2022 25th International Conference on Computer and Information Technology (ICCIT)
17-19 December, Cox's Bazar, Bangladesh

Detecting Faulty Machinery of Waste Water Treatment Plant Using Statistical Analysis & Machine Learning

Md. Mazed Ul Islam
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
md.mazed.ul.islam@g.bracu.ac.bd

Joyanta Jyoti Mondal
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
joyanta.jyoti.mondal@g.bracu.ac.bd

Ibne Farabi Shihab
Department of Computer Science
Iowa State University
Ames, IA, United States
ishihab@iastate.edu

***Abstract*—The goal of wastewater treatment is to eliminate contaminants from wastewater and convert them into effluent/discharge that can be reintroduced into the water cycle. In order to monitor, analyze plant performance, and decrease environmental pollution in wastewater treatment facilities, a model for fault detection must be developed. In this study, we examine different time and cost-efficient machine learning approaches to monitor the operation of a Waste Water Treatment Plant (WWTP) and identify plant faults as an alternative to human, laboratory-based time consuming, costly, and challenging techniques. This will allow us to develop a time and cost-efficient approach to detect such problems. To discover plant defects, we collect one year of unsupervised WWTP data and convert the data into supervised data. Using several machine learning algorithms based on water quality standard measurements (pH, BOD, COD, and suspended solid), we establish whether or not the data is valid.**

***Index Terms*—Waste Water Treatment Plant, Machine Learning, Fault Detection, Unsupervised data, Supervised data**

I. Introduction

The term "wastewater treatment plant" (often abbreviated as "WWTP") refers to a facility that utilizes a number of different processes (such as physical, chemical, and biological) in order to treat wastewater from industrial facilities and eliminate contaminants. This treatment plant's operations are broken up into three distinct phases: pre-treatment, primary treatment, and secondary treatment. It is a dynamic and complicated system that is used in a broad variety of sectors, including the chemical, textile, medical, and leather industries, amongst others. The goal of the treatment process at a WWTP is to purify industrial waste water so that it may be reused in the manufacturing process. It is required for businesses to have a Wastewater Treatment Plant (WWTP) or an Effluent Treatment Plant (ETP) in different nations. As will be shown in the future, several industries have constructed their own WWTPs, however, it is challenging to manage such a vast, dynamic, and complicated system. If the system sustains damage or stops functioning as it should, locating the source of the problem becomes more challenging, as well as time- and labor-intensive. So, in this research, we investigate a feasible, cost-efficient, and automated solution to solve this issue. In our study, we analyze approaches to determine whether a WWTP is functional or not. Occasionally, a WWTP may experience disruption and may not work efficiently. Hence, we use Multivariate Statistical approaches to identify faults. Moreover, we use different Machine Learning (ML) algorithms to predict the fault pattern of the treatment plant.

Based on our study, we make the following set of contributions to this paper:

- We predict the performance of wastewater plants which can significantly contribute to reducing environmental pollution.
- We evaluate different effective machine-learning approaches to monitor WWTP's performance.

II. Background

A. Important Parameters:

1. Potential of Hydrogen (pH): The pH scale measures how acidic or basic the water is. The range of pH is between 0-14 where from 0 to below 7 is represented as acidic. And if pH is above 7, then it is represented as basic.

2. Biological Oxygen Demand (BOD): The term "Biological Oxygen Demand" refers to the average amount of oxygen required for the breakdown of microorganisms. It is regarded to be in excellent condition when the BOD level is between 1-2 ppm. If it has a concentration of more than 100 ppm, then it is considered to be highly contaminated.

3. Chemical Oxygen Demand (COD): The procedure for measuring the chemical oxygen demand is both quick and accurate. The maximum concentration of COD that is allowed is 250 mg/l.

4. Conductivity: The capacity of water to allow electrons to flow through it, also known as its ability to conduct electricity, is measured by the water's conductivity. Conductivity standard levels below 800 are safe for human consumption, whereas conductivity standard levels between 800 and 2,500 are safe for all animals. Above 10,000, the level is unsafe for both human habitation and agricultural irrigation.

5. Sediments: The presence of sediment is what determines how clean the water is or how much clarity there is in the

979-8-3503-4602-2/22/$31.00

water. The removal of sediments is accomplished by a process called filtering.

6. Suspended Solids: Filtration is a process that may remove suspended solids from water, which includes things like plant matter, leaves, decomposed organic matter, and a variety of other things.

B. Related Works

In general, the discharge of wastewater into different water sources, such as rivers, creates an adverse and dangerous scenario; thus, there has been a drive to develop more targeted solutions for the treatment of sewage. Combining sedimentation with another process, "synthetic precipitation" is one of the most used methods for treating wastewater. The earliest experiments on the microbiology of slime assimilation are conducted in England in 1865 [2]. The earliest tests on the uneven filtering of wastewater were conducted in 1868. The next year, in 1870, the first sand filtration studies using discontinuous filtration were conducted in England. England was the site of the first research on air circulation in 1882. The United States of America was the first nation to adopt bar racks in 1884. The United States of America constructed its first drug precipitation treatment center in 1887. In Massachusetts' Lawrence Experiment Station in 1889, contact bed filtration was attempted for the first time. In 1891, Germany was the first nation to establish the technology of digesting sludge in tidal ponds. 1895: England collects methane gas from sewage tanks and uses it to light plants. This custom is observed in England. In 1898, the first rotary sprinklers for use with rotary filters were created. In 1904, the United States of America witnessed the introduction of the first grit chambers. The hostile character of sedimentation's effluent necessitated the use of septic tanks in which the sediments are made relatively harmless. However, a variety of problems necessitated the widespread adoption of the Travis two-story septic tank in England in 1904 and the Imhoff tank that was certified in Germany. 1904 saw the introduction of both of these tanks. In 1912 and 1913, at the Lawrence Experiment Station, the air was pumped through containers of wastewater containing slate. Arden and Lockett conducted studies in 1914 that culminated in the refining of the actuated slop process, which is how a high level of decontamination is obtained. Arden and Lockett's leadership prompted these examinations. In 1916, the procedure was used for the first time in a municipal facility in San Marcos, Texas, which was responsible for sewage treatment [2]. In 1925, Buswell started manufacturing contact aerators in the United States. In 1912 and 1913, slate-lined tanks were used to oxygenize wastewater at the Lawrence Experiment station. Arden and Lockett conduct research in 1914 that ultimately leads to the development of the activated sludge process, which enables a high degree of purification to be achieved. San Marces, Texas, used the technology for the first time in 1916 to filter sewage. Diverse advances occurring in the wastewater treatment sector today are largely attributable to the fact that the properties of wastewater are changing as a consequence of the release of a broad array of contaminants. One of these methods is the treatment of wastewater or sewage. Several strategies are created and tested with the ultimate goal of mimicking the natural treatment processes in order to lower the pollution load to a level that nature can handle. Consequently, it is crucial to pay particular attention to the surveying of the natural consequences of the wastewater treatment facilities that are currently in operation.

Researchers from the last decade also started working in this sector. Guo et al. [17] and Wang et al. [18] propose different machine-learning approaches to predict effluent prediction and its economical benefits of it. Moreover, in the last couple of years, researchers are focusing on advanced deep-learning approaches that depend less on using a high amount of computational power to solve problems which make the more accessible [19,20] to mass people. So, their models can be more available regardless of the processing power of a computer.

III. Research Methodology

This section covers the experiment used to find problems with wastewater treatment. Using a data-driven method that is applied to the online data of a plant throughout this process, we discover that the output is dependent on the quality as well as the amount of the data. Data-driven method comprises two sections which are the univariate approach and the multivariate technique. We use the multivariate methodology because, using the univariate method, we would have been unable to evaluate a huge dataset in which each variable is dependent on the others. On the other hand, the multivariate method is able to manage a vast number of variables that are interrelated. There are several different types of multivariate techniques, including factor analysis, cluster analysis, multidimensional scaling, and principal component analysis [3]. One method of learning from a particular dataset is known as supervised learning, whereas the other is known as unsupervised learning. In our study, we compile an unsupervised dataset. Therefore, we use cluster analysis in order to determine whether or not a wastewater treatment facility has excellent data or poor data. In order to prepare for cluster analysis, we first build a graph that illustrates a correlation between the parameters. When we look at such graphs, we see small flocculated points, which are deemed to be false data since they are produced as a result of the disruption of the machinery that is found in a WWTP. A method like clustering is an example of a multivariate approach that gives a main variable list that represents another variable of the process.

We experiment with a variety of methodologies including variable distribution and neural networks which includes Linear Regression, Logistic Regression, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decision Tree, and Random Forest.

A linear technique for modeling the connection between a scalar response and one or more explanatory factors, linear regression models the relationship in a linear fashion.

Logistic regression is a statistical model that predicts the likelihood of one occurrence out of two options by expressing

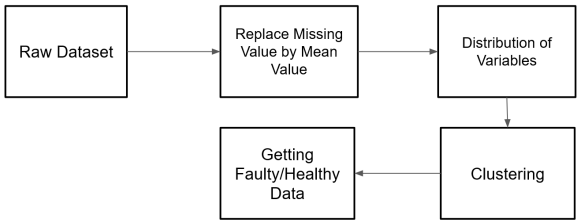

Fig. 1. Workflow

the log chances (the logarithm of the odds) as a linear combination of one or more independent variables. Logistic regression predicts the likelihood that one of two possible events will occur ("predictors"). The objective of the regression method known as logistic regression is to estimate the parameters of a logistic model.

Application of decision trees as a non-parametric supervised learning technique for classification and regression. The objective of this project is to construct a model capable of predicting the value of a target variable using basic decision rules drawn from the data's features. A tree may be seen as an illustration of a piecewise constant approximation.

Random forests, also known as random choice forests, are an ensemble learning strategy for classification, regression, and other issues. Random forests operate by producing a huge number of decision trees during the training phase of the procedure. Choice-based random forests are a subset of random forests. The output of the random forest when applied to classification issues is the category selected by the majority of trees. When doing regression tasks, the output is the mean or average of the various trees' predictions. Random decision forests address to counteract the propensity of decision trees to overfit to their training sets.

The Support Vector Machine, often known as SVM, is one of the most well-known Machine Learning algorithms that is based on Supervised Learning. It is used for Classification as well as Regression issues. The purpose of the Support Vector Machine (SVM) technique is to generate the optimal line or decision boundary that can divide an n-dimensional space into classes. This will allow us to simply place any new data points in the appropriate category in the future.

Multilayer Perceptron (MLP) is a Neural Network approach responsible for understanding the associations between linear and non-linear data sets. It is a kind of feed-forward neural network that may be separated into three different layers called the input layer, output layer, and hidden layer. The signal that will be processed is transferred to the input layer. It is the output layer's responsibility to do the appropriate tasks. The actual computational engine of the MLP consists of an arbitrary number of hidden layers located between the input and output layers. An MLP functions in a way comparable to a feed-forward network. Data moves forward from the input layer to the output layer.

IV. Experimental Evaluation

A. Dataset

The goal of this research is to examine the condition of the wastewater treatment plant using statistical analysis and machine learning. Manel Poah, Unitat d'Enginyeria Quinica, Universitat Autnoma de Barcelona, produce a dataset for the respiratory machine learning project [3] at UCI. This data collection includes information collected over a period of 527 days. This plant processes 3,500 m^3 of residential and industrial wastewater per day. This plant consists of three sections: Pre-treatment, Primary Treatment, and Secondary Treatment. This dataset is created using daily sensor readings from an urban sewage treatment plant. To forecast defects using the state variables of the plant at each step of the treatment process, we classify the condition of the operating plants. This domain contains 38 different characteristics and is poorly structured, according to the description provided.

B. Experimental Findings

In the process of our investigation, we look into the question of whether or not the water treatment facilities have any machinery that is inoperable. We are doing research to evaluate the influence on plant production at various stages of the plant. To be more specific, our primary objective is to investigate the dataset, monitor the various parameters (pH, dissolved oxygen, sediments, and solid particles) [4], and determine which parameter or collection of characteristics exhibits anomalous behavior. Our objective is to provide an accurate estimation of:

1) If there are any problems with the water treatment facility or if it is running correctly.
2) Identify which machines or portions of plants are experiencing problems or may be experiencing problems based on the dysfunctional behavior of certain parameters in the dataset and make a prediction.

1) pH Analysis: The distribution of values for the four variables PH-E [4], which represents the input pH to the plant, PH-P [4], which represents the input pH to the primary settler, and PH-D [4], which represents the input pH to the secondary settler, and PH-S [4], which represents the output pH, is shown in the Figure 2. Do any kinds of patterns emerge from their behavior? Are there any specific behaviors that are represented by the PH-E and PH-P charts? Each figure in this article illustrates the possible value distributions that we may encounter as well as the relationships between two variables (PH-E, PH-P). Due to the fact that it is a representation of a 4×4 matrix, this indicates that we are able to see 16 charts. This graph's distribution shows that, with very few exceptions, the majority of pH values lie between 7-8.

The distribution of PH-E, PH-P, PH-D, and PH-S is shown in Figure 2.

2) Conductivity analysis: COND-E (input conductivity to plant), COND-P (input conductivity to the main settler), COND-D (input conductivity to a secondary settler), and COND-S (output conductivity) are shown in Figure 3.

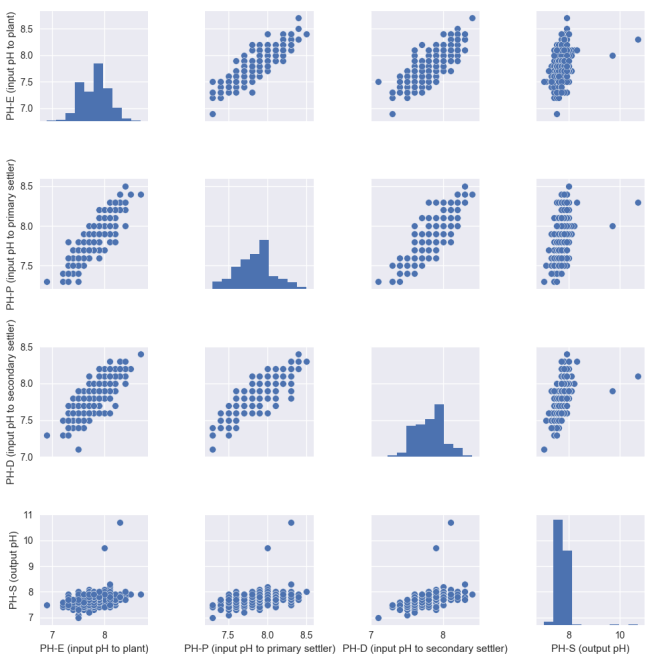

Fig. 2. PH Analysis

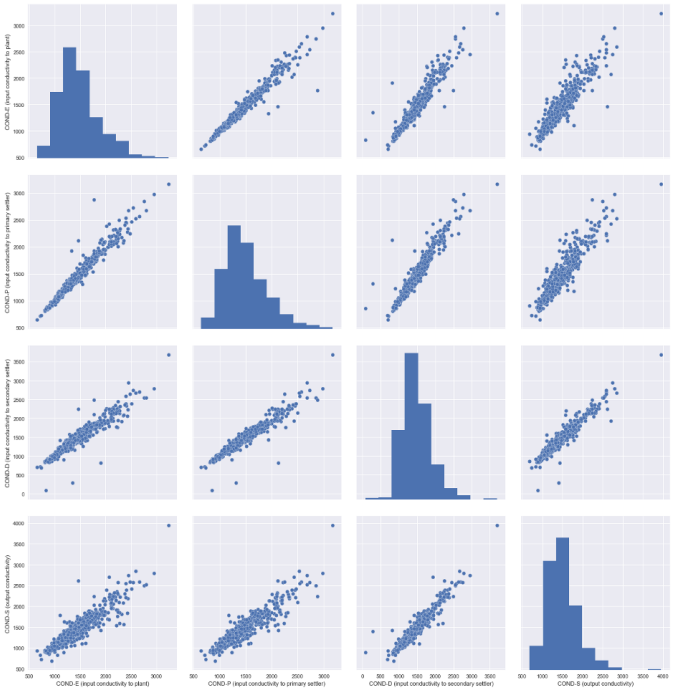

Fig. 3. Conductivity analysis

A scatter plot matrix is shown on the graph; this kind of plot illustrates how several variables are related to one another. The graph at the top depicts the distribution of the four variables COND-E, COND-P, COND-D, and COND-S. These variables are as follows: In this example, each chart displays two different variables. Due to the fact that this is an illustration of a 4×4 matrix, there are a total of 16 charts that may be seen here. The conductivity values are clearly visible in this graph distribution to be spread between the ranges of

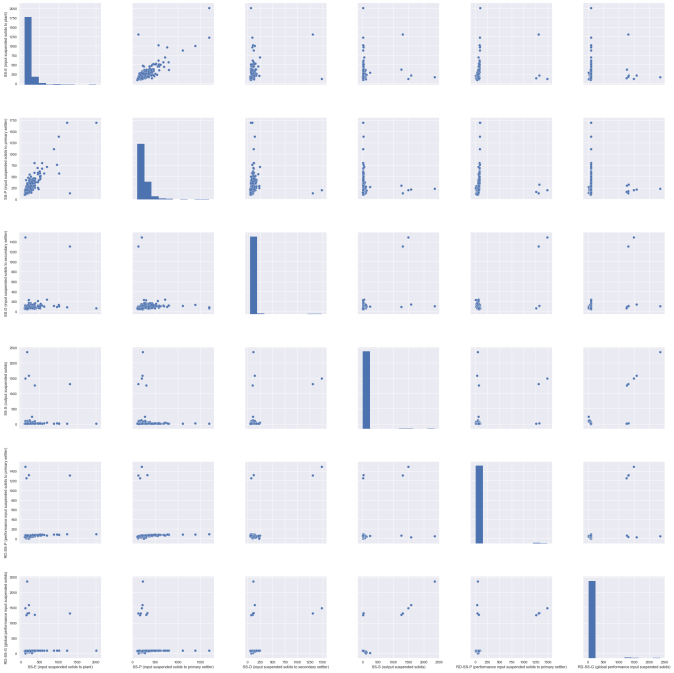

Fig. 4. Suspended Solid Analysis

700 to 3000, with the majority falling between 1000 and 2000. The range of conductivity values is clearly visible in this graph distribution.

3) Suspended Solid Analysis: Figure 4 SS-E (input suspended solids to plant), SS-P (input suspended solids to a primary settler), SS-D (input suspended solids to a secondary settler), SS-S (output suspended solids), and RD-SS-G distribution (global performance input suspended solids).

This graph's distribution shows that the values of the input suspended solids for the output are typically dispersed between 0 and 300, whilst the values of the input for the other side are mainly scattered between 100 and 500. We can observe this by comparing the two sides of the graph.

4) Biological oxygen demand analysis: Figure 5 represents the distribution of DBO-D, DBO-S (output Biological Demand of Oxygen), RD-DBO-G, DBO-E (input Biological Demand of Oxygen to Plant), DBO-P (input Biological Demand of Oxygen to Primary Settler) (global performance input Biological demand of oxygen).

A scatter plot matrix is shown in Figure 5, and it shows the variable relationships that exist between each other. In addition to this, it illustrates the distribution of the seven variables known as DBO-E, DBO-P, DBO-D, DBO-S, RD-DBO-P (performance input Biological demand of oxygen in primary settler), RD-DBO-S (performance input Biological demand of oxygen to a secondary settler), and RD-DBO-G. In this example, each chart displays two different variables. Because this is a depiction of a seven-by-seven matrix, we may see all 49 charts that are included inside it. The distribution of this graph makes it quite evident that the range of values for the input of the biological oxygen requirement is between 0 and 500 and between 1000 and 2000. Despite this, output and global performance are dependent on the input. Values for the

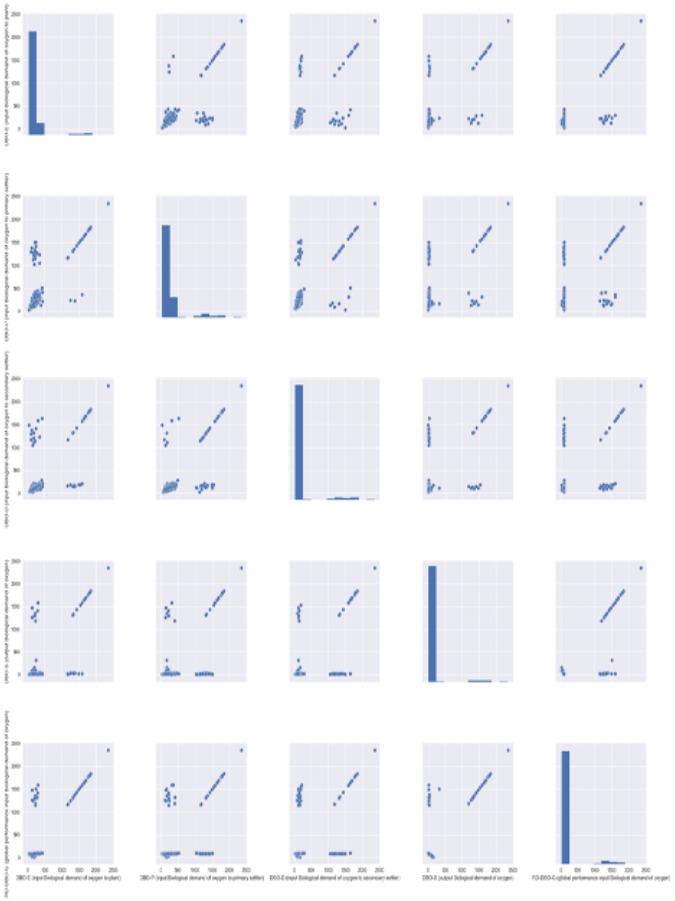

Fig. 5. Biological oxygen demand analysis

biological oxygen requirement may vary anywhere from 0 to 250 to between 1000 and 1800.

5) Chemical oxygen demand analysis: Figure 6 represents RD-DQO-G (output chemical demand of oxygen), DQO-S (output chemical demand of oxygen), DQO-E (input chemical demand of oxygen to plant), and DQO-D (input chemical demand of oxygen to a secondary settler) (global performance input chemical demand of oxygen).

The Figure that can be seen above is a scatter plot matrix, which illustrates how several variables are related to one another. The following image depicts the distribution of the five variables DQO-P (performance input chemical demand of oxygen to a secondary settler), DQO-D, DQO-S, RD-DQO-S, and RD-DQO-G. DQO-P refers to the input chemical demand of oxygen to the plant. DQO-D and DQO-S refer to the distribution of DQO-P. DQO-S refers to the distribution of DQO In this example, each chart displays two different variables. A depiction of a 5×5 matrix may be seen here in the form of twenty-five charts. The majority of the output and global performance input chemical demand of oxygen values are within the range of 0-250, whereas the majority of the input chemical demand of oxygen values is within the range of 100-700. The majority of the output chemical demand of oxygen values is also within this range.

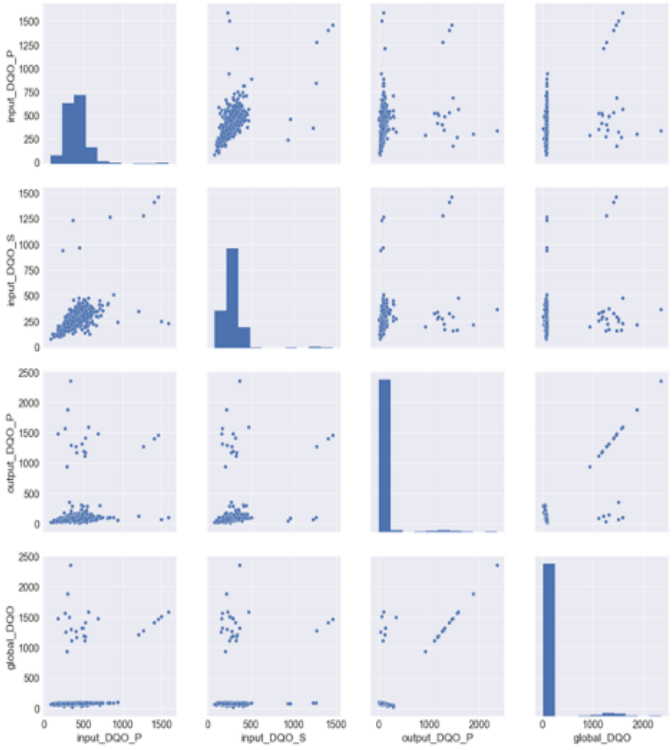

Fig. 6. Chemical oxygen demand analysis

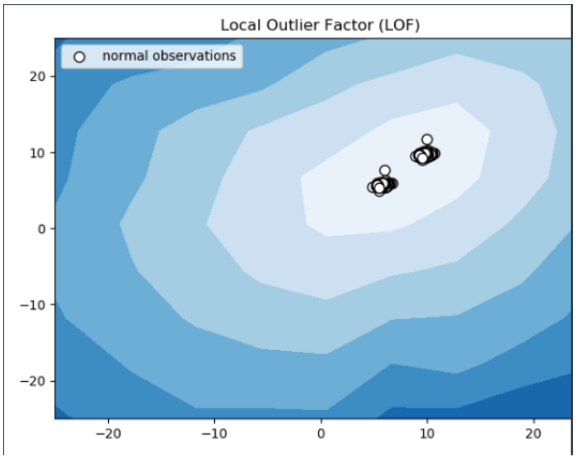

Fig. 7. Local Outlier Factor

C. Results

The effects of output pH density on a WWTP are shown for our study in Figure 7. In this case, a very dense region indicates correct data or a pH value that is close to the standard value. On the other hand, a value that is farther out from the dense area indicates incorrect data or a pH value that is either higher or lower than the standard value of pH.

The comparison of the algorithms that we test in Table I suggests that a variety of algorithms each provide the greatest accuracy result for a distinct parameter. The SVM method is the one that works best with pH. On the other hand, the best results may be obtained using the random forest method for BOD and COD. Within the context of suspended solids, all three algorithms— SVM, MLP, and random forest — show

Parameters	Linear Regression	Logistic Regresion	SVM	MLP	Decision Tree	Random Forest
pH	34.6	94.03	95.12	94.03	85.9	89.43
BOD	34.6	85.9	62.6	71.81	80.85	86.17
COD	45.42	85.63	56.09	75.88	84.01	86.72
Suspended Solids	21.51	98.91	99.18	99.18	99.45	99.18

TABLE I
ACCURACY TABLE (%)

the same level of accuracy.

V. DISCUSSION

When dealing with this dataset, there are certain missing variables that might have an effect on the outcome. These values could be important. We replace any null values with the mean or average value of the corresponding parameter. The dimensionality of the dataset is decreased as a result of the difficulty associated with classifying a huge dataset.

When we evaluate the data, we see that only a very tiny percentage of each value is slipping outside of this range in a dispersed way. This is something that we notice when we look at the data. The magnitude of the differences across all of the numbers may be seen rather well in this graph distribution. The most reasonable explanation that we can come up with is that these few findings are inaccurate data that were produced as a result of a machine failing to perform properly during peak business hours.

VI. CONCLUSION

Using a range of parameter sets, this study evaluates multiple ways of finding defective equipment inside a WWTP. Additionally, there are several subcategories for each of the other requirements. These factors include subcategories such as pH, DBO, DQO, SS Input PH, Primary Settler PH, Secondary Settler PH, and Output PH, among others. Following the execution of a variety of algorithms, we determined that each algorithm is capable of detecting a distinct parameter; however, no one parameter can detect all of the other parameters. According to our results, each individual set of parameters requires a new classifier. This study's major objective is to detect, by statistical analysis and machine learning, which components of a WWTP's equipment are malfunctioning. In addition, the examination of big datasets is incredibly challenging for humans, but clustering and machine-learning algorithms make this process quite easy. Nonetheless, this procedure is not devoid of disadvantages. Data play a significant part in this process. Therefore, accuracy may not be sufficient if there is an abnormality in the data or if there are missing values. We have opted to substitute any missing data with the average value for each parameter in order to resolve this problem. A substantial correlation coefficient between the input variable and the output variable demonstrates the veracity of this research's result. Future work includes examining more advanced algorithms that may be used to build a more precise model. In addition, future studies will concentrate on the construction of a single algorithm that will be the optimal match for each unique parameter.

ACKNOWLEDGMENT

We have our most sincere gratitude to Dr. Iftekharul Mobin for his supervision throughout the research work.

REFERENCES

[1] Benazzi, F., Gernaey, K. V., Jeppsson, U., Katebi, R., 2007. Online estimation and detection of abnormal substrate concentrations in WWTPS using a software sensor: A benchmark Study. Environmental Technology 28 (8), 871–882.

[2] Bernholt, T., Fried, R., Gather, U., Wegener, I., 2006. Modified repeated median filters. Statistics and Computing 16 (2), 177–192.

[3] Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

[4] Carstensen, J., Harremo¨es, P., Strube, R., 1996. Software sensors based on the grey-box Modelling approach. Water Science and Technology 33 (1), 117–126.

[5] Carstensen, J., Madsen, H., Poulsen, N. K., Nielsen, M. K., 1994. Identification of wastewater treatment processes for nutrient removal on a full-scale WWTP by statistical methods Water Research 28 (10), 2055–2066.

[6] Cecil, D., Kozlowska, M., 2010. Software sensors are a real alternative to true sensors. Environmental Modelling & Software 25 (5), 622–625.

[7] C´er´eghino, R., Park, Y.-S., 2009. Review of the Self-Organizing Map (SOM) approach in Water resources: Commentary. Environmental Modelling & Software 24 (8), 945–947.

[8] Ravi Kumar, P, Liza Britta Pinto, Somashekar, R.K. (November 2010), Assessment of the Efficiency Of Sewage Treatment Plants: A comparative study between Nagasandra and Mailasandra Sewage Treatment Plants, Kathmandu University Journal of Science, Engineering and Technology Vol. 6, No. II, pp 115-125.

[9] Victor Chipofya, Andrzej Kraslawski and YuryAvramenko (June 2010), Comparison of pollutant levels in effluent from wastewater treatment plants in Blantyre

[10] Malawi International Journal of Water Resources and Environmental Engineering Vol. 2(4), pp. 79-86

[11] Nobel Francisco Rovirosa Morell 1997, Performance Evaluation of an on-site domestic sewage treatment plant for individual residences.

[12] Majed M. Ghannam 2006, Performance Evaluation of Gaza Waste water Treatment Plant, Islamic University-Gaza

[13] E. Awuah & K. A. Abrokwa 2008, Performance evaluation of the USAB sewage treatment plant at James Town (Mudor), Accra, 33rd WEDC International Conference, Accra, Ghana.

[14] Sushil Kumar Shah Teli, (December 2008), Performance Evaluation of Central Wastewater Treatment Plant: a Case Study of Hetauda Industrial District, Nepal, 36/Environment and Natural Resources Journal Vol.6, No.2.

[15] Al-Zboon, Kamel and Al-Ananzeh, Nada (August 2008), Performance of wastewater treatment plants in Jordan and suitability for reuse, African Journal of Biotechnology Vol. 7 (15), pp. 2621-2629.

[16] Yahaya Mijinyawa and Nurudeen Samuel Lawal (June 2008), Treatment efficiency and economic benefit of Zartech poultry slaughter house waste water treatment plant, Ibadan, Nigeria, Scientific Research and Essay Vol. 3 (6), pp. 219-223.

[17] H. Guo et al., "Prediction of effluent concentration in a wastewater treatment plant using machine learning models," J. Environ. Sci. (China), vol. 32, pp. 90–101, 2015, doi: 10.1016/j.jes.2015.01.007.

[18] D. Wang et al., "A machine learning framework to improve effluent quality control in wastewater treatment plants," Sci. Total Environ., vol. 784, no. 147138, p. 147138, 2021, doi: 10.1016/j.scitotenv.2021.147138.

[19] J. J. Mondal, M. F. Islam, S. Zabeen, A. B. M. A. A. Islam, and J. Noor, "Note: Plant leaf disease network (PLeaD-net): Identifying plant leaf diseases through leveraging limited-resource deep convolutional neural network," 2022, doi: 10.1145/3530190.3534844.

[20] B. Mamandipoor, M. Majd, S. Sheikhalishahi, C. Modena, and V. Osmani, "Monitoring and detecting faults in wastewater treatment plants using deep learning," Environ. Monit. Assess., vol. 192, no. 2, p. 148, 2020, doi: 10.1007/s10661-020-8064-1.