Possible self-loops must be avoided in this network, thus a Kronecker delta must be subtracted from the recurrence matrix and as a consequence, the elements of the adjacency matrix for an ε-recurrence network are:

$$A_{i,j}(\varepsilon) = R_{i,j}(\varepsilon) - \delta_{i,j}, \tag{1}$$

where the ε-dependence is considered explicitly. There is not a universal criterion for choosing ε but the choice must be made avoiding too small values that lead to a situation in which there are not enough recurrence points, or too large values implying that every vertex is connected with many other vertices irrespective of their actual mutual proximity in phase space [9]. Having reconstructed the adjacency matrix $\mathbf{A}$ from a time series, we can apply appropriate networks characteristics to analyze and obtain information of the underlying system. In this work we focus our interest in four global network measures: the *Average path length* ($\mathcal{L}$), that is the mean value of the shortest geodetic path lengths $l_{i,j}$ considering all pair of vertices (i, j); the *Mean coreness* ($\mathcal{C}\imath$), that is the average of the corenesses (significance of a node and its "popularity" in the network) of all the vertices [10]; the *Global clustering coefficient* ($\mathcal{C}$), that is the average of the clustering coefficient of each vertex (ratio of triangles including vertex i and the number of triples centered on vertex i where triple refers to a pair (j, k) of vertices that are both linked with i, but not necessarily mutually linked); and the *Scale local transitivity dimension* ($\mathrm{D}_{\mathcal{T}}$), defined as $\mathrm{D}_{\mathcal{T}} = \frac{\log \mathcal{T}}{\log(3/4)}$, being $\mathcal{T}$ the transitivity (ratio of the number of triangles in the network times three and the number of linked triples of vertices). These four measures depend on ε and have a global character. A detailed description of networks and their properties can be found in [11].

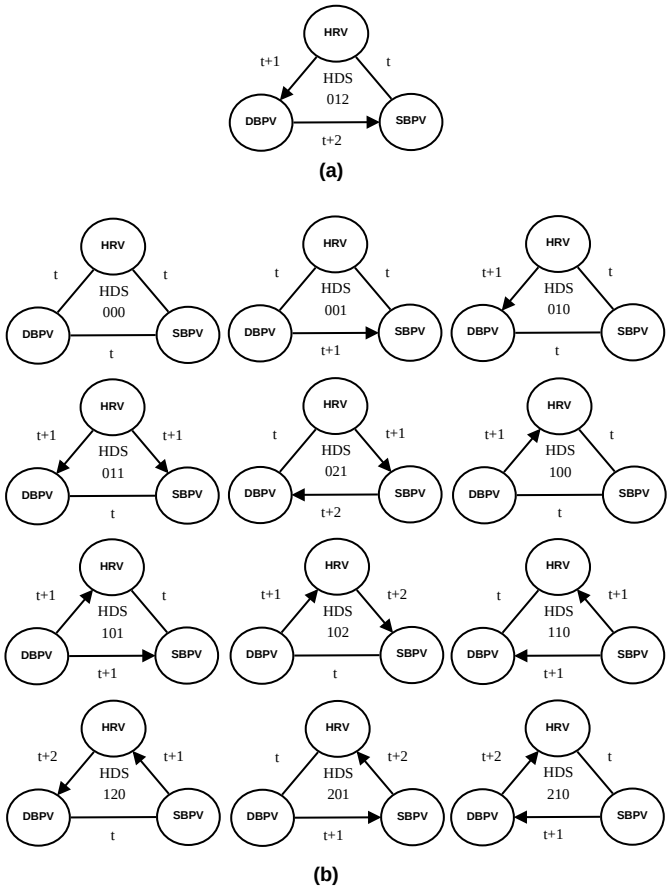

Fig. 1. (a) Coupling structure considering that HRV drives the DBPV and this in turn the SBPV (directed arrows from HRV to DBPV, and from DBPV to SBPV). Note that when the variables are linked only for a line, it means that these are coupled but without any delay. This might be written schematically as $H(t)D(t+1)S(t+2) \equiv 012$; the latter number can change according to the delay among the sequential variables HRV, DBPV, and SBPV, represented as HDS. (b) All the other possibilities of coupling structures.

C. Data processing and statistics

In order to avoid artifacts such as double recognition of beats, the original RR time series were filtered using a preprocessing algorithm which first removes obvious recognition errors; then applies an adaptive percent filter, and finally, an adapting controlling filter [12]. With the aim of using a recurrence network approach, we consider the three CV markers and some possible embeddings. An estimation of the coupling structure of CV markers has been performed using nonlinear additive autoregressive models with external input following the idea of Granger causality [13]. This coupling analysis shows that HRV, DBPV, and SBPV respond to respiration; SBPV respond to DBPV and the latter to HRV. In our case, we do not consider respiration; thus, the coupling structure might be represented as in Fig. 1(a) where according to the coupling scheme, there is a delay between the HRV, the DBPV, and the SBPV. For simplicity we write down the coupling structure as (HRV(t),DBPV$(t+1)$,SBPV$(t+2)$) or simply $H(t)D(t+1)S(t+2) \equiv 012$.

We sought to predict whether or not a patient develops PE using the CV markers embedded in a phase space determined by the structure of coupling. We consider a minimalist assumption in which the structure of coupling between HRV, DBPV, and SBPV is equal in each subject of a group and that this structure does not change during the measurement. In this study, we set out to test all the possible structures of coupling shown in Fig. 1 and a wide range of the threshold ε going from 0.01σ to 0.99σ being σ the standard deviation of the underlying process in the embedded phase space. From a simple CV time series corresponding to each patient, we construct a complex network for each possible structure of coupling and each value of ε. Then, we compute the four network measures: ($\mathcal{C}$,$\mathcal{L}$,$\mathcal{C}\imath$,$\mathrm{D}_{\mathcal{T}}$), and with these new measures we perform an analysis to classify the groups of patients. For that purpose, we firstly verify whether or not these new parameters are significant by means of a Mann-Whitney *U*-test and considering a significance level of 5%; being the null hypothesis that data in the vectors corresponding to control and preeclamptic patients are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians.

III. RESULTS

As the approach is based on recurrence complex networks, firstly we obtain the matrices $\mathbf{R}$ and $\mathbf{A}$. A visualization of

the associated networks, obtained using the medians of the time series are shown in Fig. 2. These representations are constructed using the coordinates of the nodes. An inspection of these networks (PE and control) allows us to perceive some differences between them as for example the existence of more free nodes (more outliers from a statistical point of view) in the case of the control group network compared to the PE group network, and the apparent node degree that seems to be higher in the control group network. Nonetheless, this visualization inspection is just a first checkup that cannot replace the quantification of the network measures.

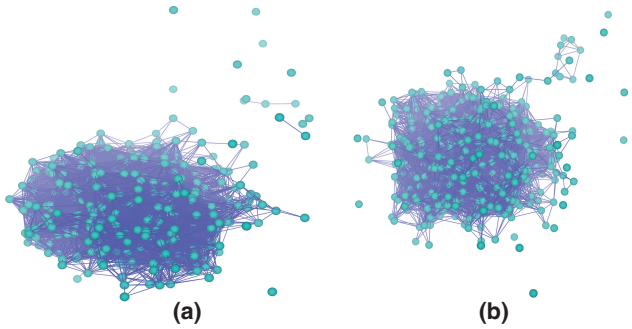

Fig. 2. (Color online) Visualization of the networks obtained using the time series of the medians for both groups of patients (a) PE, and (b) control. The visualization has been obtained by means of the software Pajek [14], with a 3-dim perspective and using all the nodes and their corresponding coordinates into the phase space $HDS(t)$.

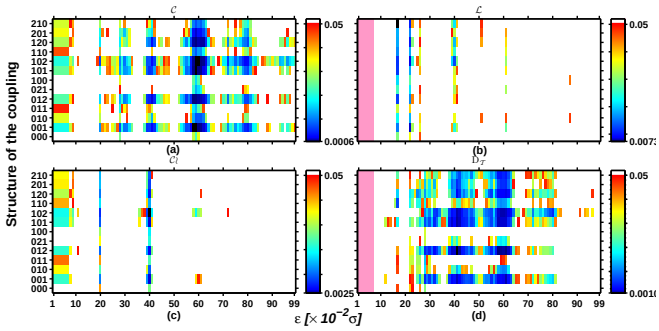

Fig. 3. (Color online) Phase plane structure of the coupling vs. ε showing the significance level p computed by means of a Mann-Whitney U-test for establishing differences between the control and PE groups and using the network measures (a) $\mathcal{C}$, (b) $\mathcal{L}$, (c) $\mathcal{C}l$, and (d) $\mathrm{D}_{\mathcal{T}}$. The color code indicates the p-values. Notice that some special pixel are used such as white ($p \geq 0.05$; H_0 cannot be rejected), pink (it is not possible to compute p-value; thus, the p-value is undetermined), and black (minimum p-value).

The results for each network measure are represented in the phase plane, embedding (structure of coupling) vs. ε as shown in Fig. 3. The color code indicates the p-values of the statistical test when the null hypothesis H_0 of equal medians at 5% significance level is rejected. The white pixels denote that there is no difference between both groups ($p \geq 0.05$), and pink ones, the impossibility to compute p. On the contrary, the black pixels represent the minimum p-value among all the possibilities on the phase plane.

According to Fig. 3, the significant values for each network measure occur only for some coupling structures and thresholds ε. Fig. 4 shows the same plane as in Fig. 3 but considering the cases in which all the four network measures are simultaneously significant, i.e. $p < 0.05$ (black pixels).

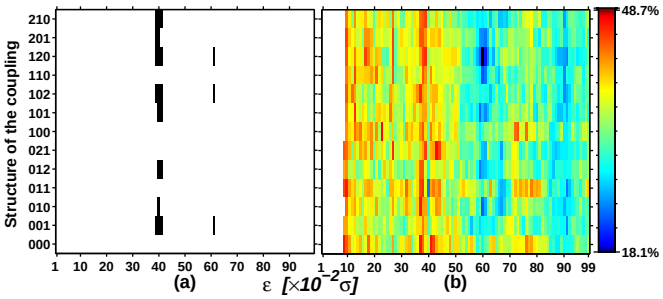

Fig. 4. (Color online) Same phase plane as in Fig. 3 showing the situations in which (a) the four considered network measures satisfy simultaneously the condition $p < 0.05$ (black pixels). (b) Misclassification errors (color code) in the classification of control and PE groups after a quadratic discriminant analysis for the four network measures. The white pixels indicate that the discriminant analysis cannot be performed and it is related to the fact that for these cases, at least one of the network measures has an undetermined p-value. The black pixel indicates the minimum value of the error.

The inspection of Fig. 4(a) shows that there are 22 situations in which the four network measures satisfy simultaneously the statistical significance test and we further restrict the analysis to these selected cases which do not necessary correspond to the lower p-values. Now, considering these four measures as the parameters for the classification of control and PE groups, we perform a quadratic discriminant analysis for all the possible structures of the coupling and ε (Fig. 4(b)).

Table. I shows the statistical measures of the performance of a binary classification test for the 22 selected cases. Such measures are misclassification error rate (percentage of observations that are misclassified), sensitivity (proportion of true positives that are correctly identified by the test), specificity (proportion of true negatives correctly identified by the test), positive predictive value (PPV), i.e. the proportion of patients with positive test results who are correctly diagnosed, and negative predictive value (NPV), i.e. the proportion of patients with negative test results who are correctly diagnosed.

From Table I, we select the situation corresponding to a coupling structure 120 and $\varepsilon = 0.61\sigma$ (bold fonts) whose misclassification error is 20.1% giving consequently the best values for the classification results, i. e. a sensitivity of 91.7%, a specificity of 68.1%, a PPV of 48.9%, and a NPV of 96.1%.

IV. Conclusion

The essential aspect of the approach used in this work lies in its novelty when applying to CV signals, i.e. complex biosignals time series that in their raw form are not useful for classification, are transformed into recurrence networks from which we extract several measures that allow a classification with suitable results. In fact, after the choice of an adequate structure of the coupling and the threshold ε, only one complex network is constructed from the three CV markers for each person and then, we quantify the network features that constitute the parameters for the classification analysis. In summary, our exploratory results show that the used approach constitutes a useful tool to study such a classification problem.

Note that the analysis presented here is in some sense only a first approximation of the recurrence networks approach. We see for future research several ways of improvements,

TABLE I
STATISTICAL MEASURES OF THE PERFORMANCE OF A BINARY CLASSIFICATION TEST CONSIDERING THE 22 POSSIBLE SITUATIONS IN WHICH THE FOUR NETWORK MEASURES SATISFY SIMULTANEOUSLY THE CONDITION $p < 0.05$.

Structure of coupling	ε [$\times\sigma$]	Misclassification [%]	Sensitivity [%]	Specificity [%]	PPV [%]	NPV [%]
001	0.39	40.3	79.2	40.3	30.7	85.3
	0.40	31.2	79.2	58.3	38.78	89.4
	0.41	34.0	89.5	44.4	34.4	91.4
	0.61	29.2	79.2	62.5	41.3	90.0
010	0.40	28.5	83.3	59.7	40.8	91.5
012	0.40	29.2	83.3	58.3	40.0	91.3
	0.41	36.8	87.5	38.9	32.3	90.3
101	0.40	34.0	83.3	48.6	35.1	89.7
	0.41	38.2	87.5	36.1	31.3	89.7
102	0.39	41.0	79.2	38.9	30.2	84.8
	0.40	34.7	83.3	47.2	34.5	89.5
	0.41	40.3	83.3	36.1	30.3	86.7
	0.61	29.2	83.3	58.3	40.0	91.3
120	0.39	45.8	79.2	29.2	27.1	80.8
	0.40	34.0	91.7	40.3	33.9	93.6
	0.41	36.1	91.7	36.1	32.3	92.9
	0.61	**20.1**	**91.7**	**68.1**	**48.9**	**96.1**
201	0.39	42.4	83.3	31.9	29.0	85.2
	0.40	38.2	75.0	48.6	32.7	85.4
210	0.39	46.1	79.2	30.6	27.5	81.5
	0.40	32.6	87.5	47.2	35.6	91.9
	0.41	36.1	87.5	40.3	32.8	90.6

as explained in the following. In spite of the minimalist assumptions concerning the structure of the coupling, and just one value of ε in order to avoid the ambiguities stated in [15], our results give useful information for the classification and are similar to those obtained in [6], thus validating our method. The consideration of dynamic structures of the coupling (i.e. temporal variations in the coupling structure) could improve our results and also give us a deeper insight in the underlying physiological processes. For that, it is necessary to design an adaptive algorithm taking into account possible transitions in several time windows. This method could be also useful as an alternative to find the adequate structure of coupling.

The consideration of other qualitative aspects related to the history of the patients (age, ethnicity, body mass index) [16] and in general predisposing factors such as genetic [17], behavioral [18] or environmental [19] could give additional information to improve the classification analysis combined with the technique used in this paper.

Our study follows the same line as previous works [6] in which the biosignal analysis (in our case, the associated recurrence complex network analysis) constitutes a noninvasive, cheap and universal diagnostic approach whose utilization offers new possibilities both in the understanding of PE pathogenesis and on the envisaging of new therapeutic strategies.

REFERENCES

[1] B. Sibai, G. Dekker, and M. Kupferminc, *Pre-eclampsia*, Lancet. 365(9461):785–99, 2005.

[2] S. Rana, S. A. Karumanchi, R. J. Levine, S. Venkatesha, J. A. Rauh-Hain, H. Tamez, and R. Thadhani, *Sequential changes in antiangiogenic factors in early pregnancy and risk of developing preeclampsia*, Hypertension. 50(1):137–42, 2007.

[3] A. Ohkuchi, C. Hirashima, S. Matsubara, K. Takahashi, Y. Matsuda, and M. Suzuki, *Threshold of soluble fms-like tyrosine kinase 1/placental growth factor ratio for the imminent onset of preeclampsia*, Hypertension. 58(5):859–66, 2011.

[4] A. H. Siddiqui, R. A. Irani, S. C. Blackwell, S. M. Ramin, R. E. Kellems, and Y. Xia, *Angiotensin receptor agonistic autoantibody is highly prevalent in preeclampsia*, Hypertension. 55(2):386–93, 2010.

[5] D. M. Carty, J. Siwy, J. E. Brennand, P. Zürbig, W. Mullen, J. Franke, J. W. McCulloch, R. A. North, L. C. Chappell, H. Mischak, L. Poston, A. F. Dominiczak, and C. Delles, *Urinary proteomics for prediction of preeclampsia*, Hypertension. 57(3):561–9, 2011.

[6] H. Malberg, R. Bauernschmitt, A. Voss, T. Walther, R. Faber, H. Stepan, and N. Wessel, *Analysis of cardiovascular oscillations: A new approach to the early prediction of pre-eclampsia*, Chaos. 17(1):015113, 2007.

[7] H. Stepan, A. Geipel, F. Schwarz, T. Krämer, N. Wessel, and R. Faber, *Circulatory soluble endoglin and its predictive value for preeclampsia in second-trimester pregnancies with abnormal uterine perfusion*, Am. J. Obstet. Gynecol. 198(2):175.e1–e6, 2008.

[8] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, *Complex network approach for recurrence analysis of time series*, Phys. Lett. A. 373(46):4246–54, 2009.

[9] R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths, *Recurrence networks – a novel paradigm for nonlinear time series analysis*, New J. Phys. 12(3):033025, 2010.

[10] V. Batagelj and M. Zaveršnik, *An O(m) algorithm for cores decomposition of networks*. University of Ljubljana, Institute of Mathematics, Physics and Mechanics, Dept. of Mathematics, 2002.

[11] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, *Complex networks: Structure and dynamics*, Phys. Rep. 424(4-5):175–308, 2006.

[12] N. Wessel, A. Voss, H. Malberg, C. Ziehmann, H. U. Voss, A. Schirdewan, U. Meyerfeldt, and J. Kurths, *Nonlinear analysis of complex phenomena in cardiological data*, Herzschr. Elektrophys. 11(3):159–73, 2000.

[13] M. Riedl, A. Suhrbier, H. Stepan, J. Kurths, and N. Wessel, *Short-term couplings of the cardiovascular system in pregnant women suffering from pre-eclampsia*, Philos. T. Roy. Soc. A. 368(1918):2237–50, 2010.

[14] http://pajek.imfm.si/doku.php?id=pajek.

[15] R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths, *Ambiguities in recurrence-based complex network representations of time series*, Phys. Rev. E. 81(1):015101, 2010.

[16] L. C. Kenny, D. I. Broadhurst, W. Dunn, M. Brown, R. A. North, L. McCowan, C. Roberts, G. J. S. Cooper, D. B. Kell, and P. N. Baker, *Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers*, Hypertension. 56(4):741–9, 2010.

[17] J. M. Roberts and D. W. Cooper, *Pathogenesis and genetics of pre-eclampsia*, Lancet. 357(9249):53–6, 2001.

[18] A.-K. Wikström, O. Stephansson, and S. Cnattingius, *Tobacco use during pregnancy and preeclampsia risk*, Hypertension. 55(5):1254–9, 2010.

[19] P. Magnus and A. Eskild, *Seasonal variation in the occurrence of pre-eclampsia*, BJOG: Int. J. Obstet. Gynaecol. 108(11):1116–9, 2001.