World Academy of Science, Engineering and Technology
International Journal of Environmental and Ecological Engineering
Vol:15, No:3, 2021

A Machine Learning Approach for Anomaly Detection in Environmental IoT-Driven Wastewater Purification Systems

Giovanni Cicceri, Roberta Maisano, Nathalie Morey, Salvatore Distefano

Open Science Index, Environmental and Ecological Engineering Vol:15, No:3, 2021 publications.waset.org/10011922.pdf

Abstract—The main goal of this paper is to present a solution for a water purification system based on an Environmental Internet of Things (EIoT) platform to monitor and control water quality and machine learning (ML) models to support decision making and speed up the processes of purification of water. A real case study has been implemented by deploying an EIoT platform and a network of devices, called Gramb meters and belonging to the Gramb project, on wastewater purification systems located in Calabria, south of Italy. The data thus collected are used to control the wastewater quality, detect anomalies and predict the behaviour of the purification system. To this extent, three different statistical and machine learning models have been adopted and thus compared: Autoregressive Integrated Moving Average (ARIMA), Long Short Term Memory (LSTM) autoencoder, and Facebook Prophet (FP). The results demonstrated that the ML solution (LSTM) out-perform classical statistical approaches (ARIMA, FP), in terms of both accuracy, efficiency and effectiveness in monitoring and controlling the wastewater purification processes.

Keywords—EIoT, machine learning, anomaly detection, environment monitoring.

I. Introduction

THIS study is part of a regional research project (P.O.R Calabria) focusing on the development of an innovative EIoT (Environmental Internet of Thing) platform and a Smart Meter network, called Gramb, for monitoring, analyzing and control the quality of civil and industrial wastewater, upstream and downstream of the Public and/or Private purification circuits. The proposed solution integrates and interfaces with independently powered IoT devices, equipped with a Global Positioning System (GPS) and environmental sensors that allow real-time data collection on wastewater quality monitoring for the operation of the purification system. Thanks to the early warning algorithms, the proposed solution allows to monitor, control and optimise the quality, efficiency and effectiveness of environmental wastewater treatment processes. The platform can aggregate and process incoming data, then enforcing control policies and related actions.

The Calabria Region has a nominal purification capacity of 75% overall. In the Province of Catanzaro, 17 out of 22 purification plants are not working properly and only very few plants have quantities of sludge disposed compatible with the quantities of treated water, hence the need to monitor the inlet and outlet flow of wastewater at each plant to alert the authorities in the case of out of bound values in measurements. Monitoring of the flows entering and leaving the plant and, if necessary, those by-passed, will allow constant control not only of the treated flows, but also of what is discharged directly, over the norms, or illegally. A recurring problem in all the plants examined is the origin of the incoming wastewater, as the sewerage system is “hybrid”, since rainwater, irrigation water or even groundwater can be mixed within. Moreover, the consequent dilution of the load entering the plants, with pollutant concentrations much lower than those envisaged in the design phase, easily causes problems of settling as a result of the sludge exceeding the optimal rate of ascent compared to that envisaged in the design phase.

One of the most frequent situations is the drainage into the public sewerage system of effluents from various types of production activities with significant quantities of water from the alluvial aquifer. In many cases, the sewage treatment plant operator does not or cannot control the quantities of effluent discharged into the sewer. Another possible cause of hydraulic overload in a purification plant may be a discrepancy between the water consumption forecast and the actual one. Therefore, since it is not actually possible to check the state of the network upstream of the plant, monitor the inlet and outlet of the purification plants in real-time is a valid solution for triggering an immediate alarm to the authorities or simply to collect information useful for an analysis of the functionality of the plants or the identification of plants with anomalous inlet effluents in certain weather conditions. In such circumstances, the explosion of new technologies like Artificial Intelligence (AI) and different Machine Learning (ML) techniques play an important role and can be used to maintain complex systems. ML models allow to manage of large amounts of non-linear data and provide great accuracy in prediction, even with small volumes of observations. The goal of our analysis is to adopt and test different statistical and ML models for predicting the input and output values of sewage treatment plants in real-time, to also find anomalies in the system. We investigate how such algorithms can predict future values and find failures in the operating sewage plant, also identifying which parameters can improve the prediction process.

In summary, the main contribution of the paper is threefold: i) implementing real-time monitoring of system efficiency, quality and certification to better control the reliability

G. Cicceri is with the Department of Engineering, University of Messina, Messina, Italy (e-mail: gcicceri@unime.it).

R. Maisano is with C.I.A.M., University of Messina, Messina Italy (e-mail: robertam@unime.it).

N. Morey is with Microzone srl, Lamezia Terme, Italy (e-mail: natmorey@gmail.com).

S. Distefano is with the Department of Mathematics and Computer Science, University of Messina, Messina, Italy (e-mail: sdistefano@unime.it).

and robustness of the wastewater plant process; ii) create a system that helps the wastewater data review process while minimizing energy resources and reduce the exposure of personnel to time-consuming, repetitive and dangerous operations; and iii) demonstrate the ability of the ML approaches to overcome ARIMA and Prophet models to detect anomalies at the wastewater treatment plant on time series data;

The remainder of the paper is organized as follows: related works are discussed in Section II. In Section III, we introduce the techniques of data collection and the prediction models. Section IV presents the experimental tests and results, while Section V concludes the paper with some final remarks.

II. Related Works

Over the last decade, much research has been conducted to assess the wastewater monitoring and control to contribute to the balance of the environmental system [1]. Although the quality of depuration also depends on how the system is designed and implemented, the role of instrumentation, control and automation has become essential for the cost-effective and safe process operation [2]. In this direction, some of these methods have been used to control the process by verifying the correct functioning and reliability of the instruments through the prediction of the behavior and the detection of anomalies in the system.

The application of modern computer and online sensors technologies has vastly improved the performance in many wastewater treatment plants. Numerous approaches have been made to more accurately monitor and control wastewater quality analysis through forecasting environmental time series by the use of predictive modeling also with the use of neural networks. Advances combining sensor technology and information science have been becoming a focus in recent years [3] regarding also real-time data acquisition [4]. The information generated from online sensor devices can be used in forecasting models to accurately predict the sewage process treatment behavior for reuse applications. A neural network model may be useful in such cases because of its self-learning capability [5] [6] [7] [8]. Most of these models need several different input data, which are not easily accessible and make it a very expensive and time-consuming process. The application of a neural network model as a prediction tool to facilitate decision-making in effluent reuse applications was conducted using chemical monitoring parameters to build the neural network model, including also meteorological parameters such as rainfall index [9].

Wastewater monitoring and control patterns depend on a wide variety of factors. Determining an appropriate prediction model for wastewater charging behaviour is a highly specialised task. In literature, there are many models that rely on network specifications rather than generalization; among favorites include the Long-Short Term Memory (LSTM), the Auto-Regressive-Integrated-Moving-Average (ARIMA) [10] and other complementary models like recently Prophet [11]. Although the ARIMA model is proficient in forecasting daily load based upon the linear aspect of the data, it is not able to take into account for the non-linear aspects of the load time series, which represent the randomness induced by unaccounted emergencies and weather conditions [12].

The authors in [13], show the efficiency of ARIMA and ANN (Artificial Neural Network) models in predicting water quality parameters at a wastewater treatment plant, concluding that in all error estimates, ANNs models performed better than the ARIMA model. To uncover the non-linear aspects of time series, LSTM is usually preferred, which is a deep learning technique based on the RNN (Recurrent Neural Network) structure [14]. Authors in [12] show how this RNN architecture usually outperformed all other models in terms of accuracy when using metric like mean absolute percent error (MAPE). In this manuscripts [15], LSTM neural network has been used with the aim of detecting faults in wastewater treatment plants by overcoming traditional methods and enabling timely detection of collective faults. The recent model, named Prophet, has been recently used for the prediction and detection of anomalies in time series[16] [17], and with respect to ARIMA and LSTM, applications of this model have not been found in relation to sewage water treatment. Others specific algorithms based on principal component analysis [18] have been developed to support the efficient operation of wastewater treatment plants.

Although all these studies mainly focus on comparing different methods, as mentioned above, for time series prediction and anomaly detection like [19], they have not been applied to wastewater treatment plant monitoring in real-time fashion that can help the wastewater data review process, support decision making and speed up the processes of purification of water.

III. System Description

A. Sensing Infrastructure and Data Collection

For the experience, several purification plants with a capacity between 2,000 and 10,000 p.e. (population equivalent) were viewed. The selected treatment plant, which has recently undergone some technological modernisation, is located in the municipality of Briatico in Calabria (Fig. 1). It has a treatment capacity of 3.500 p.e.

Fig. 1 Wastewater treatment plant

The sewage water plant is designed for a maximum flow rates of 80-100 m3/h at the inlet of the treatment plant during winter. As the wastewater comes from the surrounding coastal tourist areas, this capacity can reach 400 m3/h during the summer period. The plant is equipped with some instruments for the acquisition of chemical-physical measures as:

- No. 1 Chemitec inlet flow meter model S103 Flow Meter (S103C):
  - Electromagnetic meter;
  - Output: 4 x 20mA outputs (configure 1 per flow value output);
  - RS485 Modbus.
- No. 1 Chemitec model 42Series (4204P) outlet flow meter:
  - Measuring unit: m3/h
  - Measuring range: 0 to 9999m3/h
  - Output: 4 x 20mA outputs (configure 1 per flow rate output)
  - RS485 Modbus
- No. 1 newly installed temperature/pH meter located at the inlet of the system
  - Unit of measurement: pH
  - Measuring range: 0 - 14pH
  - Working temperature: 0 - 80C
  - Output: 4 outputs 4..20mA (configured 1 for pH and 1 for temperature)
  - Output: RS485 Modbus
- No. 1 NO3/NH4 meter located in the purification treatment tank
  - Unit of measurement: mg/l
  - Output: 4 outputs 4..20mA (configured 1 for NO3 and 1 for NH4)
  - Output: RS485 Modbus

A Meter, called Gramb, was designed to acquire the chemical and physical parameters of the plant by connecting directly to the measuring PLCs installed in the purifier. The connection with the PLCs is established via an industrial protocol MODBUS on a standard RS485 bus (Fig. 2). The data received by the Meter are managed by the EIoT platform dedicated to the purification processes monitoring. The communication between the platform and the Gramb Meter is assumed via a virtual port on the server configured with the following parameters:

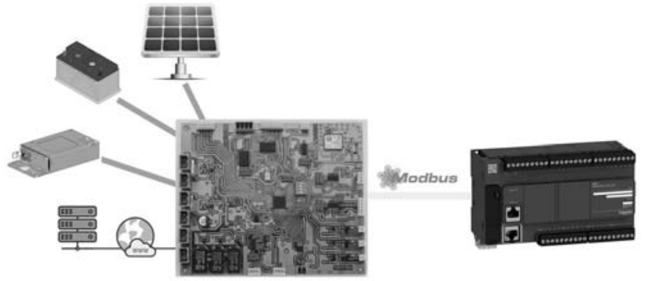


Fig. 2 Communication system between PLC and Gramb Meter

BAUD RATE= 19,200 - DATA BITS= 8 - STOP BITS= 1 - HANDSHAKE= NONE - PARITY= NONE

The platform acquires the information every half an hour during the day and every hour during the night. The collected data cover three fall-winter months of process operation. The parameters taken into account are inlet and outlet flow rates, temperature, ammonium and nitrates with the time of measurement. Data stored in a cloud are analyzed within the EIoT platform which, if necessary, alerts a possible alarm. Anomalies are detected thanks to particular predictive algorithms as described furthermore.

For the experiment, collected data consist of 2166 records.

B. Data Processing and Prediction Models

For our flow analyses, we chose a common time series model that is popular among data scientists and is represented by *Autoregressive Integrated Moving Average* (ARIMA). This model, compared to classical exponential smoothing statistical models based on a description of the trend and seasonality of the data, aims to make a description of the correlation between the data [20] with the ability to forecast future behaviour and, where this is not possible, to use linear models to predict the flow level in the future and on data with high variations. This model takes into account *three* key aspects of temporal information: *Auto-Regression* (AR), where observations have regressed on their previous values, *Integration* (I), where data values are replaced by the difference between values, and *Moving Average* (MA), where regression errors depend on lagged observations. ARIMA is used in forecasting both stationary and non-stationary time series data [21]. In the presence of non-stationary data, it is possible to make them stationary by putting them through differentiation responsible for the *integration* part [22]. Evaluation of auto-correlation is necessary to address the relative variability of the data.

The forecast problem in the time series was addressed by also comparing the *Prophet* model developed by Facebooks Core Data Science team [23]. Prophet is a forecast model for time series data based on an additive algorithm in which non-linear trends fit the seasonality of annual, weekly and daily, and is also suitable in all situations in which it is possible to make forecasts and detect anomalies [24]. In our approach, we leverage this model to try to see and predict the performance of the sewage plant and detect possible anomalies caused by external factors or malfunctioning of the same plant. The parameters used by Prophet take into account the linear or logistic growth curves for modeling non-periodic changes in the time series data, periodic changes like seasonality and also holidays. Prophet uses time as a regressor trying to fit different linear and non-linear functions of time as components, using their predictions as training in curve fitting rather than looking explicitly at the time-based dependence of each observation within a time series [25]. This model presents some advantages such as capturing seasonality over different periods; measurements do not have to be spaced regularly (and this is a good formula for our dataset), the interpolation of missing values are not necessary and the fit is very fast.

To complete our analysis, in this work we propose an *autoencoder* model based on a *Long Short Term Memory*

(LSTM) Recurrent Neural network (RNN). This method uses a LSTM Encoder-Decoder (LSTM-AE) architecture for sequence-to-sequence data [26]. Autoencoders are a type of self-supervised learning model that can learn a compressed representation of input data. It is widely used in time series, where anomalies provide significant information in critical situations and useful in the detection of unusual and unexpected records, but also in predictions and pattern recognition [27]. The configuration of a predictive LSTM-AE model is set to read an input sequence, encode, decode and recreate it. The ability of this model is evaluated by its good ability to perfectly recreate the input sequence [28]. This predictive model, in our case, could help us to predict the value of the future range by estimating it with the current data. Again, being a supervised approach, our goal is, once the prediction is made, to detect anomalies by comparing them with the actuals. For this use case, we used a neural network architecture. In particular, we applied a simple feed-forward neural network, in which the flow of information moves from the input to the hidden layers, to the output. The advantage that allowed us to use this type of architecture to detect anomalies in the system lies in the fact that we train a model on sequential data taking into account the error in the reconstruction phase, so when the model sees ”abnormal” data, it can tell by the increase in error in the reconstruction phase. As we will see in the Experimental Results section, this model is able to signal the failure of the plant by detecting when the flow rate readings begin to diverge from normal operating values.

IV. Experimental Results

A. Descriptive Analysis

The parameters took into analysis concern inlet and outlet flow rates, ammonium, nitrates and temperature. Data have been pre-processed for three months with an acquisition frequency taken every half hour (from 0:00 to 10:00) and others every hour (from 10:30 to 23:30) for a total of 38 daily observations daily. The descriptive analysis shows that the data have a high dispersion of the measured values with the presence of several outliers. All the variables are uncorrelated from each one except for the two flow rates, which are clearly correlated (Fig. 3). The water flow arriving in the inlet is processed within 30-45 minutes. In the case of an efficient wastewater treatment process, a smaller quantity is measured at the outlet after that time.

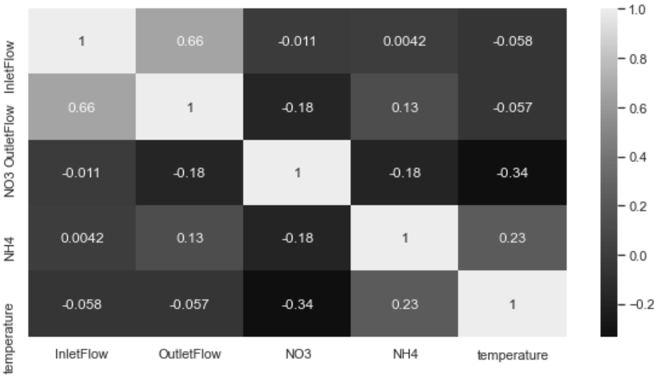


Fig. 3 Correlations among parameters

As illustrated in the introduction, control of the volumes entering the plant will allow constant monitoring not only of the malfunctioning of the wastewater treatment system but also of what is discharged directly, out of standard, or illegally. As the correlations between the parameters collected were low, subsequent analyses were carried out using only the inlet flow variable to focus on process behaviour over time as an indicator of the operation of the purification plant. Date of measurements and inlet flow rates will be the only variables taken into account in the dataset used for our experimentation.

B. Model Outputs

1) ARIMA: To perform the ARIMA model, python statsmodel library [29] was used. We first checked whether the time-series was stationary (a precondition for using this model) to apply it. Specifically, performing the augmented Dickey-Fuller statistic test, we verified the stationarity of the InletFlow variable (p-value $<$ 1%) as we can see also in Fig. 4 that shows the three components trend, seasonality and residuals of the time-series. We note that data are distributed with a stable trend with some high peaks of the Inlet Flow during December 2020.

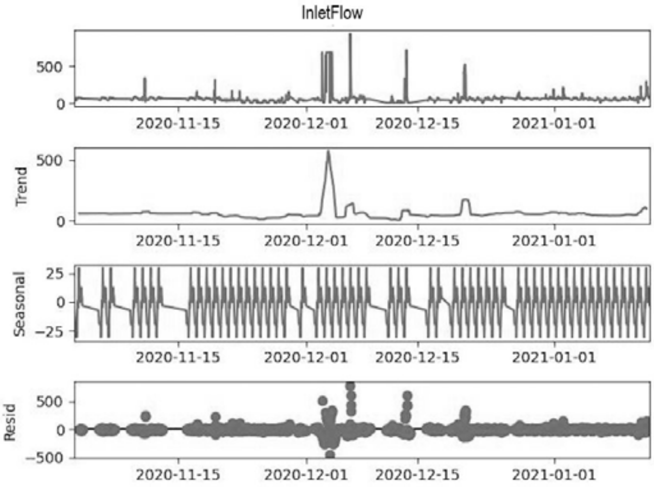


Fig. 4 Time-series components

The 2166 samples have been divided 80% into train set and 20% into a test set for the period: 2020-11-02 to 2021-01-12. Through the use of the *auto_arima* function, the optimal model that minimizes the *Akaike Information Criterion* (AIC) error was found, which allowed the range of the different parameters of p (autoregressive terms), d (order of difference) and q (order of moving average terms) to be optimised with values of (4,0,3) respectively. These parameters were used to adjust the final model. Fig. 5 shows the time-series trend with the test set prediction with 95% of the confidence interval. We observe that the ARIMA model is not a good predictor of future inlet flow trends for predicting changes (in intensity), especially for unpredictable out of range values, so we don’t predict anomalies for this data. ARIMA turns out to be a straightforward method by design but not powerful enough to predict signals and find anomalies in them.

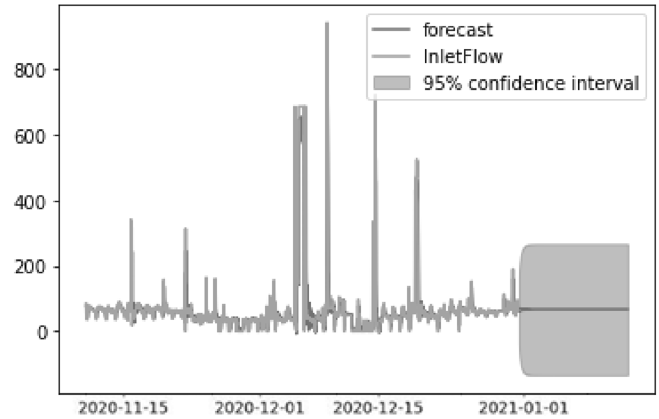


Fig. 5 ARIMA forecast

2) Prophet: An alternative to our anomaly detection task was the use of the Prophet model, whose forecasting task is more performant, especially for seasonality in the data. This model let us detect automatically change points in the time series, allowing us to factor hourly, daily, weekly and monthly trends. We performed hourly average resampling for the whole dataset thus, obtaining a training set of 1382 samples (up to day 2020-12-28) and a test set of 346 samples (from the day 2020-12-29). We set default *changepoints* that are deducted for the first 80% of the time series in order to project the trend-forward and avoid overfitting problems at the end of the time series.

In Fig. 6 we can see the result of the prediction where black dots are original data, blue line is the predicted values and the shaded blue area is the confidence interval of the predictions.

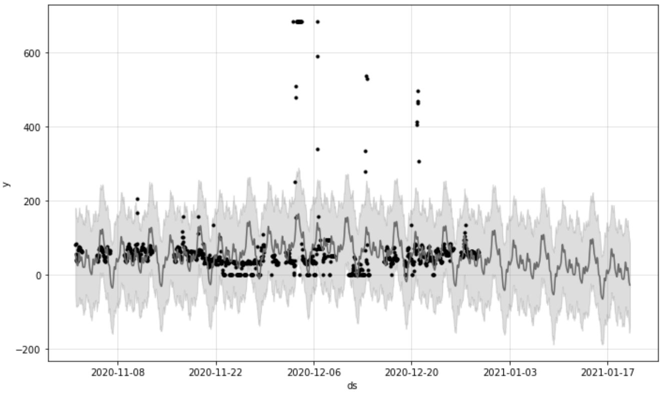


Fig. 6 Prophet forecast

Prophet predictions are not good due to the low amount of samples, even though it shows a reasonable seasonal trend, as visualized in the individual forecast components (Fig. 7). In fact, the forecast and components visualization shows that Prophet was able to accurately model both the underlying trend of the data as well as the weekly and daily seasonality. They highlight the daily trend of the quantities of water detected at the entrance of the treatment plant. This flow is low during the night periods, increases around 8 a.m., then peaks at periods 10-11 and 12-14 before decreasing in the evening. This trend is different at the weekend when the flows decrease overall over the whole day. This trend is verified in reality and effectively reflects the consumption habits of the households of the neighbouring population.

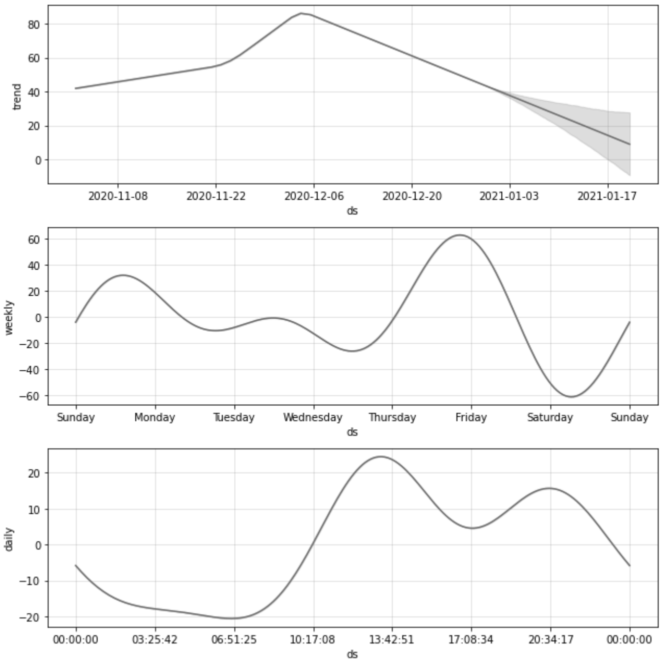
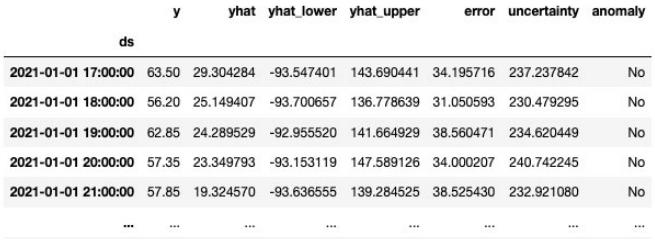


Fig. 7 Prophet components

At this point, Prophet creates a new dataframe assigned to the forecast variable that contains real data (**y**), the predicted values for test set data under the *yhat* column, as well as the *uncertainty* ranges and error of the forecast. Based on the *uncertainty* rate, we label a value as an anomaly when the absolute value of the error is greater than 30% of the uncertainty, as shown in Fig. 8.



	y	yhat	yhat_lower	yhat_upper	error	uncertainty	anomaly
ds							
2021-01-01 17:00:00	63.50	29.304284	-93.547401	143.690441	34.195716	237.237842	No
2021-01-01 18:00:00	56.20	25.149407	-93.700657	136.778639	31.050593	230.479295	No
2021-01-01 19:00:00	62.85	24.289529	-92.955520	141.664929	38.560471	234.620449	No
2021-01-01 20:00:00	57.35	23.349793	-93.153119	147.589126	34.000207	240.742245	No
2021-01-01 21:00:00	57.85	19.324570	-93.636555	139.284525	38.525430	232.921080	No
...	...	...	...	...	...	...	...

Fig. 8 Prophet anomaly dataframe

The graph in Fig. 9 corresponds to the time series with the anomalies highlighted in red. It shows that the model detects anomalies given by values far outside the range of regular operation of the plant and also detects other anomalies that, despite having a ”regular” value, are far from the model’s prediction in terms of predicted seasonality described above. The unexpected zero value, probably due to a plant fault, is not identified as an anomaly by the model, as well as some others high values observations close to 100m3/h (highlighted in Fig. 9).

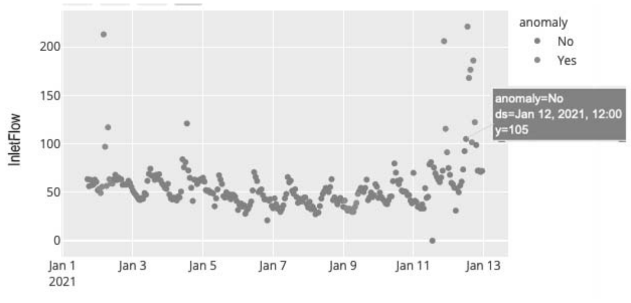


Fig. 9 Prophet anomaly detection

3) LSTM-Autoencoder (LSTM-AE): As the last model for our univariate analysis, we decided to use LSTM-AE implementation, which proves to be a good combination for regression forecasting and detect anomalies with sequential time series data. This model's architecture allows the use of encoder and decoder as LSTM networks, which allows learning from data over long sequences (thus capturing temporal dependencies) and is suitable for making predictions or detecting anomalies. The main assumption for this model is that on the distribution of the data there must be significantly different normal and abnormal data for the model to be proficient at taking a sample of input data, extracting all the important information and reconstruct the input to output, thus also being able to discriminate on outliers. To build our model, we used the same division of time series data of other models by taking 80% as the train set and the remainder as the test set (20%). Observations were taken every half hour and others every hour by choosing a temporal sequence of 10 data values (window) so that the model can learn from 10 previous values (from 5 to 10 hours before), trying to predict the next sequence values of the flow rate. To extract the temporal dependencies of one instance to another, in our implementation, we used 2 LSTM hidden layers for both encoder and decoder (Fig. 10). Being in a supervised approach, we feed the network with anomalous sequences. This turns out to be an advantage in that we help the network so that it can recognize these anomalies during input reconstruction, leading and error less than in a fully unsupervised approach. In this way, at autoencoder part is required to learn the most salient features of the training data. For our sequence reconstruction error, so in order to detect anomalies, we proceeded with calculating the Mean Absolute Error (MAE) on the test data and looking at the error distribution, we picked a threshold of 80% (Fig. 11). We declared an anomaly when the error is larger than this threshold.

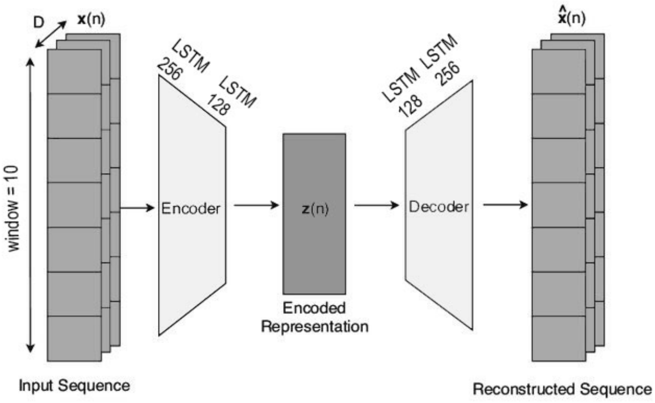


Fig. 10 LSTM-AE network

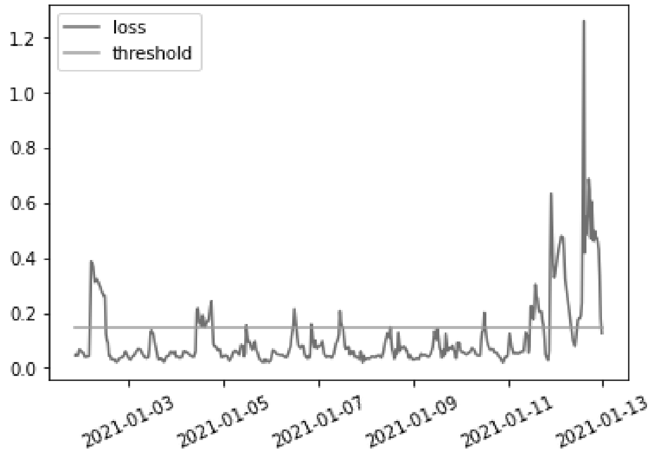


Fig. 11 LSTM - Loss with threshold

In Fig. 12, we can see the anomalies detected in the testing data.

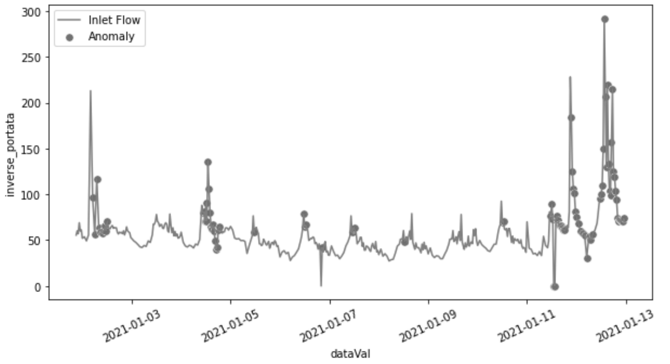


Fig. 12 LSTM-AE - Anomalies detection

The graph shows excess flow anomalies at the inlet of the purification plant. In reality, these high values are often linked to heavy rainfall that occurred in the previous half-hour or two hours (on December 27th, more than 17mm of water fell during the day, on January 4th more than 12mm and on January 12th, more than 33mm). Other anomalies, sporadic, for slightly high values are reported on days without rain. This could be an unusual and uncontrolled discharged of effluent into the sewer. January 1st is characterised by anomalies attributed to the identification of unexpectedly low values. The model encounters more insufficient data than those measured on working days. In reality, these values are consistent with those observed during the weekend marked by decreasing inflows. Finally, some station malfunctions with unexpected zero values are identified (like January 11th), demonstrating how the LSTM-AE approach is also well suited for predicting failures in the plant.

TABLE I
LSTM NETWORK CONFIGURATION

LSTM-AE architecture	
$Windows\ size$	10
$Layers$	4
$Neurons$	$[256, 128]$ $[128, 256]$
$Activation\ functions$	*Relu*
$Dropout$	$[0.3, 0.3]$
Hyperparameters	
$Maximum\ Training\ Epochs$	200
$Early\ Stopping$	$Pat = 10\ Monitor = loss$
$Optimizer$	*Adam*
$Batch\ Size$	32

C. Model Performance

Table II shows the comparison between the three models we performed for anomaly detection in terms of their metrics. In particular, we see that LSTM-AE achieves better RMSE, MSE, and MAE scores. These results confirm that applied LSTM networks with autoencoders are a promising approach for both anomaly detection and predictive maintenance. This approach can be adopted in modern smart industries for environmental protection, and it can be improved with careful network architecture and choice of hyperparameters to achieve better results.

TABLE II
PERFORMANCE METRICS COMPARISON

Forecasting metrics			
Models	**RMSE**	**MSE**	**MAE**
$ARIMA$	30.79	948.10	21.92
$Prophet$	50.70	2571.50	38.62
$LSTM - AE$	23.45	550.03	10.61

V. CONCLUSIONS

In this paper, the task of monitor and control a water purification system based on an EIoT platform was performed with the application of ML models with the aim to detect anomalies in the inlet flow entering and on the plant operation, in order to support decision making and speed up the whole purification system. Three different statistical and ML models have been applied with a subsequent comparison of them. In particular, we proposed and compared ARIMA, Prophet and LSTM-AE models to analyze failures and detect possible anomalies in the plant. Thanks to the Gramb Meter connected to the measuring PLCs installed in the plant, acquiring the chemical and physical parameters, we used the real-time dataset for testing both statistical and ML models. The ARIMA statistical model weakly adapts to the observations of the dataset, making it unsuitable for the detection of anomalies. The recent Facebook model Prophet is able to capture seasonality, with a weekly and daily trend in our dataset. The resulting Prophet prediction components really reflect the consumption habits of the households close to the plant area. Although Prophet is able to detect anomalies outside the range of regularity of operation, it cannot find some observations such as zero-value given by the system not working or others with values above the designed capacity limits of the system. LSTM-AE model, with its composite architecture, outperforms the first two models in terms of predictions and anomaly detection. From the results in fact we have seen that proper tuning of the hyperparameters of an LSTM-AE can lower the errors by having a good understanding of the input data and detect anomalies accurately, even not in the presence of a high volume of data. We were able to design a good LSTM-AE model, which confirm the superior performance in terms of RMSE, MSE and MAE error if compared with the other approaches. These results are encouraging, considering the limited data available and the possibility to improve the LSTM-AE approach with the addition of data and other features. In fact, we are planning to collect an extension of the dataset with a consequent multivariate analysis. Future work regards the improvement of the model architecture with the goal of reducing the error, compare it with other ML approaches that also make use of ANN, and test it in the same facility to verify the capabilities in monitoring and controlling wastewater treatment processes.

ACKNOWLEDGMENT

The authors would like to thank the managers of the S. Giorgio purification plant in the municipality of Briatico for allowing them continuous access to real data for the realisation of this study.

REFERENCES

[1] T. Zarra, V. Naddeo, V. Belgiorno, M. Reiser, and M. Kranert, "Odour monitoring of small wastewater treatment plant located in sensitive environment," *Water Science and Technology*, vol. 58, no. 1, pp. 89–94, 2008.

[2] G. Olsson, M. Nielsen, Z. Yuan, A. Lynggaard-Jensen, and J.-P. Steyer, *Instrumentation, control and automation in wastewater systems*. IWA publishing, 2005.

[3] R. Martínez, N. Vela, A. e. Aatik, E. Murray, P. Roche, and J. M. Navarro, "On the use of an iot integrated system for water quality monitoring and management in wastewater treatment plants," *Water*, vol. 12, no. 4, p. 1096, 2020.

[4] V. Edmondson, M. Cerny, M. Lim, B. Gledson, S. Lockley, and J. Woodward, "A smart sewer asset information model to enable an internet of things for operational wastewater management," *Automation in Construction*, vol. 91, pp. 193–205, 2018.

[5] C. Fu and M. Poch, "System identification and real-time pattern recognition by neural networks for an activated sludge process," *Environment International*, vol. 21, no. 1, pp. 57–69, 1995.

[6] M. Ct, B. P. Grandjean, P. Lessard, and J. Thibault, "Dynamic modelling of the activated sludge process: Improving prediction using neural networks," *Water Research*, vol. 29, no. 4, pp. 995 – 1004, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/004313549593250W

[7] J. C. Spall and J. A. Cristion, "A neural network controller for systems with unmodeled dynamics with applications to wastewater treatment," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 27, no. 3, pp. 369–375, 1997.

[8] K. Oliveira-Esquerre, M. Mori, and R. Bruns, "Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis," *Brazilian Journal of Chemical Engineering*, vol. 19, pp. 365 – 370, 12 2002.

[9] J. Chen, N. Chang, and W. Shieh, "Assessing wastewater reclamation potential by neural network model," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 2, pp. 149 – 157, 2003, applications of Artificial Intelligence for Management and Control of Pollution Minimization and Mitigation Processes. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197603000563

[10] E. Karakoyun and A. Cibikdiken, "Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting," in *The 13th Multidisciplinary Academic Conference in Prague*, vol. 2018, 2018, pp. 171–180.

[11] N. Zhao, Y. Liu, J. K. Vanos, and G. Cao, "Day-of-week and seasonal patterns of pm2. 5 concentrations over the united states: Time-series analyses using the prophet procedure," *Atmospheric environment*, vol. 192, pp. 116–127, 2018.

[12] V. Jadhav and V. Ligay, "Forecasting energy consumption using machine learning," *ResearchGate*, 2016.

[13] h. chioma, I. Howard, and E. Etuk, "Evaluation of arima and artificial neural networks in prediction of effluent quality of waste water treatment system." 10 2020.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] B. Mamandipoor, M. Majd, S. Sheikhalishahi, C. Modena, and V. Osmani, "Monitoring and detecting faults in wastewater treatment plants using deep learning," *Environmental Monitoring and Assessment*, vol. 192, no. 2, p. 148, 2020.

[16] K. Thiyagarajan, N. Ulapane *et al.*, "A temporal forecasting driven approach using facebooks prophet method for anomaly detection in sewer air temperature sensor system," 2020.

[17] B. Vishwas and A. Patel, *Prophet*, 08 2020, pp. 375–394.

[18] H. Haimi, M. Mulas, F. Corona, S. Marsili-Libelli, P. Lindell, M. Heinonen, and R. Vahala, "Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 65–80, 2016.

[19] H. Weytjens, E. Lohmann, and M. Kleinsteuber, "Cash flow prediction: Mlp and lstm compared to arima and prophet," *Electronic Commerce Research*, pp. 1–21, 2019.

[20] Y. Lai and D. A. Dzombak, "Use of the autoregressive integrated moving average (arima) model to forecast near-term regional temperature and precipitation," *Weather and Forecasting*, vol. 35, no. 3, pp. 959–976, 2020.

[21] K. Hollingsworth, K. Rouse, J. Cho, A. Harris, M. Sartipi, S. Sozer, and B. Enevoldson, "Energy anomaly detection with forecasting and deep learning," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4921–4925.

[22] V. Panasa, R. V. Kumari, G. Ramakrishna, and S. Kaviraju, "Maize price forecasting using auto regressive integrated moving average (arima) model," *Int. J. Curr. Microbiol. App. Sci*, vol. 6, no. 8, pp. 2887–2895, 2017.

[23] I. Yenidoan, A. ayir, O. Kozan, T. Da, and C. Arslan, "Bitcoin forecasting using arima and prophet," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018, pp. 621–624.

[24] K. K. R. Samal, K. S. Babu, S. K. Das, and A. Acharaya, "Time series based air pollution forecasting using sarima and prophet model," in *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, 2019, pp. 80–85.

[25] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[26] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[27] H. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management," *International Journal of Information Management*, p. 102282, 2020.

[28] S. Saumya, J. P. Singh *et al.*, "Spam review detection using lstm autoencoder: an unsupervised approach," *Electronic Commerce Research*, pp. 1–21, 2020.

[29] J. Perktold, S. Seabold, J. Taylor *et al.*, "Statsmodels: Statistics in python," *Internet: http://www. statsmodels. org/devel/generated/statsmodels. tsa. stattools. adfuller. html [August 12, 2019]*, 2017.

Giovanni Cicceri is currently a PhD student in Cyber Physical Systems (CPS) at the Department of Engineering, University of Messina (Italy). He is a member of the National Interuniversity Consortium for Computer Science (CINI) for the COVID19/IT Task Force. His research activity is interdisciplinary in Computer Engineering and Computational Finance and is focused on the study of Machine Learning and Deep Learning techniques targeted for multi-risk and multi-objective analyses applied in different contexts with particular reference to financial engineering, embedded systems and Internet of Things, and intelligent computing systems applied to health and life sciences. He is a collaborator on several multidisciplinary and application-oriented research projects.

Roberta Maisano is a Software Engineer and currently is assistant manager of Web and Mobile Technologies Staff Unit at University of Messina. She has a PhD in Technology and Economics of Processes and Products for Environmental Protection. Her research activity is interdisciplinary in Computer Science, Statistics and Chemical-physical processes. Her scientific articles focus on Data Science, Image Processing and Machine Learning topics. She is a member of CVPL(Italian Association for Computer Vision, Pattern Recognition and Machine Learning).

Nathalie Morey is a Software Engineer. She is currently collaborating for the Microzone S.R.L.S society where she participated to the Gramb project. She works on EIoT for environmental and territorial applications and on Google integration solutions for public administrations. Her research activity includes systemic, seismic and hydrogeological risk with the application of technologies like IoT and positioning system. She participated in various European project regarding industrial risk and environment.

Salvatore Distefano is an Associate Professor at the University of Messina (Italy). His research interests include Cloud, Fog, Edge, continuum computing, IoT, crowd-sourcing, Big Data, distributed ledgers, software and service engineering, performance and reliability evaluation and QoS. He is involved in several national and international projects. He is a member of international conference committees and journal editorial boards such as IEEE Trans. on Dependable and Secure Computing. He is the coordinator of the Italian CINI Task Force on Covid19. He has also co-founded the SmartMe.io startup.