Evolutionary computation to explain deep learning models for time series forecasting

A. R. Troncoso-García
Data Science and Big Data Lab
Pablo de Olavide University
ES-41013, Seville, Spain
artrogar@upo.es

M. Martínez-Ballesteros
Department of Computer Science
University of Seville
ES-41012, Seville, Spain
mariamartinez@us.es

F. Martínez-Álvarez
Data Science and Big Data Lab
Pablo de Olavide University
ES-41013, Seville, Spain
fmaralv@upo.es

A. Troncoso
Data Science and Big Data Lab
Pablo de Olavide University
ES-41013, Seville, Spain
atrolor@upo.es

ABSTRACT

Deep learning has become one of the most useful tools in the last years to mine information from large datasets. Despite the successful application to many research fields, deep learning is known as a black box approach and most experts experience difficulties to explain and interpret deep learning results. In this context, explainable artificial intelligence (XAI) is emerging with the aim of providing black box models with sufficient interpretability so that models can be easily understood by humans. The use of an evolutionary-based association rules extraction algorithm to explain deep learning models for multi-step time series forecasting is addressed in this work. This evolutionary application is proposed to be used with the predictions obtained by long-short term memory (LSTM) deep learning network. Data from Spanish electricity energy consumption has been used to assess the suitability of the proposal, showing that almost 98% of the model can be explained.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**;

ACM Reference Format:
A. R. Troncoso-García, M. Martínez-Ballesteros, F. Martínez-Álvarez, and A. Troncoso. 2023. Evolutionary computation to explain deep learning models for time series forecasting. *Proceedings of ACM SAC Conference (SAC'23).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3555776.3578994

1 INTRODUCTION

Nowadays, deep learning (DL) is an essential tool used to make predictions or classify large and heterogeneous data in different fields. DL is the technology behind artificial intelligence applications in a wide range of industries. One of its most serious disadvantages is that DL models are considered black-box, meaning that it is impossible to know how the model obtains the output by applying inner nonlinear operations to the input. Explainability could be defined as a relation between the input data and the prediction of a model, in such a way that it can be comprehended by humans [1]. This concept is crucial because models are used today to make high-stakes decisions in essential sectors, namely health, security, or economy [4].

In this paper, the focus is on model-agnostic explainability. Model-agnostic techniques are applied to the results or predictions that have been obtained after training the model and are independent of the model. A new approach to interpret deep learning models applied to time series forecasting is presented. Predictions are explained through an evolutionary-based quantitative association rules (QARs) algorithm, hereinafter referred to as MOQAR [9]. The method is tested by using a long-short-term memory (LSTM) network for time series forecasting as a case study. The idea is explaining how LSTM generates the predicted values, with reference to the input features (past samples from the historical data, in this context). Predictions are used as input for the evolutionary MOQAR model, and QARs are shown as a reliable way to understand the internal behavior of the model.

The remainder of the article is structured as follows. In Section 2, recent advances in DL explainability are reviewed. Section 3 illustrates evolutionary algorithm MOQAR. Section 4 describes the experiments carried out, and Section 5 presents the results. Section 6 concludes the paper.

2 RELATED WORK

Nowadays, obtaining interpretable DL models is a hot topic research field. There are several research projects in the literature exploring the use of association rules for machine learning and deep learning interpretability applied to essential areas, such as disease detection in health care [16] or electric vehicle load demand [5].

The application of association rules (ARs) to achieve the goal of interpretability is a major trend in the field of XAI [11]. First, there are some examples creating interpretable models by using ARs. In [6], a model based on ARs and Bayesian analysis is built

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC'23, March 27 –April 2, 2023, Tallinn, Estonia

https://doi.org/10.1145/3555776.3578994

and tested in personalized medicine and health. Furthermore, in [14] a multi-objective optimization for multiple ARs is developed for interpretable classification. Experiments showed that the model obtains better performance and faster execution time than other ARs mining models. Another example is found in *Takagi-Sugeno-Kang*. Interpretability is added to an existing model by generating ARs. Interpretable intervals are calculated by finding overlapping values in common between antecedent and consequent data in ARs.

Then, the point is on increasing the interpretability of the preexisting ones. In [3], ARs are extracted using a model based on the well-known *Apriori* algorithm for explainability in an omic-data neural network. The rules are evaluated regarding a set of quantitative quality measures such as *Confidence, Support, Lift* or *Conviction*. Then, they are validated by human experts. Another example is shown in [7], where ARs are extracted from a decision tree model with high values of accuracy. Lastly, ARs explaining the predictions produced for a tabular classification dataset are provided in [12]. A neighbourhood of similar instances is built, and predictions are made for those perturbated instances. After that, the *k*-optimal ARs are selected. The focus is on ARs that cover more instances rather than the highest predictive rules.

3 EVOLUTIONARY MODEL FOR ASSOCIATION RULES EXTRACTION

In the field of data mining, ARs are a popular and well-known method to discover interesting relations among variables in large databases [2]. ARs are known as one of the highest interpretable techniques providing high-accuracy results [10]. An AR is known as QAR when the domain is continuous. The multi-objective algorithm based on the evolutionary NSGA-II called MOQAR has been applied in this work due to the effectiveness presented in previous researches. MOQAR mines QARs in datasets with continuous attributes without discretizing the attributes of the dataset trying to find the best trade-off among all the measures optimized. A detailed description of MOQAR can be found in [9]. MOQAR is described in a general way in this Section.

Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of features or attributes, with values in $\mathbb{R}$. Let S and T be two disjoint subsets of A, that is, $S \subset A$, $T \subset A$ and $S \cap T = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in S belong to the antecedent X, and features in T belong to the consequent Y, such that X and Y are formed by a conjunction of multiple Boolean expressions of the form $a_i \in [l, u]$, (with $l, u \in \mathbb{R}$). Thus, in a QAR, the features or attributes of the antecedent are related with the features of the consequent, establishing an interval of membership values for each attribute involved in the rule [8].

Many measures could be found in the literature to assess the quality of ARs. Definition and mathematical equation of the main quality measures can be found in [9]. In particular, *support* (Equation 1), *confidence* (Equation 2), and *gain* (Equation 3) have been the objective to be optimized by the evolutionary association rule extraction model. The objective is assessing the generality, reliability and information gain of the rules, respectively.

$$Support(X \Longrightarrow Y) = \frac{n(X \cap Y)}{N} \tag{1}$$

$$Confidence(X \Longrightarrow Y) = \frac{support(X \Longrightarrow Y)}{support(X)} \tag{2}$$

$$Gain(X \Longrightarrow Y) = confidence(X \Longrightarrow Y) - support(Y) \tag{3}$$

4 METHODOLOGY

The main goal of this work is to explore the ability of QARs to interpret the predictions of a time series made by DL models. Given a time series with previous values up to time t, $[X_1, \ldots, X_t]$, the task is to predict the h next values of the time series, from a window of w past values. This multi-step forecasting problem can be formulated as below, where f is the model to be learnt by the deep learning model in the training phase:

$$[X_{t+1}, X_{t+2}, \ldots, X_{t+h}] = f\left(X_t, X_{t-1}, \ldots, X_{t-(w-1)}\right) \tag{4}$$

First, we use the training set to train deep learning models obtaining the prediction model $\hat{f}$. The aim is learning *how* and *why* the model $\hat{f}$ makes a prediction. Consequently, we use the model $\hat{f}$ to make predictions for the same data that have been used to train it. That is:

$$\left[\hat{X}_{t+1}, \ldots, \hat{X}_{t+h}\right] = \hat{f}\left(X_t, \ldots, X_{t-(w-1)}\right) \tag{5}$$

where $[\hat{X}_{t+1}, \ldots, \hat{X}_{t+h}]$ are the predicted values by the deep learning model $\hat{f}$.

Then, the input dataset D for the MOQAR rule extraction algorithm is constructed as follows:

$$D = \{(X^{(i)}, Y^{(i)}) : i = 1, 2, ..., N\} \tag{6}$$

where N is the number of instances, $X^{(i)}$ and $Y^{(i)}$ are the features belonging to the antecedent and the consequent of the rule, respectively. These features are defined as follows:

$$X^{(i)} = [X_{t-(w-1)}, ..., X_{t-1}, X_t] \tag{7}$$

$$Y^{(i)} = [\hat{X}_{t+1}, \hat{X}_{t+2}, ..., \hat{X}_{t+h}] \tag{8}$$

where $t = w + (i-1) * h$.

In order to ensure that rules with all prediction horizons in the consequent are obtained, the input dataset D is divided into subsets D_j with $j = 1, ..., h$.

$$D_j = \{(X^{(i)}, Y_j^{(i)}) : i = 1, 2, ..., N\} \tag{9}$$

where the attributes forming the consequent of the rule are made up of a single attribute:

$$Y_j^{(i)} = \hat{X}_{t+j} \tag{10}$$

Then, we apply the evolutionary model MOQAR to all subsets D_j separately for the extraction of QARs. The parameters are configured for only retrieving the rules with a certain values of confidence, accuracy and support, just to minimize the number of valid rules for each iteration. As a result, we will obtain a comprehensible explanation of how the model $\hat{f}$ makes predictions using the output, which means that we will be able to interpret the deep learning algorithm. In this work, the interpretability of a recurrent neural network

Table 1: A selection of QARs obtained by MOQAR and quality measures for each prediction horizon for LSTM predictions.

h	Rule	Support	Confidence	Gain
1	IF $X_{t-1} \in [25122, 32611]$ AND $X_t \in [25618, 33561] \Longrightarrow \hat{X}_{t+1} \in [25230.04, 33063.40]$	0.42	0.98	0.52
2	IF $X_{t-2} \in [29810, 39018]$ $\Longrightarrow \hat{X}_{t+2} \in [28149.37, 38952.71]$	0.45	0.98	0.44
3	IF $X_t \in [26833, 32630]$ $\Longrightarrow \hat{X}_{t+3} \in [25947.24, 33570.65]$	0.35	0.98	0.52
4	IF $X_{t-114} \in [23085, 39275]$ AND $X_t \in [26786, 36046] \Longrightarrow \hat{X}_{t+4} \in [26556.69, 36436.02]$	0.46	0.97	0.41
5	IF $X_t \in [29971, 38813]$ $\Longrightarrow \hat{X}_{t+5} \in [28089.77, 38531.58]$	0.43	0.97	0.43
6	IF $X_{t-2} \in [22237, 29724]$ AND $X_{t-1} \in [22812, 29042] \Longrightarrow \hat{X}_{t+6} \in [22059.42, 30741.35]$	0.32	0.93	0.46
7	IF $X_t \in [23435, 29565]$ $\Longrightarrow \hat{X}_{t+7} \in [23286.75, 31999.55]$	0.32	0.93	0.45
8	IF $X_{t-3} \in [19503, 29267]$ $\Longrightarrow \hat{X}_{t+8} \in [19866.40, 31687.74]$	0.44	0.94	0.34
9	IF $X_{t-4} \in [21948, 33583]$ $\Longrightarrow \hat{X}_{t+9} \in [21678.36, 33638.91]$	0.62	0.89	0.20
10	IF $X_{t-129} \in [29781, 40611]$ AND $X_{t-1} \in [28913, 42075] \Longrightarrow \hat{X}_{t+10} \in [29408.60, 40850.50]$	0.37	0.98	0.47
11	IF $X_{t-1} \in [17846, 26748]$ $\Longrightarrow \hat{X}_{t+11} \in [18447.88, 29838.03]$	0.31	0.97	0.46
12	IF $X_{t-1} \in [20486, 28110]$ $\Longrightarrow \hat{X}_{t+12} \in [20873.39, 31424.08]$	0.36	0.96	0.41
13	IF $X_{t-134} \in [29047, 35642]$ AND $X_{t-8} \in [19893, 35062] \Longrightarrow \hat{X}_{t+13} \in [25787.62, 36114.02]$	0.32	0.96	0.37
14	IF $X_{t-3} \in [19330, 29017]$ $\Longrightarrow \hat{X}_{t+14} \in [19459.90, 31157.35]$	0.38	0.84	0.32
15	IF $X_{t-130} \in [19709, 28064]$ $\Longrightarrow \hat{X}_{t+15} \in [19733.12, 32743.51]$	0.38	0.89	0.27
16	IF $X_{t-133} \in [19523, 28556]$ AND $X_{t-85} \in [25115, 41521] \Longrightarrow \hat{X}_{t+16} \in [19602.96, 30048.36]$	0.31	0.90	0.41
17	IF $X_{t-133} \in [29176, 39694]$ AND $X_{t-128} \in [28498, 37063] \Longrightarrow \hat{X}_{t+17} \in [26667.44, 38203.40]$	0.42	0.96	0.33
18	IF $X_{t-121} \in [26852, 36382]$ AND $X_{t-115} \in [24036, 36978] \Longrightarrow \hat{X}_{t+18} \in [26940.19, 37320.54]$	0.45	0.81	0.20
19	IF $X_t \in [31778, 41140]$ $\Longrightarrow \hat{X}_{t+19} \in [28402.35, 42863.99]$	0.32	0.91	0.34
20	IF $X_{t-125} \in [29314, 38134]$ AND $X_{t-121} \in [30329, 37335] \Longrightarrow \hat{X}_{t+20} \in [24542.16, 39102.50]$	0.36	0.98	0.20
21	IF $X_{t-124} \in [28495, 35152]$ $\Longrightarrow \hat{X}_{t+21} \in [26016.44, 35727.54]$	0.38	0.89	0.29
22	IF $X_{t-113} \in [18036, 26626]$ AND $X_{t-103} \in [18916, 30240] \Longrightarrow \hat{X}_{t+22} \in [18987.06, 31309.65]$	0.32	0.89	0.29
23	IF $X_{t-109} \in [18628, 27189]$ $\Longrightarrow \hat{X}_{t+23} \in [18326.44, 30991.96]$	0.35	0.87	0.28
24	IF $X_{t-122} \in [31268, 41236]$ $\Longrightarrow \hat{X}_{t+24} \in [29353.53, 42480.26]$	0.34	0.88	0.37

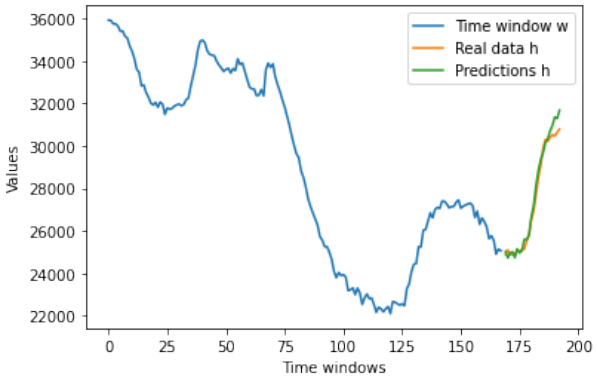


Figure 1: Example of a time series instance. Predictions made with LSTM model.

model, LSTM and its ability to predict the electricity demand in Spain is studied [15].

5 RESULTS

5.1 Input data

The input data used for this experiment is a time series of electrical energy consumption in Spain [13]. Data have been collected with 10-minute frequency during nine years and six months, specifically between January 1st 2007 and June 21st 2016. The time window used to predict has been set to 168 (1 day and 4 hours) whereas the value of the horizon of prediction h is 24, that is, 4 hours, as done and discussed in [15]. Figure 1 shows one instance of the dataset.

5.2 Quantitative association rules

The evolutionary MOQAR algorithm is used for obtaining a set of rules for each of the 24 samples forming the prediction horizon. For each one, between 12 and 20 rules are obtained, resulting in a total of 400 rules across all horizons.

The most representative QAR with regard to metrics such as confidence and support are presented in Table 1. In predictions of the first two hours of the horizon (X_{t+1} and X_{t+2}) the LSTM model gives a greater weight to the most recent values of the time series, approximately the two or three hours before, while for the third and fourth hour of the horizon (except for the horizons $t + 14$ and $t + 19$) it gives more weight to time series values farther away in time, coinciding with hours close to the previous day.

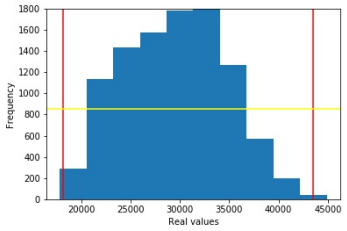
Table 1 presents both an example of QAR of each of the prediction horizons X_{t+h} and the quality measures of the QARs prevously mentioned. QARs show high confidence covering from 30% to 62% of the records in the dataset.

5.3 Explaining LSTM model

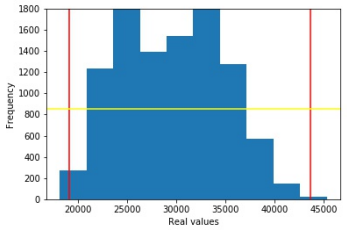
QARs are used as a comprehensible and understandable tool to explain *how* the predictions made by DL model LSTM. The amount

of real data explained by the set of rules of each prediction horizon X_{t+h} is calculated concerning the data range that QARs are covering.

Histograms for each prediction horizon are generated in order to analyze graphically the distribution of the actual values of the time series in different intervals. *Hist. range* is computed. *Hist. range* means the range of values with more than 50% of the samples, in other words, the range of values of time series X_t with more than 50% of the frequency in the histogram. Then, the percentage of range covered from histograms (*Hist. range covered (%)*) is about the relation of this interval and the *Range covered* by the set of rules. Overall, MOQAR explained more than 98% on average for all the samples in the prediction horizon, for the dataset used in this work.



(a) Best result $\hat{X}_{t+22}$, with 99.89% of values explained



(b) Worst result $\hat{X}_{t+17}$, with 96.03% of values explained.

Figure 2: Histograms of percentage of real data values covered of two different prediction horizons.

Examples of graphical representations are illustrated in Figure 2. Histograms represent the frequency of the real values of the time series. The red lines show the interval covered by the rules, while the yellow ones identify the interval with more than 50% of the frequency. In (a), it is shown the best result obtained for the prediction horizon $\hat{X}_{t+22}$ (99.89% covered), whereas in (b), the worst one, for $\hat{X}_{t+17}$ (96.03% covered).

6 CONCLUSIONS

In this work, QARs have been used to explain how a deep learning model, namely the well-known recurrent network model LSTM, makes predictions. For this purpose, the model obtained in the training of the LSTM network has been used first to obtain predictions. Then, QARs have been obtained using an evolutionary algorithm. QARs' antecedent (*IF* statement) is formed by the network input, that is, the past values of the time series. The consequent (*THEN* statement) is formed by the network output, i.e. the predictions obtained by the LSTM. Results have been reported using a real-world time series consisting of electricity consumption in Spain measured every 10 minutes. Overall, the results obtained show that the rules are a useful tool for explaining DL predictions. Using the QARs it is possible to get information about which time past values ($X_{t(w-1)}$) and in which range are important to calculate the prediction. Each prediction horizon X_{t+h} (from 1 to 24) has enough rules to cover the greater part of their data interval. Even the worst set of rules, concerning $\hat{X}_{t+17}$, is covering almost all the data in the real data range. Future work will be focused on creating graphical representations using information from QARs.

ACKNOWLEDGMENTS

The authors would like to thank the Spanish Ministry of Science and Innovation for the support under the project PID2020-117954RB-C21 and the European Regional Development Fund and Junta de Andalucía for projects PY20-00870 and UPO-138516.

REFERENCES

[1] A. Abanda. 2021. *Contributions to Time Series Classification: Meta-Learning and Explainability.* Ph.D. Dissertation. University of the Basque Country.
[2] R. Agrawal, T. Imielinski, and A. Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data.* 207–216.
[3] A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M. Aguilera, and J. Alcalá-Fdez. 2020. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Computational Biology* 16, 4 (2020), e1007792.
[4] A. Barredo-Arrieta, N. Díaz-Rodríguez, J. del Ser, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
[5] J. A. Gallardo-Gómez, F. Divina, A. Troncoso, and F. Martínez-Álvarez. 2022. Explainable Artificial Intelligence for the Electric Vehicle Load Demand Forecasting Problem. In *Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications.* 413–422.
[6] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.
[7] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano. 2021. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity* 2021 (2021), 6634811.
[8] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme. 2011. An evolutionary algorithm to discover quantitative association rules in multidimensional time series. *Soft Computing* 15, 10 (2011), 2065–2084.
[9] M. Martínez Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme. 2016. Improving a multi-objective evolutionary algorithm to discover quantitative association rules. *Knowledge and Information Systems* 49 (2016), 481–509.
[10] G. Pandey, S. Chawla, S. Poon, B. Arunasalam, and J. G. Davis. 2009. Association rules network: Definition and applications. *Statistical Analysis and Data Mining* 1, 4 (2009), 260–279.
[11] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He. 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association* 27, 7 (2020), 1173–1185.
[12] D. Rajapaksha, C. Bergmeir, and W. Buntine. 2020. LoRMIkA: Local rule-based model interpretability with k-optimal associations. *Information Sciences* 540 (2020), 221–241.
[13] R. Talavera, R. Pérez-Chacón, M. Martínez-Ballesteros, A. Troncoso, and F. Martínez-Álvarez. 2016. A nearest neighbours-based algorithm for big time series data forecasting. *Lecture Notes in Computer Science* 5391 (2016), 674–679.
[14] D. Thi, P. Meysman, and K. Laukens. 2022. MoMAC: Multi-objective optimization to combine multiple association rules into an interpretable classification. *Applied Intelligence* 52 (2022), 3090–3102.
[15] J. F. Torres, M. J. Jiménez-Navarro, F. Martínez-Álvarez, and A. Troncoso. 2021. Electricity Consumption Time Series Forecasting Using Temporal Convolutional Networks. In *Proceedings of the Conference of the Spanish Association for Artificial Intelligence.* 216–225.
[16] A. R. Troncoso-García, M. Martinez-Ballesteros, F Martínez-Álvarez, and A. Troncoso. 2022. Explainable machine learning for sleep apnea prediction. *Procedia Computer Science* 207 (2022), 2930–2939.