Granger Causality Analysis based on Neural Networks Architectures for bivariate cases

Alvaro D. Orjuela-Cañón
School of Medicine and Health Sciences
Universidad del Rosario
Bogota D.C., Colombia
alvaro.orjuela@urosario.edu.co

Jan A. Freund
ICBM & Research Center Neurosensory Science
Carl von Ossietzky Universität Oldenburg
Oldenburg, Germany
jan.freund@uni-oldenburg.de

Andres Jutinico
Mechanical, Electronics and Biomedical Faculty
Universidad Antonio Nariño
Bogota D.C., Colombia
ajutinico@uan.edu.co

Alexander Cerquera
Department of Neurology-College of Medicine
University of Florida
Gainesville, FL, United States
ealexander.cerqu@ufl.edu

***Abstract*—This work deals with an analysis of Granger Causality computation based on artificial neural networks, including a nonlinear relation between the involved variables. Information about the training parameters are exhibited in order to visualize how the conditions of the chosen model to obtain the connectivity information depend on the architecture of network. Three chaotic maps with a bivariate case built from two time series were employed to see the effect of training parameters of the models. Nonlinear autoregressive and nonlinear autoregressive with exogenous inputs were used to forecast the time series, and then, obtain the causality information based on differences of errors between both approximations. Results show that the causality computation is sensible to neural network parameters previously untreated in a detailed mode.**

Keywords—Granger causality, nonlinear models, transfer entropy, artificial neural networks, time series forecasting

I. INTRODUCTION

Causality is a proposed concept that aims to determine if some process or system contributes to other in some way. For this, a measure of this effect can be obtained through information from time series that represents any two or more systems involved [1]. In this way, when a first time series X_1 improves the prediction of a second time series X_2, based on knowledge of the first one, means that there is a causality relation from X_1 to X_2.

One way to capture the contribution or flow of information is carried out by Granger causality (GC), which quantifies the connectivity between two or more systems represented by the mentioned time series. Due to this, a classical computation of the GC measure is based on the use of multivariate autoregressive (MVAR) models [2], [3]. The utility of this tool is the extracted information related to the flow of the connections of the systems implicated.

In spite of the wide use of GC in different applications, mainly in electroencephalographic signals [4], [5], some limitations have been reported, due to the model assumes linearity inside data and taking as basis parametric tests. Moreover, approximations that follow the same objective of the connectivity quantification have been emerged. An alternative based on information theory is known as transfer entropy (TE) [6]. In this case, TE does not employ a parametric basis as GC does and quantifies the flow of information through the computation of the communication of the processes under analysis. In addition, TE is a nonlinear and nonparametric mathematical tool that is employed in different applications, mainly in neurosciences [7]–[9]. Nevertheless, in some cases the GC and TE metrics can provide the same information about the systems or subsystems implicated. This is how some authors evinced that when the time series obey to a Gaussian distribution, the two measures are equivalent [10].

Some years ago, the artificial neural networks (NN) have been proposed to modify the traditional computation of GC, allowing to formulate a Granger causality in a neural network sense (NNGC) [11], [12]. In this way, this updated approximation provides a nonlinearity that the basic model does not hold. For this, different architectures of NN have been utilized, such as radial basis function (RBF) networks in [13], [14], where the nonlinearity is a task of a hidden layer composed by Gaussian functions. For example, in the chapter five from the book [15], authors compared the RBF architecture and other models to analyze the Mackey-Glass system and to obtain a causal machine. The results showed that RBF and multilayer perceptron (MLP) models had comparable results. However, in most of the reports that employ MLP networks, the authors constrain the parameters to take into account of this type of implementations. One important parameter is the number of neurons in the hidden layer, which is not analyzed as it must.

Other structures like recurrent neural networks, as echo state networks (ESN) or long short term memory (LSTM), have been employed due to its advantages in time series forecasting in specific contexts [16], [17]. Also, additional works related with the use of NN for GC computation, show the relevance of the computation through a interpretable way. This can be seen in [18], where a MLP and a LSTM networks were employed to understand the data in terms of GC computation in a nonlinear mode. For this, the connections inside the network are proposed in a sparse method, where the architecture named the *componentwise* allow to the MLP training the inclusion of penalties in the error minimization process. However, details with relevance in the NN architectures and its training have not been reported. For example, mostly of the employed models are adjusted to a fixed number of neurons in the hidden layer, modifying the number of hidden layers of the architecture. This

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

is opposite to the universal approximation theorem, which defends the use of just one hidden layer for MLP proposals [19]. This is a supported by authors that see these computational models as architectures based on kernels, where the hidden layer develops this function, making a nonlinear mapping of the data input. Examples of nonlinear GC at employing kernel have been reported in [13], [20].

The aim of the present work is to determine what training aspects of neural networks can affect the computation of GC connectivity. For this, the employment of nonlinear autoregressive (NAR) models is proposed joint to the same representation, but including an additional exogenous input approximation. Experiments were developed considering just a bivariate case, showing some aspects that have not been detailed for some authors in previous works. For example, the GC computation in relation to the number of parameters implicated in the NN architectures and its capacity of learning from data. The parameters considered here are: i) number of inputs used to compute the NNGC, ii) number of neurons in hidden layer of the models, and, iii) the quantity of data employed to train the network. The paper is divided into a section II that explains the time series, models and tools used to develop the comparison. Section III shows the results in terms of computed causalities emphasizing in the NN models with a discussion in the section IV. Finally, section V describes some conclusions obtained after of realization of this work.

II. Materials and Methods

A. Time series

For developing the experiments and to compare the proposed models for connectivity identification, three scenarios with different generation of time series were simulated: i) A multichaotic map employed in [21], ii) Hénon chaotic map, and iii) Ikeda chaotic map. First one, an approximation consist of two couple chaotic maps employed by Montalto in [21]. There, the series were generated by:

$$\begin{aligned} x_n &= 1 - \beta b_1^2 + d\varepsilon_n \\ y_n &= (1 - C_1)(1 - \beta b_1^2) + C_1(1 - \beta b_1^2) + d\varepsilon_n \end{aligned} \quad (1)$$

where the influence of two time series with n samples and determined as from x to y, is given by a coupling coefficient $C_1 = 0.2$, $b_1=|means(x_{,n-1})|$, $b_2=|means(y_{,n-1})|$, $\beta = 1.8$, $d = 0.03$ is the value that control the noise and ε is the Gaussian noise. The data were obtained by a function given by the same toolbox described in [21]. For the present paper, this map is named *Montalto map* for comparison of the results.

Second map was built through the simulation from Hénon map, according to the expression:

$$\begin{aligned} x_{n+1} &= 1 - ax_n^2 + y_n \\ y_{n+1} &= bx_n \end{aligned} \quad (2)$$

considering a classical map where $a=1.4$ and $b=0.3$, which produces a chaotic behavior.

Finally, the Ikeda map was employed for simulating a coupled time series in the way:

$$x_{n+1} = 1 + u(x_n sint_n - y_n cost_n) \quad (3)$$

$$y_{n+1} = u(x_n cost_n - y_n sint_n)$$

where $u=0.9$ and $t_n = 0.4 - \frac{6}{1+x_n^2+y_n^2}$.

For all scenarios, the time series were composed by 500 data points, allowing to develop the computational experimentation with different algorithms without to take into account the hardware for the processing. In addition, time series were normalized to interval [-1, 1] before applying the used methods.

B. Transfer entropy approximation

The process to compute the TE connectivity measure between two systems x and y is based on the quantification of the mount of uncertainty. This method employs the quantity given by Shannon's entropy. In our case, the MuTE toolbox computed an embedding stage, applying parameters as dimension and delay for reconstruction in a phase space. Then, the optimization the paramenters selection was obtained for each time series. This allows to reduce the redundant information, compared to when the embedding is complete fixed parameters.

For x and y, which represent two subsystems (source and target respectively), are sampled at a present time n. Then, the TE from x to y is given by:

$$\begin{aligned} TE_{x\to y} = H(y_n \backslash y_{n-1:n-L}) \\ - H(y_n \backslash y_{n-1:n-L}, x_{n-1:n-L}) \end{aligned} \quad (4)$$

where $H(a)$ is the Shannon entropy of a and the '\' operator means dependence from other samples or time series. Surrogate data analysis was applied to evaluate the significance of the results. At last, the quantification of connectivity is obtained, showing the relation between the subsystems and its direction. The employed toolbox computes the TE connectivity through different techniques, proposing embedding before the obtain the metric [21]. The employed techniques for the embedding were: *i) linear estimation*, where an autoregressive (AR) model is used to compute the relation between the series, *ii)* binning *estimation*, based on fixed state space partitioning, and *iii) neural network estimation,* where models based on NN with limitations about the manipulation of the parameters is used. The first two estimators were applied with a nonuniform embedding, according to [21]. The third method was used as it was explained in [11], [22], and in this way, compare to the detailed use of NN in the present work.

C. Granger causality computation

For the GC computation, the multivariate GC (MVGC) toolbox was used to obtain the connectivity information [3]. The time series were passed by a process of order computation, stability and finally, a test that estimated the significance of differences between the time series forecasting [3]. Models were obtained through the computation of a bivariate AR system with order ρ. This system can be represented by:

$$\begin{aligned} x_n &= \sum_{j=1}^{\rho} A_{11,j} x_{n-j} + \sum_{j=1}^{\rho} A_{12,j} y_{n-j} + \varepsilon_1(t) \\ y_n &= \sum_{j=1}^{\rho} A_{21,j} x_{n-j} + \sum_{j=1}^{\rho} A_{22,j} y_{n-j} + \varepsilon_2(t) \end{aligned} \quad (5)$$

where x_n and y_n are two time series that represent two process components and ρ the order of the model. A_j is a matrix containing the regression coefficients for each j=1... ρ, and ε_1 and ε_2 are the residual (or prediction error) for each variable, respectively.

Two models were considered: *i) full model*, this includes information from two time series, and *ii) reduced model,* which holds the information from just one series. The computation of GC from *y* to *x* is obtained through:

$$GC_{y\to x} = ln\frac{var(\varepsilon_R)}{var(\varepsilon_F)} \tag{6}$$

where ε_R is the error of the restricted model, and ε_F indicates the error of the full model.

D. Granger causality in the neural network sense.

For determining the Granger causality in a NN sense (NNGC), NAR and NARX models were employed. First one uses just one time series to obtain the forecasting of itself such us the linear AR model. Second model includes an exogenous input given by the other time series, and in this way, develop the forecasting. Then, the error of both proposals is compared, finding what forecasting was better and determining the causality. Capacity of NN to learn patterns and achieve approximations of nonlinear functions was taken in advantage [23]. This compared to the linear approximation that GC does when develop the model from data.

In a similar mode that the reduced and full models from the GC approximation, it is possible to have a nonlinear representation of x_i, adjusting the a_i coefficients such as the AR model was obtained. For this, the computation is given by the way:

$$x_i = tanh\left(\sum_{k=1}^{p} a_i x_{i-L} + b\right) \tag{7}$$

where x_i is the time series to be modeled, a_i are the coefficients of the AR model and known as synaptic weights (w_{ij}). The nonlinearity that changes the AR model to a NAR model is hold in the hyperbolic tangent (tanh) in (7). When information from other time series is included, the NAR model is modified to an exogenous with this data in the way:

$$x_i = tanh\left(\sum_{k=1}^{p} a_i x_{i-L} + b_i y_{i-L}\right) \tag{8}$$

Here, it is included the data from the *y* time series. In this case, the *b* parameter is related to the synaptic weights that will be computed to do the forecasting in the *x* series. Figure 1 shows the architectures of NAR and NARX models.

NN models have the particularity that its dependence of the synaptic weights. In this way, the number of inputs were adjusted in an experimental mode. For this, lags of the input was modified from one to five, searching a model with low error and low complexity. In addition, the number of units in the hidden layer was changed from one to five. For training, a subset with 50, 60, 70, 80 and 90% from the time series was employed to train the model. The resilient backpropagation algorithm was used due to speed performance compared with other training algorithms [24], [25]. Best model was considered based on computation of errors using the developing subset.

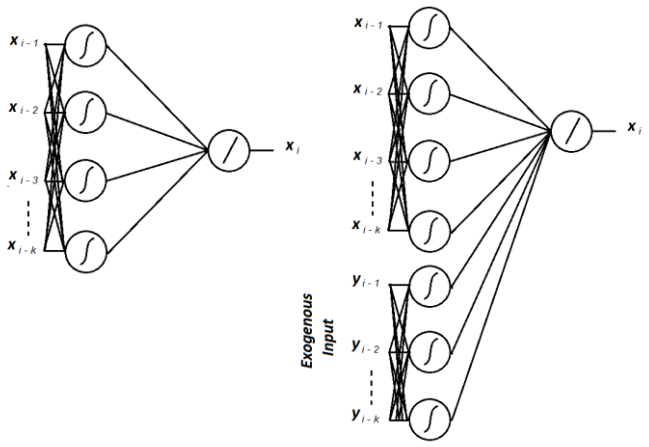

Fig. 1. NAR and NARX employed architectures.

To deal with the initialization problem from NN models, in each architecture 100 initializations were tested. Also, the portion of time series to train and validate was modified from 50% to 90% in the training set. This allowed to observe the overfitting of the models, when the network learnt the time series in a specialized way.

The computation of GC in terms of NN models was obtained in a similar way as in [11], given by:

$$NNGC_{y\to x} = \varepsilon_R - \varepsilon_F \tag{9}$$

where ε_R is forecasting error obtained by a reduced model that just uses the NAR for modeling the time series *x* and ε_F is the NARX full model that includes the information of time series *y*.

Finally, the Mann-Whitley Wilcoxon test was employed to determine if the differences from NAR and NARX models were statistically significant [26]. This allowed to detect the connectivity between the subsystems given by the *x* and *y* time series.

III. Results

Results for the connectivity according to GC, TE and NNGC are shown in a separate mode. The emphasis was put on the approximation based on NAR and NARX architectures, where the aspects of the error based on quantity of data for training the models, lags, neurons in the hidden layer are indicated.

A. Transfer Entropy

Results obtained through the use of information of entropy were obtained employing the toolbox shown in [21]. Table I visualize the entropy values for each couple of time series in each of the three scenarios studied.

It is possible to see how the binning and linear methods can find the connectivity from *x* to *y* series, but the flow of information in the opposite way was not found for the Montalto map, where TE value is zero. This is supported by same authors in the work where the toolbox is described [21]. In spite of this, the both methods based on neural networks reported connectivity from *y* to *x* (see Table I), which show that the use

of this kind of models can be sensitive if its parameters did not studied.

TABLE I. RESULTS FOR THE CONNECTIVITY BETWEEN TIME SERIES *X* AND *Y* USING TRANSFER ENTROPY

Method	Map	Connectivity	*TE value*
Binning Nonuniform Embedding	Montalto	$x \rightarrow y$	0.2941
		$x \leftarrow y$	0
	Hénon	$x \rightarrow y$	1.7248
		$x \leftarrow y$	0
	Ikeda	$x \rightarrow y$	0.6752
		$x \leftarrow y$	1.0120
Linear Nonuniform Embedding	Montalto	$x \rightarrow y$	0.0453
		$x \leftarrow y$	0
	Hénon	$x \rightarrow y$	13.9278
		$x \leftarrow y$	0.0503
	Ikeda	$x \rightarrow y$	0.9704
		$x \leftarrow y$	0.1074
Neural Networks Uniform Embedding	Montalto	$x \rightarrow y$	0.5110
		$x \leftarrow y$	0.1035
	Hénon	$x \rightarrow y$	0.0845
		$x \leftarrow y$	0.5902
	Ikeda	$x \rightarrow y$	3.2392
		$x \leftarrow y$	0.2386
Neural Networks Nonuniform Embedding	Montalto	$x \rightarrow y$	0.6016
		$x \leftarrow y$	0.0192
	Hénon	$x \rightarrow y$	1.1059
		$x \leftarrow y$	0.0579
	Ikeda	$x \rightarrow y$	2.8542
		$x \leftarrow y$	1.1986

The Hénon map evinced same effect, showing flow of information from *y* to *x*, direction that was not reported in a previous work that dealt with five systems [11]. In this case, the toolbox show disagreement in this point with same methods mentioned for the Montalto case.

B. Granger Causality

Results obtained through the use of MVGC models, employing the toolbox shown in [3] can be seen in Table II. There, the ones indicate that exists a flow of information from *y* to *x* for all the cases. This computation was given by a statistical test that obtains as a result the connectivity or not between subsystems, according to expression (6). For the Montalto and Hénon scenarios, there were not established causalities in a wrong way as a previous method. However, the results depict that this linear method is not enough to find the nonlinear relations between data for Hénon map. Order for the AR model employed were three for two first maps and five for the last. Here, it is possible to see that this method achieved to find this representation of data.

C. Granger Causality with Neural Networks models

Employing only NAR models, the forecasting for *x* time series for the Montalto map had a minimum error when the training was developed employing 60% of the series. Therefore, the error value was 0.0381 ± 0.0505 for the *x* time series and 0.0286 ± 0.0079 (using 70% for training) for the *y* time series (see figures 2 and 3) with specific values for lags and neurons in the hidden layer as shown in Table III. In a similar way, for the Hénon chaotic map the minimum mean error value was 0.0408 ± 0.0465 for the *x* time series and 0.0382 ± 0.0489 for the *y* time series. Finally, for Ikeda map this value was 0.1476 ± 0.0343 and 0.0920 ± 0.0423 for *x* and *y* time series respectively.

TABLE II. RESULTS FOR THE CONNECTIVITY BETWEEN TIME SERIES *X* AND *Y* USING GRANGER CAUSALITY

Map	*Connectivity*	*GC*
Montalto	$x \rightarrow y$	0
	$x \leftarrow y$	1
Hénon	$x \rightarrow y$	0
	$x \leftarrow y$	1
Ikeda	$x \rightarrow y$	1
	$x \leftarrow y$	1

Figures 2 and 3 visualizes how each time series has a different condition for its training. For the *x* and *y* forecasting applying NAR models the map that presents the highest level of error was the obtained by Ikeda. At the same time, the figures show that the Hénon map has time series that need more quantity in the data to do the best forecasting. In this case, it was necessary the 90% of the time series to do the forecasting.

According to the information from Table III, NARX models with the same parameters were used to developed the forecasting. Here, the information from the couple of time series was attached in the input of the model, observing the differences between the errors from both approaches. Figures 4 to 6 exhibit the histograms for the error from 100 training results given by architectures mentioned. Table IV resumes the connectivity devised with NNGC computation. There, it is seen that the Montalto map has only the flow $x \leftarrow y$ as reported in the literature. For the Hénon and Ikeda scenarios, NNGC obtained the connectivity for both directions in the bivariate cases.

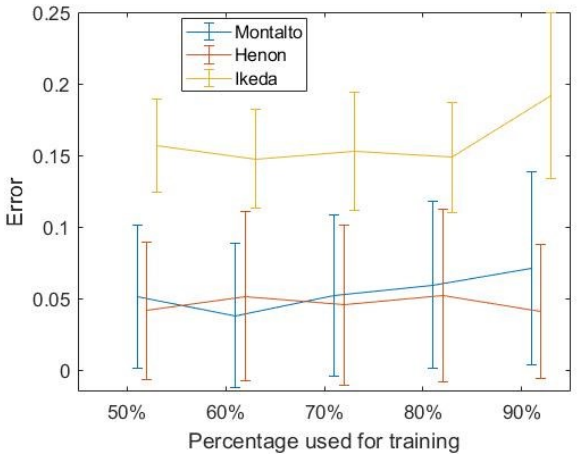

Fig. 2. Results of NAR models for the *x* time series forecasting for the three studies scenarios.

IV. DISCUSSION

The Ikeda map time series presented the highest error value, showing that this series is more difficult to forecast (see Fig. 2). In spite of that, models trained with 60% of the series was enough to find the best result. At the same time, it is possible to see that the Montalto and Ikeda maps exhibit opposite behavior, while one of these presents minimum trendy, the other show higher values for the error. This evidence allows to determine that depending on the time series type, the segments used to train and validate the NN play an important role for the forecasting, making that the interpretation of the NNGC changes for each context.

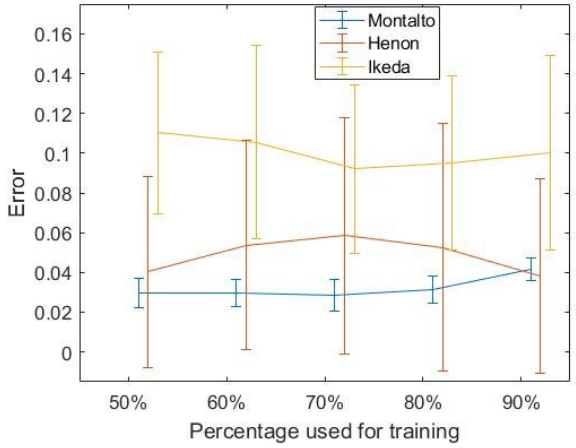

Fig. 3. Results of NAR models for the *y* time series forecasting for the three studies scenarios.

TABLE III. RESUME FOR THE RESULTS FOR INDIVIDUAL TIME SERIES FORECASTING

Chaotic Map	Time Series	Training parameters			
		Percentage	*Lags*	*Neurons*	*Error*
Montalto	*X*	60	4	5	0.0009±0.0017
	Y	70	1	4	0.0009±0.0023
Hénon	*X*	90	2	5	0.0006±0.0005
	Y	90	3	3	0.0009±0.0021
Ikeda	*X*	60	4	5	0.0046±0.0029
	Y	70	2	4	0.0024±0.0294

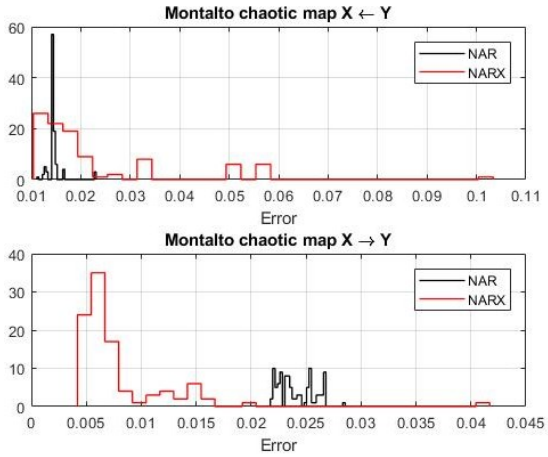

Fig. 4. Comparison of the errors when NAR and NARX models were applied to the Montalto time series map. Histograms corresponds to the errors for trainings of the model with characteristics in Table III.

An interesting aspect established was the effect of the MuTE toolbox. There the binning and linear methods determined the connectivity $x \rightarrow y$, which was reported by the same authors, but the approach given by NN models could not obtain this behavior, indicating causality for both directions. At the same time, more details about this outcome could be not studied. The election of the parameters for NN are restrictive and an exploration as made in this work is not easy to implement. In addition, for the present case, the use of visual information related to the histograms obtained allow to understand better the performance of models, and to permit to assess the error in a complementary way. This can be seen in Figure 5, where in spite of visual differences observed in the histogram, the statistical test confirmed that the error for both models was different.

In relation to the GC computation base on linear models, it was seen that for the Hénon map the connectivity $x \leftarrow y$ was not obtained. This corresponds to different studies about the relevance of nonlinear GC computation, something that was dealt in this work. As a complement of this study, and making use of the presented information with a deeper analyses, it could be helpful to understand how the nonlinearity works in this level.

Further analysis with other NN models must be explored. Examples of this can be seen in [17], where the use of LSTM networks was compared with other NN models. However, aspects related to training parameters and architecture of the models were missing. Also, other recurrent neural networks can be included in order to compare the advantages of the LSTM models in this specific application.

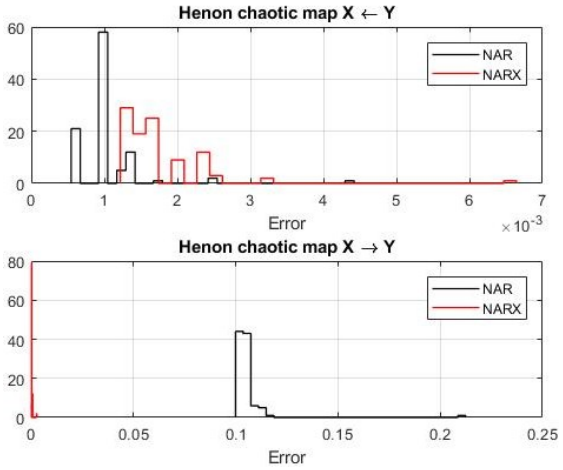

Fig. 5. Comparison of the errors when NAR and NARX models were applied to the Hénon time series map. Histograms corresponds to the errors for trainings of the model with characteristics in Table III.

V. CONCLUSIONS

Neural networks are sensitive to different parameters that modify the capacity of learning. In this work, these aspects were analyzed in order to compute a nonlinear Granger causality (NNGC). Results show how each case of NNGC computation needs special attention of NN parameters, due to the forecasting, which is an important step of the NNGC measure, is determined for the specific characteristics of the time series used. As future work, the exploration of other architectures, as for example recurrent neural networks, must be analyzed, due to its advantages for time series forecasting.

ACKNOWLEDGMENT

Authors acknowledge the participation of Universidad del Rosario, Universidad Antonio Nariño (project 2018213), Carl von Ossietzky Universität Oldenburg and University of Florida.

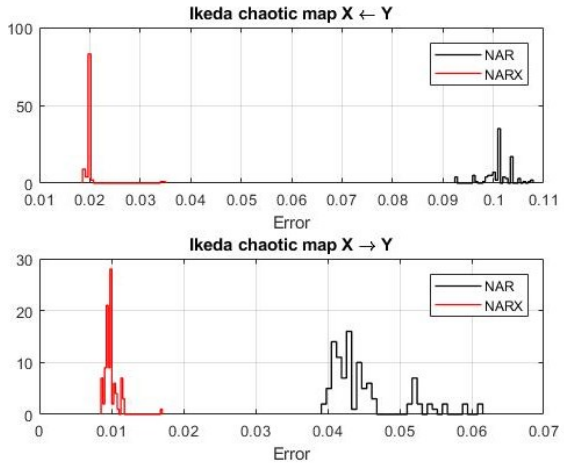

Fig. 6. Comparison of the errors when NAR and NARX models were applied for the Ikeda time series map. Histograms corresponds to the errors for trainings of the model with characteristics in Table III.

TABLE IV. RESULTS FOR THE CONNECTIVITY BETWEEN TIME SERIES *X* AND *Y* USING GRANGER CAUSALITY

Map	Connectivity	NNGC	*p-value*
Montalto	$x \rightarrow y$	0	0.0849
	$x \leftarrow y$	1	0
Hénon	$x \rightarrow y$	1	0
	$x \leftarrow y$	1	0
Ikeda	$x \rightarrow y$	1	0
	$x \leftarrow y$	1	0

REFERENCES

[1] S. L. Bressler and A. K. Seth, "Wiener--Granger causality: a well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[2] S. Guo, C. Ladroue, and J. Feng, "Granger causality: theory and applications," in *Frontiers in Computational and Systems Biology*, Springer, 2010, pp. 83–111.

[3] L. Barnett and A. K. Seth, "The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference," *J. Neurosci. Methods*, vol. 223, pp. 50–68, 2014.

[4] L. Faes, D. Marinazzo, F. Jurysta, and G. Nollo, "Granger causality analysis of sleep brain-heart interactions," in *Cardiovascular Oscillations (ESGCO), 2014 8th Conference of the European Study Group on*, 2014, pp. 5–6.

[5] L. Faes, D. Marinazzo, S. Stramaglia, F. Jurysta, A. Porta, and N. Giandomenico, "Predictability decomposition detects the impairment of brain--heart dynamical networks during sleep disorders and their recovery with treatment," *Phil. Trans. R. Soc. A*, vol. 374, no. 2067, p. 20150177, 2016.

[6] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, p. 461, 2000.

[7] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy: a model-free measure of effective connectivity for the neurosciences," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 45–67, 2011.

[8] A. Cerquera, A. Orjuela-Cañón, J. Roa-Huertas, J. A. Freund, G. Julia-Serdá, and A. Ravelo-García, "Transfer entropy to characterize brain-heart topology in sleep apnea patients treated with continuous positive airway pressure," in *12th International Symposium on Medical Information Processing and Analysis (SIPAIM 2016)*, 2016.

[9] L. Faes, G. Nollo, F. Jurysta, and D. Marinazzo, "Information dynamics of brain–heart physiological networks during sleep," *New J. Phys.*, vol. 16, no. 10, p. 105005, Oct. 2014.

[10] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.*, vol. 103, no. 23, p. 238701, 2009.

[11] A. Montalto, S. Stramaglia, L. Faes, G. Tessitore, R. Prevete, and D. Marinazzo, "Neural networks with non-uniform embedding and explicit validation phase to assess Granger causality," *Neural Networks*, vol. 71, pp. 159–171, 2015.

[12] A. Attanasio and U. Triacca, "Detecting human influence on climate using neural networks based Granger causality," *Theor. Appl. Climatol.*, vol. 103, no. 1–2, pp. 103–107, 2011.

[13] N. Ancona, D. Marinazzo, and S. Stramaglia, "Radial basis function approach to nonlinear Granger causality of time series," *Phys. Rev. E*, vol. 70, no. 5, p. 56221, 2004.

[14] H. Ma, K. Aihara, and L. Chen, "Detecting causality from nonlinear dynamics with short-term time series," *Sci. Rep.*, vol. 4, p. 7464, 2014.

[15] M. Tshilidzi, *Causality, correlation and artificial intelligence for rational decision making*. World Scientific, 2015.

[16] A. Duggento, M. Guerrisi, and N. Toschi, "Echo State Network models for nonlinear Granger causality," *bioRxiv*, p. 651679, 2019.

[17] Y. Wang, K. Lin, Y. Qi, Q. Lian, S. Feng, Z. Wu, and G. Pan, "Estimating brain connectivity with varying-length time lags using a recurrent neural network," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1953–1963, 2018.

[18] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox, "Neural granger causality for nonlinear time series," *arXiv Prepr. arXiv1802.05842*, 2018.

[19] V. Kurková, "Kolmogorov's theorem and multilayer neural networks," *Neural networks*, vol. 5, no. 3, pp. 501–506, 1992.

[20] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Phys. Rev. Lett.*, vol. 100, no. 14, p. 144103, 2008.

[21] A. Montalto, L. Faes, and D. Marinazzo, "MuTE: a MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy," *PLoS One*, vol. 9, no. 10, p. e109462, 2014.

[22] A. Montalto, L. Faes, and D. Marinazzo, "MuTE: a freeware, modular toolbox to evaluate Multivariate Transfer Entropy and Artificial Neural Networks Granger causality.," *Front. Neuroinform.*, no. 35.

[23] S. Haykin, *Neural Networks and Learning Machines*, 3ra ed. Pearson, 2009.

[24] R. S. Naoum, N. A. Abid, and Z. N. Al-Sultani, "An enhanced resilient backpropagation artificial neural network for intrusion detection system," *Int. J. Comput. Sci. Netw. Secur.*, vol. 12, no. 3, p. 11, 2012.

[25] F. Günther and S. Fritsch, "Neuralnet: Training of neural networks," *R J.*, vol. 2, no. 1, pp. 30–38, 2010.

[26] G. Cardillo, "MWWTEST: Mann-Whitney-Wilcoxon non parametric test for two unpaired samples." 2009.