Deep Learning Based Anomaly Detection in Water Distribution Systems

Kai Qian, Jie Jiang, Yulong Ding, Shuanghua Yang
Department Of Computer Science
Southern University of Science and Technology
Shenzhen, China
11849333@mail.sustech.edu.cn, {jiangj, dingyl, yangsh}@sustech.edu.cn

***Abstract*—Water distribution system (WDS) is one of the most essential infrastructures all over the world. However, incidents such as natural disasters, accidents and intentional damages are endangering the safety of drinking water. With the advance of sensor technologies, different kinds of sensors are being deployed to monitor operative and quality indicators such as flow rate, pH, turbidity, the amount of chlorine dioxide etc. This brings the possibility to detect anomalies in real time based on the data collected from the sensors and different kinds of methods have been applied to tackle this task such as the traditional machine learning methods (e.g. logistic regression, support vector machine, random forest). Recently, researchers tried to apply the deep learning methods (e.g. RNN, CNN) for WDS anomaly detection but the results are worse than that of the traditional machine learning methods. In this paper, by taking into account the characteristics of the WDS monitoring data, we integrate sequence-to-point learning and data balancing with the deep learning model Long Short-term Memory (LSTM) for the task of anomaly detection in WDSs. With a public data set, we show that by choosing an appropriate input length and balance the training data our approach achieves better F1 score than the state-of-the-art method in the literature.**

***Index Terms*—water quality, anomaly detection, sensor network, deep learning, sequence-to-point learning**

I. Introduction

Ensuring drinking water quality has always been the first priority of all countries. However, incidents such as natural disasters, accidents and intentional damages are endangering the safety of drinking water. For example, the 2014 Elk river chemical spill lead to 300,000 people without access to potable water. To keep drinking water quality to high standards, most water companies have deployed various kinds of sensors to monitor operative and quality indicators such as flow rate, pH, turbidity, the amount of chlorine dioxide etc., over different parts of the water distribution system (WDS) on a regular basis. Changes in these data serve as important indicators for detecting anomalies in WDSs, which facilitates early recognition of undesirable changes of the drinking water quality and enables the water supply companies to counteract the effects in time.

With water monitoring data publicly available, there has been a flourish in applying different computational methods for water quality anomaly detection. For example, the GECCO Industrial Challenge of Online Anomaly Detection for Drinking Water Quality[1] provided real-world monitoring data of drinking water recording both water quality and operative data, which attracted a lot of researchers. In the literature, there are in general three types of methods being applied to tackle this task. The first type of methods mostly relies on statistical models that will learn what pure water is from the history data. By comparing the quality difference between the pure water and the monitored water with a preset threshold, whether anomaly occurs can be determined. The second type of methods focuses on the application of the traditional machine learning models such as Logistic Regression, Support Vector Machine and Random Forest etc. More recently, researchers try to investigate the performance of deep learning models such as DNNs, RNNs and their variants. It was shown in [1] that the performance of the deep learning models (e.g. DNN, RNN) is worse than that of the traditional machine learning models (e.g. SVM, Logistic Regression). However, most of the existing applications of the deep learning methods to the problem of water anomaly detection do not take into account the characteristics of the WDS monitoring data.

To this end, in this paper, we investigate the characteristics of the WDS monitoring data and adapt the deep learning model Long Short-term Memory (LSTM) for the task of anomaly detection for drinking water quality. Firstly, given the fact that abnormal events are highly sparse in WDSs, i.e. WDS monitoring data are highly imbalanced, we investigate the application of data balancing techniques. Secondly, we investigate how the length of input sequences influences the model performance and show that by choosing an appropriate input sequence length the model performance can be improved. With a public data set, we show that by integrating data balancing and length selection our LSTM based approach achieves better F1 score than the state-of-the-art method in the literature.

The rest of the paper is organized as follows. Section II discusses the related work. Section III gives a formal description of the water anomaly detection problem and illustrates the data set used for evaluation in this paper. Sections IV and V introduce the baseline methods that are used for comparison and our proposed LSTM based method respectively. Section VI illustrates the process of data pre-processing, the experiment setup and analyses the experiment results. Finally, in Section

[1] http://www.spotseven.de/gecco/gecco-challenge/

VII, we conclude the paper with possibilities of future work.

II. Related Work

Traditional anomaly detection in WDSs involves sampling drinking water at random locations periodically and some following complicated manual laboratory analysis. By doing these, detailed reports about water quality can be provided to help decision makers determine if some anomalies occur in systems. However, such method apparently suffers some major drawbacks: (1) it is very labor-intensive and thus hugely increases the daily expense of monitoring departments; (2) it cannot give a real-time detection feedback, which is in fact very important for anomaly detection of drinking water since undetected contaminants will spread quickly in water pipes and possibly cause serious damages to people's life.

To overcome the drawbacks mentioned above, surrogate sensors measuring water quality indicators like pH, turbidity, the amount of chlorine dioxide and so on have been taken into account regarding drinking water quality monitoring. By placing these sensors at monitoring stations, water quality data can be collected at a low cost and transported to some central servers continuously for near real-time analysis. US Environmental Protection Agency has carried out an experimental analysis of these indicators and proved their effectiveness in anomaly detection of drinking water[2]. Since then, numerous studies have been done to develop new sensor data based anomaly detection algorithms to improve their accuracy and reliability.

In the early years, many simple threshold based detection algorithms are proposed by some researchers. For example, two anomaly detection methods for drinking water are given in [3]. In that work, four sensor indicators are considered for monitoring water quality: turbidity, oxidation reduction potential, pH, and electrical conductivity. Multivariate vectors and spider graphs are adopted to describe water quality in the two methods respectively and the difference between monitored water and pure water is calculated by Euclidean distance and area ratio respectively. Compared to preset thresholds, corresponding detection results are given.

Traditional machine learning techniques have been widely used to do anomaly detection in WDSs these years and some of them have achieved state-of-the-art results. In the work of [4], Artificial Neural Networks(ANN) are used to predict values of some indicators and the residuals between predicted values and real values can help determine whether some abnormal events happen in WDSs. In [5], SVM and Logistic Regression are both used as binary classifiers to determine if anomaly occurs. Their performance are carefully compared and the results show that Logistic Regression method has a good potential to be applied in real monitoring systems. In addition, some tree ensemble methods to detect anomalies in real-world water composition data set are proposed in [1] and among them, Extreme Gradient Boosting with Dropout gives the best and state-of-the-art results.

Recently, due to its success in many other engineering fields, deep learning has also been introduced as an anomaly detection method in WDSs. For example, in [5], RNN, LSTM and DNN are used to detect anomalies for drinking water. However, applying these techniques directly gives much worse results compared to other traditional machine learning techniques like SVM and Logistic Regression.

III. Problem Statement and Data Set

Water anomaly detection aims to estimate whether the water quality of a WDS is normal or abnormal based on the data obtained from the sensors deployed over the WDS. Formally, suppose we have a sequence of sensor readings denoted as $X = (x_1, x_2, \ldots, x_T)$ where $t \in \{1, \ldots, T\}$ is the sample index in the time domain, $x_t \in \mathbb{R}^n$ is a n-dimensional vector representing the readings from n different sensors at time t and T is the length of the sequence. The aim of water anomaly detection is to design a model to differentiate between normal and abnormal states of the water distribution system denoted as $Y = (y_1, y_2, \ldots, y_T)$, $y_t \in \{0, 1\}$, 0 indicates the normal state and 1 indicates the abnormal state.

The data set used in this paper is from the GECCO 2018 Industrial Challenge[6] supported by the water company Thüringer Fernwasserversorgung (TFW). The objective of this challenge is to develop an online monitoring system to detect remarkable changes or anomalies in WDSs. The monitoring system is expected to take online sensor data as input and give immediate results indicating whether there are abnormal events in the WDS in real time. To accomplish this, TFW takes measurements at significant points throughout the whole WDS, in particular at the outflow of the waterworks and the in- and out- flow of the water towers. For this purpose, a part of the water at these locations is bypassed through a sensor system where some water quality indicators, as well as some operational indicators, are measured on a minutely basis. After obtaining all the data measured in the real-world environment, TFW published the data after performing labelling, which allows researchers to apply existing or propose new binary classification algorithms to help them build effective and reliable monitoring systems.

Table I gives a summary of the indicators measured by TFW. The EVENT variable is the target variable y_t that needs to be predicted by learning the patterns of the other 9 indicators denoted as $x_t = \{x_t^1, x_t^2, \ldots, x_t^9\}$. Among these 9 indicators, the amount of chlorine oxide, the pH value, the redox potential, the electric conductivity and the turbidity are used to indicate water quality while the temperature and the flow rate are used to indicate operational situations. The time series data set used in this paper is provided by the GECCO 2018 Industrial Challenge, covering a time period from August 3, 2016 to February 13, 2017 with a total number of 279132 records. In this data set, 275077 records are labelled as False while only 4055 are labelled as True abnormal events, with a ratio of 1.452% positive instances. Thus the data set is extremely imbalanced. Figure 1 shows a one day snapshot of the 9 indicators with the abnormal events marked in red.

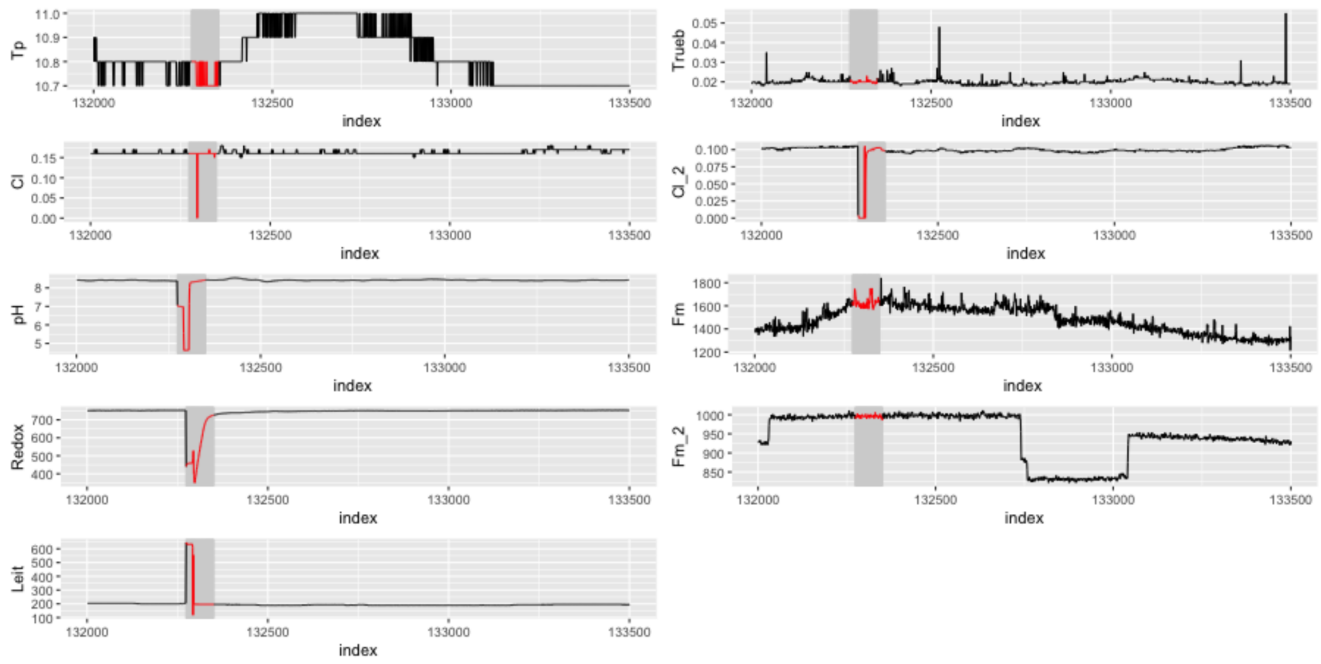

Fig. 1. Example of sensor data of one day with original event marked in red [6].

TABLE I
MEASURED INDICATORS IN GIVEN DATA SET AND THEIR DESCRIPTIONS

Indicators	Description
Time	Time of measurement, given in following format: yyyy-mm-dd HH:MM:SS
Tp	The temperature of the water, given in °C
Cl	Amount of chlorine dioxide in the water, given in mg/L (MS1)
pH	PH value of the water
Redox	Redox potential, given in mV
Leit	Electric conductivity of the water, given in µS/cm
Trueb	Turbidity of the water, given in NTU
Cl_2	Amount of chlorine dioxide in the water, given in mg/L (MS2)
Fm	Flow rate at water line 1, given in m3/h
Fm_2	Flow rate at water line 1, given in m3/h
EVENT	Boolean labels to indicate if it is an event

IV. BASELINES

In this paper, Logistic Regression and some tree ensemble methods are selected as baseline anomaly detection algorithms in WDSs as they have been proved to achieve the state-of-the-art performance[1, 5]. Logistic Regression (LR) is a popular machine learning technique that has been widely used in many engineering fields. When tackling classification tasks, it can provide probabilities to describe how much we can believe that each record in the data set should be classified into each corresponding category. Tree ensemble methods are also among the algorithms which achieve state-of-the-art results according to our knowledge. In this paper, tree ensemble methods refer to some methods in which different decision tree classifiers are combined into one single classifier for anomaly detection in WDSs. By taking results of different models into account, this ensemble classifier can significantly reduce bias and thus outperform all the single decision tree classifiers. Currently, there are many tree ensemble methods proposed and applied in many different engineering fields. Among them, Random Forest (RF)[7], Extreme Gradient Boosting (xgbTree)[8] and Extreme Gradient Boosting with dropouts (xgbDART)[9] have been proved to be effective in anomaly detection for drinking water[1] and will be tested in our work.

V. PROPOSED METHOD

A. Long Short-term Memory

Recurrent neural networks (RNNs) have many successful applications in modeling temporal signals, e.g., audio and speech signal processing [10] and natural language processing [11]. Similar to the fully connected neural networks, each input sample x_t is mapped to a hidden unit h_t by a transformation matrix. In addition, there are connections between adjacent hidden units to carry on the information from previous samples. In a non-causal system, a RNN can be bidirectional so as to use information from both history and future. A recurrent layer of a RNN can be described as:

$$h_t = \phi(Wx_t + Uh_{t-1} + b) \tag{1}$$

where W, U and b respectively represent the transformation matrix between input samples and hidden units, the transformation matrix between adjacent hidden units, and a bias term; ϕ represents a non-linear function. A RNN may consist of several recurrent layers. The backpropagation through time algorithm [12] is used for training a RNN.

One problem of the conventional RNNs is gradient vanishing/explosion [13]. This is because the depth of a RNN is proportional to the length of the input sequence. When training a RNN, the gradient will accumulate exponentially, which makes the training unstable. To solve the gradient explosion/vanishing problem, Long Short-term Memory (LSTM) was proposed, which introduces a memory cell with forget, update and output gates to control the information flow [14]. A LSTM is described as follows:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \odot \sigma_h(c_t). \end{aligned} \tag{2}$$

where f_t indicates the forget gate at time step t, i_t indicates the update gate at time step t, o_t indicates the output gate at time step t, c_t indicates the cell state at time step t, h_t indicates the final value for the memory cell at time step t, and σ represents a sigmoid (non-linear) function. The units that learn to capture short-term dependencies will tend to have forget gates frequently active while the units that learn to capture long-term dependencies will tend to have update gates frequently active.

The LSTM model we adopt in this paper consists of 5 LSTM layers with 128 hidden units for each layer and a fully connected layer followed by a sigmoid nonlinearity. A dropout of 0.5 is applied after each LSTM layer.

B. Sequence-to-point Learning

Sequence-to-point learning has been shown to work well on time series data [15]. As shown in Figure 2, a variant of sequence-to-point learning aims at finding a mapping from an input sequence $\mathbf{x} = (x_t, ..., x_{t+L-1})$ to a single target point y_{t+L-1} the index of which corresponds to the last point of the input sequence, where L indicates the length of the input sequence. This variant of sequence-to-point learning is only using the data from the past to ensure real time applications. By sliding through the whole input time series one step at a time, the target values are estimated. The length of the input sequences indicates how much history is used to infer the target value. When the input sequences are too short important information might be missing, while when the input sequences are too long there might be too much noise. To this end, it is necessary to investigate the influence of the input sequence length for training models.

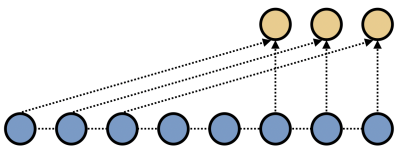

Fig. 2. Sequence to point learning.

For training a sequence-to-point binary classifier for WDS anomaly detection, the last layer of a LSTM is a fully connected layer followed by a sigmoid nonlinearity to represent the probability that there is anomaly in the WDS. Assuming an output and the corresponding target value are denoted as $\hat{\mathbf{y}}_t$ and $\mathbf{y}_t$ respectively, the loss can then be calculated using the binary cross-entropy:

$$\text{loss}(\hat{\mathbf{y}}_t, \mathbf{y}_t) = -(y_t \ \text{ln} \ \hat{y}_t + (1 - y_t)\text{ln}(1 - \hat{y}_t)). \tag{3}$$

The loss function is calculated on mini-batch data. After obtaining the loss, the gradient can be calculated and used to update the parameters of the model.

C. Data Balancing

Class imbalance indicates the situation where examples in training data belonging to one class heavily outnumber the examples in the other class. It has been reported that such situations might influence the performance achieved by learning systems [16]. As can be seen from Section III, the GECCO data set used in this paper is extremely imbalanced with the ratio of abnormal events being 1.452%. When training deep learning models, if most samples are negative instances it is likely that the models will have difficulties to learn the concept related to the minority class (in our case the abnormal events in WDSs). In the literature, there has been a lot of research on applying data balancing techniques for different problems [17]. For training LSTM models, a standard way of data balancing is to include more positive samples in the data that are fed into the LSTM models. For example, in this paper, we set a fixed ratio of 10% such that in each batch of data that are fed into the LSTM models there are 10%the positive samples.

VI. Experiments and Result Analysis

A. Data Preparation

Firstly, we filled the gaps in the data by a forward-filling method assuming that the gaps are caused by sensor communication failures, which results in a total number of 279,132 data points with each data point having 9 features as shown in Table I. Secondly, we normalised the data by subtracting the mean values and dividing by the corresponding standard deviations. Thirdly, we generate a training set, validation set and test set by splitting the data set with a ratio of 60%, 20% and 20%. As a result, we respectively have 167,479, 55,826, 55,827 data points for training, validation and test.

B. Evaluation Metric

In this paper, we use F1 score to evaluate the performance of different models as the dataset is extremely imbalanced (1.452% positive instances). F1 score [18] can be interpreted as a harmonic average of the precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{4}$$

where precision is the fraction of true positive instances among the predicted positive instances, while recall is the fraction of true positive instances over the total number of positive instances.

C. Experiment Setup

For Logistic Regression, L2 regularization is used as the penalty and a Newton-CG algorithm is applied to instruct the optimization process during training. For Random Forest, 20 decision trees are used to construct our model and the max depth for each tree is set as 5. Information gain is calculated based on Gini impurity to help generate decision trees. For Extreme Gradient Boosting with or without Dropout methods, we take 5 trees to construct the boosting models and set the max tree depth as 5. In addition, uniform sampling is applied during the training process. All these settings are carefully designed according to our rigorous experimental work.

For training the LSTM models, we use the Adam optimizer [19] with a learning rate of 0.001 to minimise the loss as shown in Equation 3. A batch size of 128 is used. These hyper-parameters are chosen experimentally. We used 1080Ti with 12 GB GPU memory to train all the LSTM models. We trained a group of LSTM models with respect to a range of sequence lengths including 10, 20, 30, 40, 50, 60, and the ratio of positive instances is set to 10% when applying data balancing.

D. Result Analysis

For both the baseline methods and the proposed LSTM method that combines the standard LSTM method with the data balancing technique, we report the F1 score evaluated on the test set with the model that achieves the best F1 score on the validation set. Table II shows the test F1 scores achieved by the standard LSTM models and the LSTM models with data balancing (balanced-LSTM) in terms of different sequence lengths. It can be seen that after applying data balancing, the performance of LSTM is improved for all sequence lengths. Among all the models, the best F1 score is achieved by the balanced-LSTM with a sequence length of 50 which corresponds to a time period of 50 minutes.

TABLE II
INFLUENCE OF SEQUENCE LENGTH AND DATA BALANCING

Sequence length	F1 - LSTM	F1 - balanced-LSTM
10	0.3505	0.4120
20	0.4084	0.4404
30	0.3923	0.4777
40	0.3483	0.6193
50	**0.4733**	**0.7819**
60	0.2206	0.4922

s

TABLE III
PERFORMANCE OF ALL TESTED ALGORITHMS ON GIVEN DATA SET

Model	F1 - validation set	F1 - test set
LR	0.3945	0.2841
RF	0.7271	0.2955
xgbDART	**0.7283**	0.2801
xgbTree	**0.7283**	0.2801
balance-LSTM	0.6219	**0.7819**

Figure 3 shows the predictions from all the methods together with the ground truth from the test set. The ground truth is indicated in orange with the value -1 while the predictions given by each method is indicated in blue with the value +1. It can be seen that the predictions from the proposed LSTM method match well with the ground truth while the three tree ensemble methods are more aggressive and output a large number of false positives.

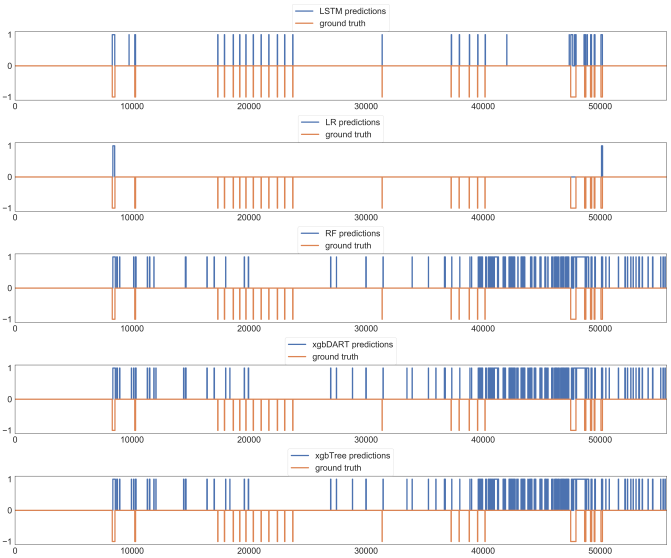

Fig. 3. Model predictions and true anomlies in the test set

VII. CONCLUSIONS

In this paper, we investigate the application of the deep learning model LSTM to the problem of anomaly detection in WDSs. In specific, we focus on the characteristics of the WDS monitoring data and integrate both sequence-to-point learning paradigm and data balancing techniques to improve the performance of LSTM models. By using a public data set, we show that our approach outperforms the state-of-the-art methods in the literature for anomaly detection for drinking water quality.

There are several directions for future work. Firstly, we intend to adopt other measures to improve the detection accuracy, e.g., building a hierarchical structure to learn patterns at different scales. Secondly, we are also considering to build an end-to-end solution from anomaly detection to source localisation for WDSs.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China under Grant Nos. 61873119 and 61911530247, and the Science and Technology Innovation Commission of Shenzhen under Grant Nos. KQJSCX20180322151418232.

REFERENCES

[1] M. Nguyen and D. Logofătu, "Applying tree ensemble to detect anomalies in real-world water composition dataset," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2018, pp. 429–438.

[2] J. Hall, A. D. Zaffiro, R. B. Marx, P. C. Kefauver, E. R. Krishnan, R. C. Haught, and J. G. Herrmann, "On–line water quality parameters as indicators of distribution system contamination," *Journal-American Water Works Association*, vol. 99, no. 1, pp. 66–77, 2007.

[3] T. P. Lambrou, C. C. Anastasiou, C. G. Panayiotou, and M. M. Polycarpou, "A low-cost sensor network for real-time monitoring and contamination detection in drinking water distribution systems," *IEEE sensors journal*, vol. 14, no. 8, pp. 2765–2772, 2014.

[4] L. Perelman, J. Arad, M. Housh, and A. Ostfeld, "Event detection in water distribution systems from multivariate water quality time series," *Environmental science & technology*, vol. 46, no. 15, pp. 8212–8219, 2012.

[5] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *Journal of Information and Telecommunication*, pp. 1–14, 2019.

[6] S. Chandrasekaran, M. Freise, J. Stork, M. Rebolledo, and T. Bartz-Beielstein, "Gecco 2018 industrial challenge: Monitoring of drinking water quality," 2018.

[7] H. Deng and G. Runger, "Gene selection with guided regularized random forest," *Pattern Recognition*, vol. 46, no. 12, pp. 3483–3489, 2013.

[8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[9] K. V. Rashmi and R. Gilad-Bachrach, "Dart: Dropouts meet multiple additive regression trees." in *AISTATS*, 2015, pp. 489–497.

[10] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE international conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proceedings of the NIPS 2014 Workshop on Deep Learning and Representation Learning*, 2014.

[12] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[13] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1310–1318.

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

[15] C. Zhang, M. Zhong, Z. Wang, N. Goddard, and C. Sutton, "Sequence-to-point learning with neural networks for nonintrusive load monitoring," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 2604–2611.

[16] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004. [Online]. Available: http://doi.acm.org/10.1145/1007730.1007735

[17] L. O. H. W. P. K. Nitesh V. Chawla, Kevin W. Bowyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, no. 16, pp. 321–357, 2002.

[18] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE Computer Society, 2013, pp. 245–251.

[19] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2014.