Data Preprocessing for ANN-based Industrial Time-Series Forecasting with Imbalanced Data

Ivan Pisa[1,2], Ignacio Santín[2], Jose Lopez Vicario[1], Antoni Morell[1], Ramon Vilanova[2]
[1]*Wireless Information Networking (WIN) Group*
[2]*Advanced Systems for Automation and Control (ASAC) Group*
Universitat Autònoma de Barcelona
08193 Bellaterra, Spain
{Ivan.Pisa, Ignacio.Santin, Jose.Vicario, Antoni.Morell, Ramon.Vilanova}@uab.cat

***Abstract*—The evolution of Industry towards the 4.0 paradigm has motivated the adoption of Artificial Neural Networks (ANNs) to deal with applications where predictive and maintenance tasks are performed. These tasks become difficult to carry out when rare events are present due to the imbalance of data. This is because training of ANN can be biased. Conventional techniques addressing this problem are mainly based on resampling-based approaches. However, these are not always feasible when dealing with time-series forecasting tasks in industrial scenarios. For that reason, this work proposes the application of data preprocessing techniques especially designed to face this scenario, a problem which has not been covered enough in the state-of-the-art. Considered techniques are applied over time-series data coming from Wastewater Treatment Plants (WWTPs). Our proposal significantly outperforms current strategies showing a 68% of improvement in terms of RMSE when rare events are addressed.**

***Index Terms*—Data preprocessing, Imbalanced data, Rare events, Artificial Neural Network**

I. Introduction

In the last years, the Industry 4.0 paradigm has arisen as one of the main motivators of industrial development in Europe. Its main objective is to interconnect industrial systems in order to cooperate between them and take decentralized decisions in the industrial plants management processes [1]. This paradigm also motivates the adoption of Artificial Neural Networks (ANNs) which are considered as an option to support some of the most popular tasks concerning industrial applications such as predictive maintenance, environmental violation prediction or stock maintenance, among others [2].

Input values considered by ANNs are usually obtained from sensors measurements which form continuous data flows. They take the form of time-series signals showing a high correlation in time. Besides, differently to typical Information Technology (IT)-based scenarios, manipulation of industrial-based signals should be addressed taking into account that all the information is important. For instance, detecting rare events (events with a very low probability to occur) becomes a task which gives highly-valuable information. In [3], the automatic detection of outliers, an example of rare events, in petrochemical industries is proposed to determine instrument faults, drifts or even process disturbances.

Time-series also face additional problems such as imbalanced data behavior. This is observed when certain range of values are overrepresented. When dealing with rare events forecasting, the imbalanced problem is even more severe: the great amount of data (usual values) will be overrepresented whilst the rare events hardly occur [4]. This phenomenon could be critical as ANNs could not be trained properly yielding to badly rare event predictions. It is worth noting that data preprocessing is a key factor in most ANN-based applications. So, in this case, the adoption of adequate data preprocessing techniques is even more crucial.

The literature presents several approaches to circumvent imbalanced data when dealing with classification problems. These are mainly based on the application of resampling techniques: subsampling and oversampling. Subsampling consists in decreasing the amount of overrepresented data achieving an equilibrium between overrepresented and underrepresented classes [5]. Opposite to subsampling, oversampling is based on the generation of new synthetic data according to the characteristics of the underrepresented class [6]. The effects of applying oversampling and subsampling techniques are widely analyzed in [7]. Synthetic Minority Over-sampling Technique (SMOTE), a combination of oversampling and subsampling, is proposed in [8]. It generates new synthetic data for the lowest represented class adopting k-nearest neighbors (K-NN) [8].

Nonetheless, a few works address the imbalanced data issue in the case of time-series forecasting, i.e., regression problems. Some adopt SMOTE extension for regression purposes [9] while other approaches face the imbalanced problem resorting to modified resampling strategies [10], [11]. For instance, in [10] a first classification into high or low presence values is performed. Later, the more represented class is subsampled in order to obtain a balanced dataset. However, resampling techniques (and their derivations) are not always adequate for industrial applications dealing with time-series. This is because these strategies entail the loss of time correlation and dependence between measurements. In [4], authors try to cope with the preservation of time correlation. Again, this

This research is supported by the Catalan Government under Projects 2017 SGR 1202 and 2017 SGR 1670, by La Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya i del Fons Social Europeu under FI grant and also by the Spanish Government under Projects TEC2017-84321-C4-4-R and DPI2016-77271-R co-funded with European Union ERDF funds.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

is performed by means of a classification of data into two types of bins (relevant or not relevant bins), but the bins now consist in successive observations of the time-series. Then, subsampling is performed only on not relevant bins, those which are overrepresented. By doing so, however, information is lost which is not adequate for industrial applications as previously commented.

For that reason, the aim of this work is to shed some light on data preprocessing for regression problems with imbalanced time-series data, especially in the case of rare events. In this case, time-correlation will be preserved at the same time imbalanced data problem is addressed. Data preprocessing will be applied over data coming from Wastewater Treatment Plants (WWTPs), where an ANN-based Soft Sensor is proposed to predict WWTPs effluent concentrations. These techniques consist in the application of a sliding window protocol, normalization of data and K-Fold based training scheme. The application is specific, but results and conclusions can be extrapolated to other industrial scenarios dealing with time-series signals and imbalanced data.

II. Problem Definition

WWTP industries are devoted to reducing the pollutant concentrations present in residual waters by means of highly nonlinear biological and biochemical processes. These processes generate some products derived from nitrogen and phosphorus which are harmful to the environment when presenting large concentrations. Therefore, certain limits are established by law [12]. Any effluent concentration exceeding those limits is treated as an effluent limit violation.

The application, in which this work is based, consists in the deployment of an ANN-based Soft Sensor able to predict WWTPs' pollutant concentrations. The main aim of this soft sensor is to help control strategies in the task of maintaining pollutant concentrations under the established limits. Thus, WWTP's performance can be improved decreasing its overall operational costs by means of predicting effluent limit violations and mitigating their effects in advance. In such a context, WWTP's influent and effluent measurements are considered as the ANN's inputs and targets, respectively. They consist in a whole year of the WWTP's influent and effluent measurements sampled every 15 minutes. These are generated by means of the well-known Benchmark Simulation Model No.2 (BSM No.2), a virtual model of a general purpose WWTP [12]. Further details about this problem can be found in [13].

In Fig.1 we schematize the problem at hand. The inputs are the influent and available measurements at the WWTPs, which are water flows and nutrient concentrations such as the ammonium concentration in the WWTP's first bioreactor tank ($S_{NH,po}$), the first clarifier's output flow (Q_{po}), the environmental temperature (T_{as}) and the recirculation flow (Q_a). Q_a corresponds to a flow transporting sludges from the last bioreactor tank to the first one. These inputs are adopted to predict the effluent concentrations, $\hat{y}_t$, by means of an ANN. Predictions are performed by two-stacked Long-Short Term Memory (LSTM) cells, a type of gated ANNs. Each cell considers 50 hidden neurons per ANN inside the cell [14, Chapter 10]. In this work, we focus on the concentrations of ammonium ($S_{NH,e}$) and total nitrogen ($S_{Ntot,e}$), two of the most pollutant nutrient concentrations present in the WWTPs. Peaks of ammonium ($S_{NH,e}$) and total nitrogen ($S_{Ntot,e}$) concentrations are rarely produced. Therefore, the prediction task carried out by the ANN-based Soft Sensor becomes a problem of predicting time-series showing rare events and highly imbalanced data.

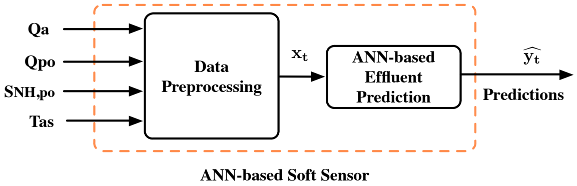

Fig. 1. ANN-based Soft Sensor. Input and output data corresponds to influent and effluent measurements.

As mentioned before, imbalanced data problem has been addressed applying resampling techniques when dealing with classification problems. In the case of regression tasks, evolutions and extensions of these techniques have been considered [4], [9]–[11]. However, they do not apply when dealing with time-series, where the time dependence of data has to be preserved. Because of this, the endeavor of this work and the applied data preprocessing techniques (sliding window, data normalization, K-Fold) will be focused on ensuring the temporal dependence of data as well as mitigating the effects of the imbalanced data problem at the training phase. In summary, the following considerations must be taken into account:

- Data Heterogeneity: inputs with different ranges will bias ANN response [15].
- Preservation of Time Correlation: most industrial applications require to preserve the time correlation of data. For instance, in WWTP scenario, influent-effluent relations depend on the temporal behavior of data. Besides, there exists a retention time which is defined as the time that residual water lasts inside the WWTP plant. It varies depending on the plant architecture.
- Preservation of information: some works address imbalanced problem by resorting to classification strategies. However, some applications need to keep the maximum amount of information as possible. In this work, for instance, predicted effluent limit violations are considered by control processes of the WWTP plant [13]: not only the detection of a violation is required, but also the value of the effluent concentration. Thus, data must be preserved as they are, making the preliminary classifications of the effluent data not feasible.
- Rare events: some industrial applications are focused on detecting rare events (e.g. predictive maintenance, faults detection, etc.). These events are difficult to be detected when dealing with imbalanced data. This is because ANN tends to learn more representative events.

III. Data Preprocessing

Data preprocessing techniques are considered in most of the ANN applications to improve the neural net performance and also to reduce its complexity [11]. As commented before, state-of-the-art techniques are not adequate enough when addressing regression problems with imbalanced data in industrial processes. For this reason, we present here data preprocessing mechanisms especially designed to fulfill four objectives: (i) reduce data heterogeneity, (ii) preserve time correlation, (iii) preserve information, and (iv) mitigate the imbalanced data problem caused by the appearance of rare events. Next, we present these mechanisms in detail:

A. Sliding Window

The implementation of a Sliding Window is proposed to organize the different time measurements and also to preserve the time correlation. The Sliding Window is defined by two parameters: the window length (WL), defined as the record data history considered for prediction in the LSTM cells; and the prediction horizon (PH), which consists in the amount of time the predictions can be given in advance. For the purpose of this work, WL and PH are configured as follows:

- A Window Length (WL) of 10 hours is considered to preserve not only the values seen at each sampling time, but also to those observed before.
- A Prediction Horizon (PH) of 4 hours is considered. It determines the amount of time that predictions of effluent concentrations will be given in advance.

The sliding window allows the ANN-based system to take into account not only the incoming but also the previously observed measurements. Performing this, the time correlation of data is considered providing more information to the ANN. Concerning its behavior, every time the sliding window slides, a new measurement is gathered removing the oldest one, i.e., a First-In First-Out policy is followed. Consequently, with each new measurement, the previously 10 hours of measurements are also considered (see Fig.2). Sliding window's parameters are set as such in order to fulfill the WWTP's retention time, which has been set to 14 hours. It is given by the considered WWTP's architecture. Further details on considered architecture can be found in [13, Fig.1].

B. Data Normalization

A data exploratory analysis was performed to get a first approach of the type of considered data. It was observed that influent and effluent measurements' ranges are widely spread showing highly heterogeneous data (see Tab.I). For example, Q_{po} values are several orders of magnitude over $S_{NH,e}$.

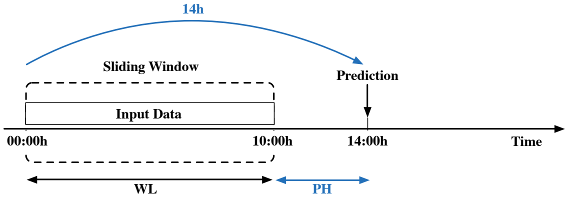

Fig. 2. Sliding window structure. Notice that the sum of the Window Length and the Prediction Horizon is equal to the WWTP retention time

TABLE I
Range of input and output signals

Input and output signals	
Parameter	Range
$S_{NH,po}$ (mg/L)	[9.84, 61.85]
Q_{po} (m^3/d)	[5593.90, 60425.00]
T_{as} $(°C)$	[9.68, 20.25]
Q_a (m^3/d)	[0.10, 309720.00]
$S_{Ntot,e}$ (mg/L)	[3.97, 24.21]
$S_{NH,e}$ (mg/L)	[0.16, 7.44]

Consequently, data normalization is proposed to instill homogeneity into data and therefore avoid the bias of ANN's output. Data normalization is performed by means of Z-score normalization technique. It is adopted instead of other simpler techniques (Min-Max Normalization) because the appearance of rare events can bias those simpler normalization techniques [15, Chapter 3]. Z-score normalization is computed as follows:

$$\mathbf{x_t} = \frac{\mathbf{x} - E[\mathbf{x}]}{\sqrt{E[(\mathbf{x} - E[\mathbf{x}])^2]}} \tag{1}$$

where $\mathbf{x}$ corresponds to the data to be normalized and $\mathbf{x_t}$ are the normalized data.

C. K-Fold based training

Histograms of effluent data (ANN's targets) were also computed to determine their distribution. They show that effluent limit violations take place with a very low probability. Consequently, predicting the effluents limit violations has become a task of rare events prediction. Histograms are shown in Fig.3 where ammonium ($S_{NH,e}$) and total nitrogen ($S_{Ntot,e}$) effluent concentration distributions are observed. Violations of ammonium (when $S_{NH,e} > 4$ mg/L) occur with a probability below the 0.23% whereas violations of total nitrogen (when

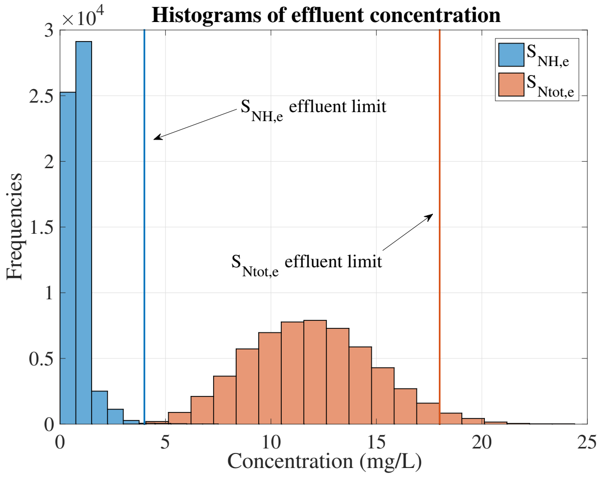

Fig. 3. Histogram of considered effluent measurements. Nearly all the values are below the established limits. Consequently, dataset is clearly imbalanced in terms of effluent concentrations exceeding the effluent limits.

$S_{Ntot,e} > 18$ mg/L) occur with a probability of 2.56%. Notice, that $S_{NH,e} = 4$ mg/L and $S_{Ntot,e} = 18$ mg/L are the effluent legal limits [12]. Consequently, effluent signals are a clear example of imbalanced data. If the imbalanced data problem is not treated accordingly, predictions can be biased toward the most represented values (those below the limits).

K-Fold has been proposed as one of the data preprocessing techniques at the learning stage to address the imbalanced data problem [16]. It is based on two principles: the division of the dataset into K equally sized data subsets; and the execution of K training processes. In our case, the dataset is defined as the influent and effluent WWTP measurements generated in BSM2 model. Besides, the number of folds (K) have been chosen in order to take a 70% of the whole dataset to train the ANNs and a 30% to test and validate them, where a 15% corresponds to the validation and another 15% to the test. The main goal is that a different prediction model is obtained for each training process, i.e., a total of K models are obtained. Each model has been derived adopting the same dataset with K-1 data subsets for training and one for testing and validation purposes. The main point here is that the subset devoted to test and validate the ANN-based Soft Sensor performance is changed at each training process. So, at the end of all the training processes, the model performing better in terms of predictions accuracy will be the one adopted in the final application. A K-Fold example is presented in Fig.4.

As a summary, K-Fold is able to determine the best data division to train the ANN-based Soft Sensor and therefore reduce the imbalanced dataset impact without resampling the data. Moreover, although K-Fold divides the data in different subsets, our proposal is respectful with the correlation of time for two reasons: (i) the consideration of the Sliding Window provides a tool to organize time measurements and maintain the time-dependence between values; (ii) data subsets considered for training tasks consist in consecutive measurements.

IV. Evaluation

The evaluation of the proposed data preprocessing techniques dealing with time-series industrial data is performed by means of the predictions carried out by the ANN-based Soft Sensor. To quantify its performance, three metrics are considered: Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (R^2). RMSE is computed as follows:

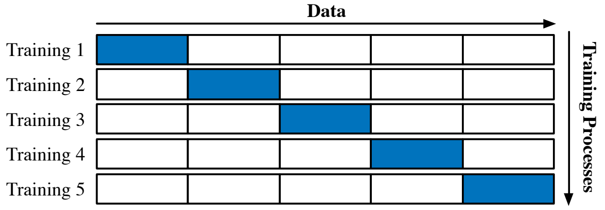

Fig. 4. K-Fold graphical example. At each training process, the colored subset is considered for ANN's testing purposes.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \tag{2}$$

where $\widehat{y_i}$ corresponds to the *ith* prediction of the effluent concentration and y_i to the target value. N corresponds to the number of effluent measurements. In terms of performance, the lower the RMSE, the better the predictions. MAPE, which is defined as the mean of the absolute percentage error performed in the prediction process, is computed as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \cdot 100 \tag{3}$$

Finally, R^2 metric measures the amount of data variance that can be explained by the ANN model. It ranges from 0 to 1, with 1 being a perfect correlation between predicted and real concentrations. It is measured as:

$$R^2 = \frac{(\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}}) \cdot (y_i - \bar{y}))^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 \cdot \sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4}$$

where $\bar{\hat{y}}$ corresponds to the mean of predicted values and $\bar{y}$ to the mean of targets.

In [4] and [10], approaches based on resampling techniques and SMOTE extensions are proposed. However, only those consisting in resampling techniques can be applied in this work. This is due to the highly correlation of time shown by our data. Besides, SMOTE extensions algorithms fail because they require more clusters than the ones achievable with the considered data. Therefore, five different scenarios have been considered during the evaluation process:

- Baseline (BL): Data have been only preprocessed considering the Z-score normalization.
- Baseline with Sliding Window (BSW): Baseline scenario with the Sliding Window. Neither K-Fold nor resampling techniques are applied.
- Oversampling (OS): Data have been preprocessed considering the Z-score normalization, the Sliding Window and oversampling techniques. It is equivalent to the oversampling strategy adopted in [10].
- Undersampling (US): Data have been preprocessed considering the Z-score normalization, the Sliding Window and subsampling techniques. It is equivalent to U_T strategy considered in [4].
- K-Fold (KF): our proposal where data have been preprocessed considering the Z-score normalization, the Sliding Window and K-Fold.

Results are shown in Tab.II. Focusing on the $S_{NH,e}$, the MAPE and RMSE values equal to 21.22% and 0.40 respectively when BL scenario is considered. These results can be improved considering the Sliding Window implementation to preserve the time correlation. Thus, MAPE and RMSE values are now equivalent to 7.87% and 0.24. However, predictions are performed taking into account imbalanced datasets. To solve this, OS, US and KF scenarios have been tested to

TABLE II
RESULTS OF THE EVALUATION PROCESS

Prediction	Scenario	RMSE	MAPE [%]	R^2
$S_{NH,e}$	BS	0.40	21.22	0.18
	BSW	0.24	7.87	0.85
	OS [10]	0.22	7.27	0.78
	US [4]	0.16	6.79	0.87
	KF	0.12	5.96	0.93
$S_{Ntot,e}$	BS	2.03	14.21	0.54
	BSW	0.76	4.69	0.95
	OS [10]	0.74	4.64	0.95
	US [4]	0.64	3.76	0.96
	KF	0.40	2.43	0.98

determine which one performs better. OS performance is similar to BSW's. US yields a MAPE and RMSE of 6.79% and 0.16 whilst KF scenario shows a MAPE and RMSE of 5.96% and 0.12, respectively. Imbalanced problem is faced with these techniques, however, time-correlation is broken in OS and US. Therefore, the scenario offering the best performance corresponds to KF, where K-Fold is considered. When K-Fold is applied, the performance of predictions is enhanced w.r.t BS: $S_{NH,e}$ RMSE and MAPE values are improved a 70% and 71.91%, respectively. R^2 is also improved showing that more variance can be explained by the ANN-based Soft Sensor's predictive model. The same applies to $S_{Ntot,e}$, where the RMSE and MAPE improvements correspond to 80.30% and 82.90%. It is worth commenting that these results are measured considering the predictions performed by the ANN-based Soft Sensor when the whole dataset is adopted as input data. Finally, the benefits of our proposal are more relevant when we focus on the rare events. This is shown in Fig.5, where it is observed that although there are some points which are not perfectly predicted, KF approach overcomes the other methods. Results of RMSE, MAPE and R^2 in this range are equal to: 0.10, 4.40% and 0.94 for $S_{NH,e}$; and 0.36, 2.34% and 0.99 for $S_{Ntot,e}$. Notice that US and OS approaches perform very badly in this case (KF RMSE for $S_{NH,e}$ is improved by a 68% w.r.t US and OS approaches).

V. CONCLUSIONS

In this work, data preprocessing techniques have been applied over industrial time-series signals corresponding to a WWTP scenario. In particular, data consist in a whole year influent and effluent signals where the latter are characterized by presenting rare events (imbalanced data) and high time correlation. These data have been considered in the training process of an ANN-based Soft Sensor whose aim is to predict effluent concentrations. In such a context, proposed techniques objective is to address the imbalanced data problem while preserving data information and time correlation.

Focusing on imbalanced data, conventional algorithms addressing this problem are not feasible since they are based on classifying and resampling the data. Consequently, their time correlation is lost. Therefore, this work proposes the adoption of a set of data preprocessing techniques to address the imbalanced data problem while preserving the time correlation. These techniques consist in Z-score normalization to homogenize data ranges, the implementation of a Sliding Window to sort the time-series measurements and the adoption of K-Fold technique to address the imbalanced data problem. The proposed approach is compared with different strategies found in the state-of-the-art. When focusing on rare events ranges, our proposal offers a 68% improvement on RMSE.

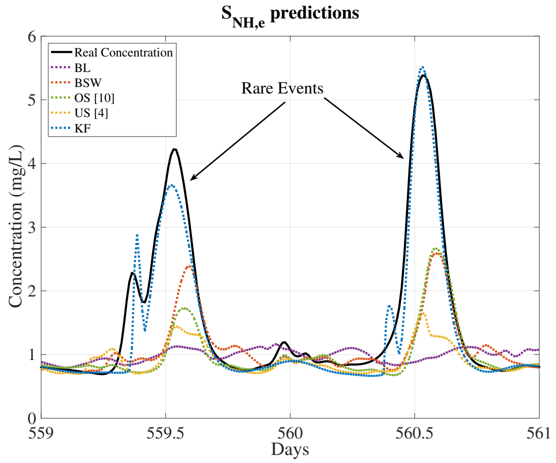

Fig. 5. $S_{NH,e}$ predictions for different considered scenarios. It is clearly shown that predictions are more accurate when K-Fold technique is adopted.

REFERENCES

[1] P. A. Sarvari, A. Ustundag, E. Cevikcan, I. Kaya, and S. Cebi, "Technology Roadmap for Industry 4.0," in *Industry 4.0: Managing The Digital Transformation*. Springer, 2018.

[2] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, 2017.

[3] S. N. Thennadil, M. Dewar, C. Herdsman, A. Nordon, and E. Becker, "Automated weighted outlier detection technique for multivariate data," *Control Engineering Practice*, vol. 70, 2018.

[4] N. Moniz, P. Branco, and L. Torgo, "Resampling strategies for imbalanced time series forecasting," *International Journal of Data Science and Analytics*, vol. 3, 2017.

[5] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, 2018.

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *IEEE International Joint Conference on Neural Networks*, 2008.

[7] S. del Río, J. M. Benítez, and F. Herrera, "Analysis of Data Preprocessing Increasing the Oversampling Ratio for Extremely Imbalanced Big Data Classification," in *Trustcom/BigDataSE/ISPA*, 2015.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE (Synthetic Minority Over-Sampling Technique)," *Journal of Artificial Intelligence Research*, vol. 16, 2002.

[9] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for regression," in *Portuguese conference on artificial intelligence*, 2013.

[10] L. Torgo, P. Branco, R. P. Ribeiro, and B. Pfahringer, "Resampling strategies for regression," *Expert Systems*, vol. 32, 2015.

[11] S. Naduvil-Vadukootu, R. A. Angryk, and P. Riley, "Evaluating preprocessing strategies for time series prediction using deep learning architectures," *The Thirtieth International Flairs Conference*, 2017.

[12] Gernaey, Krist V and Jeppsson, Ulf and Vanrolleghem, Peter A and Copp, John B, *Benchmarking of control strategies for wastewater treatment plants*. IWA Publishing, 2014.

[13] I. Santín, C. Pedret, R. Vilanova, and M. Meneses, "Advanced decision control system for effluent violations removal in wastewater treatment plants," *Control Engineering Practice*, vol. 49, no. 2, 2016.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press Cambridge, 2016, vol. 1.

[15] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[16] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, 2018.