Big Data Analytics for Advanced Fault Detection in Wastewater Treatment Plants

Morteza Zadkaramia,b, Krist V. Gernaey\*,b, Ali Akbar Safavia, Pedram Raminb

aSchool of Electrical and Computer Engineering, Shiraz University, Iran

bProcess and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), Denmark

*kvg@kt.dtu.dk*

Abstract

Fault detection in wastewater treatment plants (WWTPs) presents difficult challenges, highlighted by the nonlinear, nonstationary nature of operations and the varying fault intensities that often are neglected. While big data analytics promise transformative results, they come with their own challenges in process monitoring such as handling vast datasets, ensuring real-time responsiveness, and coping with imbalanced data distributions. With our 609-days simulation of the Benchmark Simulation Model 2 (BSM2), yielding datasets as expansive as 876,960 samples for each of the 31 measurements considered here, the inherent issues become more obvious. To address this challenging process monitoring problem, our research introduces a novel fault detection framework, handling both imbalanced data distribution and big data complications. The core of this framework includes two critical components. The first one is a wavelet-based feature analyzer which utilizes the wavelet energy and entropy information for each measurement to extract the most valuable and critical features. The other element is the enhanced neural network classifier which deals with imbalanced data distribution. This classifier partitions the data into multiple segments, subsequently determining the BSM2 operational condition (i.e. normal or fault) for each distinct segment. The proposed detection framework has demonstrated the capability to accurately identify the operational condition of the large-scale BSM2 dataset, achieving a False Alarm Rate (FAR) of less than 10%. The promising results obtained from this framework can facilitate future research on developing digital twins for WWTPs.

**Keywords**: Wastewater Treatment Plants (WWTPs), Process Monitoring, Big Data Analytics, Imbalanced Classification, Wavelet Analysis

* 1. Introduction

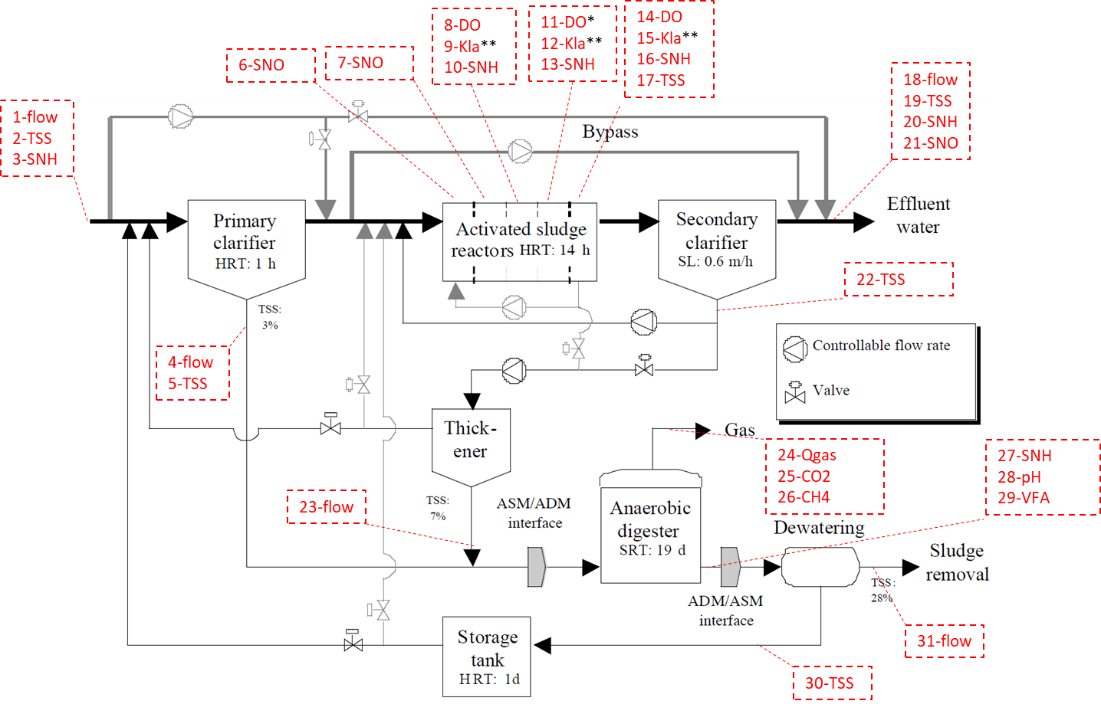
Wastewater Treatment Plants (WWTPs) play a key role in managing reclaimed water by removing nitrogen and organic matter. Therefore, it is important to develop fault detection systems for these plants to prevent harmful substances from being released into the environment. WWTPs include a series of biological, physical, and chemical reactions, leading to a plant-wide system inheriting nonlinear, nonstationary, auto- and cross-correlated characteristics (Liu et al., 2023).

Every fault detection approach possesses unique advantages and shortcomings. Therefore, the choice of a suitable fault detection framework depends on the specific case study and the considered objectives. Univariate fault detection approaches are among the most popular techniques applied on WWTPs because of the simple and straightforward nature. Sánchez-Fernández et al., (2018) adopted a modified version of the Exponentially Weighted Moving Average (EWMA) to detect various abnormalities including sensor faults, alkalinity variation, and pipe leaks. Multivariate fault detection methods, particularly based on Principal Component Analysis (PCA), are reported to be near 40% of the published fault detection studies on WWTPs (Liu et al., 2023). Ramin et al. (2021) implemented dynamic PCA on a WWTP containing process and instrument (i.e. sensor and/or actuator) faults. The study showed that the proposed dynamic PCA had promising precision. Kazemi et al., (2021) detected faults in an anaerobic digestion system by initially designing a soft sensor model for the Volatile Fatty Acids (VFA) based on influent and effluent characteristics. Once the VFA soft sensor task is established, the fault detection assessment was carried out using univariate and multivariate approaches. Despite the fact that fault detection approaches based on univariate and multivariate concepts are easy to implement, investigate and evaluate, they are mostly based on a Gaussian assumption which can lead to restrictions in practice since data typically do not follow a symmetric unimodal Gaussian distribution. To overcome the aforementioned issue, Zadkarami et al., (2023) suggested to employ discriminative classifications. Thus, these data-hungry models require big data sets to gain assuring results.

Thanks to the advancements in technology, data collection, storage, and processing can be done easily and rapidly, allowing us to reap the benefits of big data sets. The correct term of big data is a controversial topic in the literature since for different applications in the real world the definition varies. Nevertheless, based on a recent study (Gandomi and Haider, 2015), a dataset can be recognized as big data if it has at least one of the 5V characteristics as follows: Volume (i.e., large amount of data samples), Velocity (i.e., high rate of incoming data),Veracity (i.e., the data contains significant uncertainty and noise), Variety (i.e., the dataset is constructed by smaller datasets from different types and formats), Variability (i.e., the degree of complexity the data possesses, e.g. whether it contains noticeable fluctuations and peaks). In this study, big data analytics are necessary since there is a large quantity of data samples (i.e. 876,960), and the data also includes noise effects and rapid changes throughout the data pattern. Although big data can boost the modelling performance to a large degree, it comes with its own challenges including redundancy, curse of dimensionality, and high computational requirement.

In this paper, a novel fault detection framework is designed for a WWTP based on the Benchmark Simulation Model 2 (BSM2) considering big data aspects. To reflect the real-world application, the duration of faulty conditions is considerably shorter than the normal operation. This is affirmative for both process and instrument fault scenarios. Therefore, this research work deals with imbalanced big data collected from a highly sophisticated biochemical system. To address the data complexity, two main steps are introduced. First, a wavelet-based feature extraction structure is utilized which finds the best wavelet approximation and wavelet detail feature combinations to be inserted to a Multi-Layer Perceptron Neural Network (MLPNN) classifier. Since the imbalanced data (availability of more normal conditions) tend to be reflected in imbalanced MLPNN training, the second stage is to design a window-based classifier modification algorithm to put more emphasis on the smaller class (here the faulty conditions).

The rest of this paper is organized as follows: Section 2 describes the case study. The proposed fault detection methodology is thoroughly explained in Section 3. The obtained results along with discussions are presented in Section 4. Finally, Section 5 concludes the research.



**Figure 1.** The BSM2 diagram along with the location of the measurements

* 1. The Case Study

Jeppsson et al. (2007) reported a full description of the BSM2, which is the most complete model of a practical WWTP. The BSM2 was extended with the capability to consider uncertainty and noise. In this study, an updated BSM2 model (Ramin et al. (2021)) was used which is designed for monitoring and fault detection purposes. The model was simulated for 609 days with a sampling time of 1 minute for 31 measurements, resulting in a sampling size of 876,960 for each measurement. To extract data from BSM2, different locations were selected, considering process parameters that are relatively easy-to-measure in those locations (Figure 1). The simulation is implemented on two fault scenarios including a process and an instrument fault. The simulation for the sludge bulking phenomenon, which falls into the process fault category, contains 822,765 (93.82%) normal condition data points and 54,195 (6.18%) faulty data points, resulting in an imbalance ratio (number of normal data / number of faulty data) of 15.18. On the other hand, the instrument fault representing an abnormality in the aerated reactors regarding the airflow has an imbalance ratio of 15.57 and contains 824,048 (93.97%) normal condition data points and 52,912 (6.03%) faulty data points.

* 1. The Fault Detection Framework

To address the big data and imbalanced data distribution issue, a novel fault detection framework is introduced. After normalizing each measurement to zero mean and unit variance, they pass through the designed feature analysis approach (Figure 2). The extracted features are then inserted into the developed classification architecture for decision making. For further details on wavelet analysis, one may refer to the research by Pourahmadi-Nakhli and Safavi (2011). Afterwards, the feature sets are inserted into an MLPNN classifier with two hidden layers respectively containing 25 and 15 neurons with sigmoid activation functions. The stopping criteria for this network is either the error reaches less than 0.01 or the iteration number exceeds 200.

|  |
| --- |
|  |
| **Figure 2.** The designed feature extraction flowchart |

Since the MLPNN, like most of the typical discriminative classifiers, tends to provide misleading results for imbalanced data distributions, a window-based classification correction structure (Algorithm 1) is designed. The main purpose of Algorithm 1 is to pay more attention to the smaller class (Faulty class) in each segment and modify the collected initial results from the MLPNN. The main idea behind Algorithm 1 is that the working operational changes for both process and actuator faults are assumed to be considerably slow. Therefore, if for a specific period of time a certain number of faults occur, one may choose to assume that the whole time period can be assigned as faulty conditions.

* 1. Results and Discussion

The last stage in every fault detection classification method is to assess the outcome labels using the confusion matrix (see Figure 3).

|  |
| --- |
|  |
| **Figure 3.** The Confusion Matrix |

In the literature, the following criteria obtained from the confusion matrix are frequently employed (Susan and Kumar, 2021):

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |
|  | (4) |
|  | (5) |
|  | (6) |

For an ideal case, Eqs. (1-2) would be zero, while Eqs. (3-6) equals to unity. Table 1 shows the obtained fault detection results for different fault types along with considering various window sizes. It is obvious from Table 1 that for an imbalanced data distribution, such as this case study, exclusively investigating the accuracy, precision, and recall factors are not considered as reasonable criteria since they do not undertake the FAR explicitly in their calculations. On the contrary, provides plausible assessment. By examining the calculated index for each scenario, it is clear that the developed classifier correction structure can handle the imbalanced nature of the classes to a certain extent. Furthermore, in our study, the window size of 6 hours (360 samples) led to the most superior performance in detecting process and instrument faults. In fact, a promising FAR of approximately 4% and 10% for respectively the process fault and the instrument fault was obtained.

**Table 1.** Different fault detection results based on their correction window size

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fault Type** | **Window Size** | **MDR** | **FAR** | **Precision** | **Recall** | **Gmean** | **ACC** | **Time (sec.)** |
| Process Fault | -- | 0.0021 | 0.3755 | 0.9758 | 0.9979 | 0.7894 | 0.9748 | 1889 |
| 3 hours | 0.0012 | 0.058 | 0.9961 | 0.9988 | 0.9699 | 0.9953 | 318 |
| 6 hours | 0.0017 | 0.0365 | 0.9975 | 0.9983 | 0.98 | 0.9961 | 326 |
| 12 hours | 0.0057 | 0.0425 | 0.9971 | 0.9943 | 0.9757 | 0.992 | 303 |
| Instrument Fault | -- | 0.0013 | 0.5404 | 0.9664 | 0.9987 | 0.6774 | 0.9661 | 1605 |
| 3 hours | 0.0064 | 0.1036 | 0.9933 | 0.9936 | 0.9437 | 0.9877 | 276 |
| 6 hours | 0.0105 | 0.0927 | 0.994 | 0.9895 | 0.9475 | 0.9845 | 289 |
| 12 hours | 0.0087 | 0.1483 | 0.9904 | 0.9913 | 0.9188 | 0.9828 | 320 |

Only a few studies have conducted fault detection for the BSM2. Sánchez-Fernández et al., (2018) considered 140 measurements, many of which were inaccessible or redundant. Kazemi et al., (2021) focused only on faults in the anaerobic digestion unit, not the entire system. Ramin et al. (2021) suggested more realistic measurements and used dynamic PCA to examine their benchmark, achieving a precision index of 0.704 for process faults and 0.888 for instrumentation faults. However, the accuracy was unsatisfactory, with 0.651 for process faults and 0.432 for instrumentation faults. This study improves the BSM2 design of Ramin et al. (2021), making it more practical and challenging. The performance evaluation indicates that the new fault detection approach effectively handles the complexities and difficulties associated with the BSM2.

|  |
| --- |
| **Algorithm 1.** The windowed-based classifier correction algorithm |
| 1. Select a window size 2. Get the labels from the MLPNN classifier output for the current window 3. Select a threshold 4. **If** ‘the number of labels classified as faulty’ **>** threshold   Set the whole labels of that window as faulty condition  **else**  Set the whole labels of that window as Normal condition   1. Record the updated condition labelling for that window 2. **If** ‘the number of collected data’ = 876,960 samples (609 simulation days)   Set the modified output labels as the final decision condition  **else**  Move to the next window and go to step 2   1. Validation |

* 1. Conclusion

This research study introduced a novel fault detection framework for wastewater treatment plants based on the BSM2, addressing challenges of big data and imbalanced data distribution. The framework utilized a wavelet-based feature analyzer and an enhanced classifier. It was tested for both process and instrument faults individually, achieving a FAR of less than 10% in the optimum case. Future studies can be carried out in terms of simultaneous fault sources or investigating other approaches to address the imbalanced data nature of such systems.

References

A. Gandomi, M. Haider, 2015. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage, 35, 137–144.

A. Sánchez-Fernández, F.J. Baldán, G.I. Sainz-Palmero, J.M. Benítez, M.J. Fuente, 2018. Fault detection based on time series modeling and multivariate statistical process control. Chemometrics and Intelligent Laboratory Systems, 182, 57–69.

M. Pourahmadi-Nakhli, A.A. Safavi, 2011. Path characteristic frequency-based fault locating in radial distribution systems using wavelets and neural networks. IEEE Transactions on Power Delivery, 26, 772–781.

M. Zadkarami, A.A. Safavi, K.V. Gernaey, P. Ramin, O A. Prado-Rubio, 2023. Designing a fault detection classifier framework for an industrial dynamic ultrafiltration membrane process using wavelet-based feature analysis Process Safety and Environmental Protection, 174, 1-19.

P. Kazemi, C. Bengoa, J.P. Steyer, J. Giralt, 2021. Data-driven techniques for fault detection in anaerobic digestion process. Process Safety and Environmental Protection, 146, 905–915.

P. Ramin, X. Flores-Alsina, S. Olivier, N. Topalian, U. Jeppsson, K.V. Gernaey, 2021. Fault detection in a benchmark simulation model for wastewater treatment plants. Computer Aided Chemical Engineering, 49, 1363-1368.

S. Susan, A. Kumar, 2021. The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. Engineering Reports.

U. Jeppsson, M.N. Pons, I. Nopens, J. Alex, J.B. Copp, K.V. Gernaey, C. Rosen, J.P. Steyer, P.A. Vanrolleghem, 2007. Benchmark simulation model no 2: General protocol and exploratory case studies. Water Science and Technology, 56, 67–78.

Y. Liu, P. Ramin, X. Flores-Alsina, K.V. Gernaey, 2023. Transforming data into actionable knowledge for fault detection, diagnosis and prognosis in urban wastewater systems with AI techniques: A mini-review. Process Safety and Environmental Protection, 172, 501–512.