Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 7.056

IJCSMC, Vol. 12, Issue. 4, April 2023, pg.52 – 59

Water Quality Prediction Using Machine Learning

Annaji Kuthe[1]; Chaitanya Bhake[2]; Vaibhav Bhoyar[3]; Aman Yenurkar[4]; Vedant Khandekar[5]; Ketan Gawale[6]

[1, 2, 3, 4, 5, 6]Dept.of Computer Science & Engineering, KDKCE Nagpur, India

[1] a.kuthe@gmail.com; [2] chaithubhake2013@gmail.com; [3] bhoyarvaibhav15@gmail.com;
[4] amanyenurkar@gmail.com; [5] vedantkhandekar.cs@kdkce.edu.in; [6] ketangawale.cs@kdkce.edu.in

DOI: https://doi.org/10.47760/ijcsmc.2023.v12i04.006

Abstract: Different toxins have been imperiling water quality over the past decades. As a result, foreseeing and modeling water quality have gotten to be basic to minimizing water contamination. This inquiry has created a classification calculation to foreseethe water quality classification (WQC). The WQC is classified based on the water quality file (WQI) from 7 parameters in a dataset utilizing Back Vector Machine (SVM) and Extraordinary Gradient Boosting (XGBoost). The comes about from the proposed model can precisely classify the water quality based on their features. The inquire about result illustrated that the XGBoost model performed way better, with an exactness of 94%, compared to the SVM demonstrate, with as it were a 67% exactness. Indeed way better, the XGBoost brought about in as it were 6% misclassification mistake compared to SVM, which had 33%. On best of that, XGBoost too gotten consistent predominant comes about from 5-fold approval with an normal accuracy of 90%, whereas SVM with an normal exactness of 64%. Considering the upgraded execution, XGBoost is concluded to be superior at water quality classification.

Keywords: SVM, XGBoost, water quality, machine learning, classification.

I. INTRODUCTION

Water is the foremost important resource for life, because it is required for the survival of living beings and people. From The Joinedtogether Countries world water advancement report 2017, approximately 80% of the world's wastewater is released back into the environment, generally untreated, harming waterways, lakes, and oceans [1]. The far reaching issue of water contamination is imperiling our wellbeing. In case nothing is done, the issues will only decline by 2050, when worldwide freshwater request is estimated to be one-third more than it is right now . [2]. Water quality checking through classification has become an vital strategy to control the contamination level. A manual calculation is utilized to decide the water lesson when classifying water quality. The manual examination is wasteful since it takes a long time to total the operation. As a result, an progressed framework to classify water quality is essential [3]. An exertion to guarantee water quality is the improvement of a water quality observing framework. By utilizing parameters such as the concentration of dissolved oxygen, microbes levels, pH esteem, conductivity, Biochemical Oxygen Request (BOD), Chemical Oxygen Demand (COD), fecal coliform, and add up to coliform, the system can calculate and classify the contamination of the water to a certain degree [4]. The method includes utilizing an calculation to calculate the information collected from waterways or lakes.

Support vector machines (SVM) could be a part approach used to unravel classification and relapse issues [5]. Hyperplanes are utilized by SVMs, which were initially developed for twofold classification, to build choice boundaries between information points of distinctive classes. The SVM algorithm is sweet at generalizing the test, managing with nonlinear issues in a basic but exceedingly precise way, and differentiating information that contrasts from the standard; hence, it's ideal for classifying and evaluating water quality information [4][6]. Extreme angle boosting or XGBoost may be a slope boosting system from a decision-tree-based gathering Machine Learning calculation. XGBoost can be utilized to illuminate problems including relapse, classification, positioning, and user-defined expectation. In different machine learning and data mining challenges, the affect of XGBoost has been widely recognized, even for water classification [7]. Water quality should be studied comprehensively because of its significance in way oflife and its impacts on human health. Water quality observing offers the objective information needed to form educated choices around water quality management presently and within the future [8]. As a result, water quality must be checked to guarantee that no contaminants surpass levels that are perilous to human wellbeing. Water quality is as of now determined through expensive and time-consuming lab and statistical investigations. It requires test collection, transportation to labs, and a noteworthy sum of time and calculation, which is ineffectual given that water could be a profoundly communicable medium and perilous in the event that water sullied with disease-causing squander isn't avoided prior [9]. In this case, an elective strategy based on machine learning for the efficient expectation and classification of water quality levels is implemented. We propose the execution of SVM and XGBoost as an alternative method for water quality classification.

II. LITERATURE REVIEW

In recent research by Bouamar and Ladjal [10], SVM was compared with an artificial neural network (ANN) for water quality classification. The research was conducted with two classified conditions, whether the water was potable or not. SVM performed better with an error of 4% compared to ANN model with 7% error due to SVM characteristic they are more resistant to noise [10]. Further research from China, which used SVM for water quality monitoring system found that the classification result of SVM is better with an accuracy of 99.2% compared to Multi layered perceptron (MLP) Logistic Regression and Radial Basis Function (RBF) [6].In [3], SVM water quality classification competed with the K Nearest Neighbor (KNN) method, a method for object classification. Based on learning data closest to the object. The study found that SVM performed better with an accuracy of 92.40%.

In comparison, KNN achieved only 71.28% accuracy when using 10-cross-validation on 120 datasets. SVM scored the highest accuracy from the linear kernel and the highest accuracy of KNN is when K=7 [3].As mentioned in the Introduction section, next the classification method considered in this research is XGBoost. This was discovered by a study conducted by scientists from Turkey that XGBoost performed better in water quality classification for accuracy, precision and recall with 95%, 96% and 96%, respectively against Logit Boost, RF and AdaBoost [11].

Further studies on water quality classification using tree classification file model, XGBoost, is on par with Light GBM classification method with a classification accuracy of 85%. Compared to other tree-based learning models such as Classification Tree, Random Forest and Cat Boost [12]. In [13], Review paper on design and optimization of energy efficient wireless sensor network model for complex network [15]. XGBoost performed moderately with 88% accuracy in comparison on XGBoost, which had the highest accuracy of 94%, but still performed better than SVM, with only 80% accuracy. A.M Kuthe al. Prevention of Suicide Risk and Predicting Suicidal Behaviors by machine learning [14]. From the peer-reviewed literature, they have SVM and XGBoost was used for water quality classification; however their the implementation is non-trivial. Issues such as limited limitation water quality parameters and robustness of the algorithm dealing with noise are still actively investigated. So water quality classification using SVM and XGBoost more attention should be paid.

III. METHODOLOGY

This section discusses SVM classification algorithms and XGBoost, whose performances are evaluated in water quality classification. The methodology includes the project workflow, method description and evaluation performance criteria shown in Figure 1.

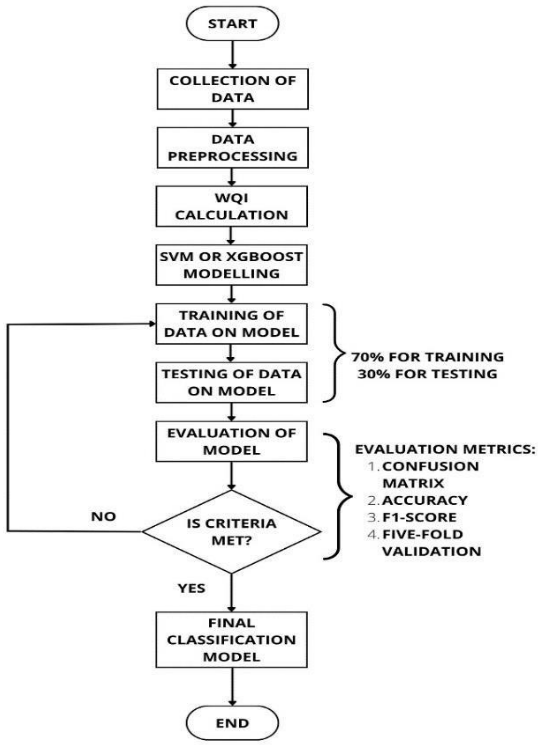

Figure no. 1 Flowchart of methodology

A. Dataset

The data set used for this research was obtained from The Kaggle website, originating from the Government of India website [18].Historical water quality data for several locations in India was combined and cleaned.

Data corresponds to current research project from characteristics required to construction the water quality index available in this dataset. Water the quality classification can be derived from the water quality index. A classification model can use a large amountdata, up to 2000 samples, for classification training and forecast.

B. Data Preprocessing

Before using the data for training, the data must be passed through pre-processing of data that concerns identification and correction errors in the data set that can negatively affect the prediction model [4]. Common errors in a dataset include missing columns and duplicate rows. The data preprocessing stage is critical data analysis to improve data quality. Using raw data from dataset for classifications may generate incorrect results; therefore, data cleaning is critical [4]. The water quality index (WQI) was calculated using the most important parameters of the data set which are dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrates, fecal coliforms and total coliforms. The WQI score was then used to categorize the water samples. This the research uses a weighted arithmetic water quality indexic calculation method. WQI is chosen because incorporates various water quality criteria into mathematics an equation that measures the health of a water body and describes acceptability of surface and undergroundwater resources for humans use [19].

C. Water quality index

WQI is one of the most effective tools for expression water quality because it provides a simple, consistent unit measurements to describe water quality. This gives an important aspect in surface water assessment and management due to its sensitivity to the presence of contaminants that affect water quality [20], [21], [22]. The WQI is a rating system that calculates the combined impact of individual water quality parameters for overall water quality. Seven important ones parameters were chosen for the water calculation study quality index. A weighted arithmetic method was chosen calculate WQI. The advantage of this approach is that itis slightly better than other methods such as Canadian methods, WQI calculation [20]. World Health Organization (WHO)Drinking water quality standards were used for the determination WQI [23]. There are several steps that must be taken before obtaining a WQI completed using the arithmetic index method in the following text steps as suggested in [24].

Calculation of sub-index of quality rating (Q_n)

Let (n) be the number of water quality parameters and q_n be the quality rating or sub-index corresponding to the n^{th} parameter. Eq.1 is used to compute the value of q_n.

$$Q_n = 100 \left[\frac{v_{n} - v_{i0}}{s_n - v_{i0}}\right]$$

Such that: q_n = rating quality for the nth water quality parameter; v_n = estimated value of the nth parameter at a given determinant; v_{i0} = ideal value of the n^{th} parameter in pure water; s_n = standard permissible value of the n^{th} parameter. For all other parameters, except for pH, which is 7.0, and dissolved oxygen = 14 mg/L, all ideal (v_{i0}) values are taken as zero (0) for drinking water [23].

Unit weight(W_n)

For various water quality parameters, the calculation of unit weight (W_n) is inversely related to the suggested standards S_n for the relevant parameters in Table 2, as shown in Eq. 2

$$w_n = k/S_n$$

Such that: W_n = unit weight for the n^{th} parameter; S_n = standard permissible value of the n^{th} parameter; K = constant forproportionality.

WQI the formula for WQI can be seen in Eq. 3:

$$WQI = \sum q_{nW_n} / \sum W_n$$

The overall water quality index was produced by linearly aggregating the standards of quality.

WQI range	Water quality	Class value
0 - 25	Excellent	3
26 - 50	Good	2
51 - 75	Poor	1
Greater than 75	Very Poor	0

Table no. 1 Range of class for water quality classification

No.	Parameter	Standard value recommended	Ideal value	Relative weight
1	Dissolved oxygen (DO) mg/L	10	14	0.2213
2	pH	8.5	7.0	0.2604
3	Clorophyll	250	0	0.0022
4	Turbidity	5	0	0.4426
5	Total Suspended Value (TSS)	45	10	0.0492
6	Fecal coliform (FC)	100	0	0.0221
7	Total coliform (TC)	1000	0	0.0022

Table no.2 Water quality parameter and its standard, ideal and relative unit value

D. Machine learning

SVMs use hyperplanes to define decision boundaries between data points of different classes originally developed for binary classification problems [4]. Hyperplanes are decision functions that distinguish between positive and negative data and have marked maximum margins, shown in Figure 2. Because of its low sensitivity to properties spatial dimensions, SVM is considered a reliable classifier for Hughes effect; classification using SVM has a min impact on the result [25].

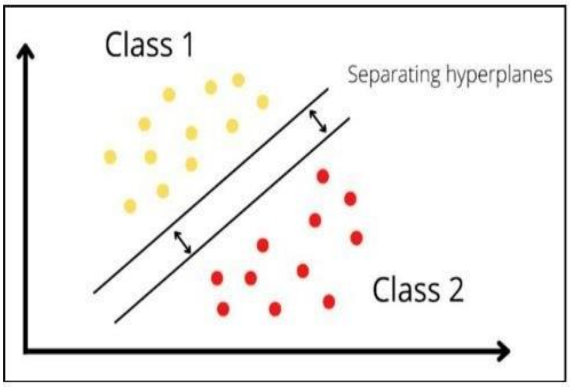

Figure no.2 Hyperplane illustration for SVM

E. Extreme gradient boosting (XGBoost)

The result of water quality classification between SVM and XGboost is covered in this section. A total of 1992 data were split into two parts, one for training the model and the other for testing. A split of 70% was used for the training and 30% for the testing validation. XGBoost is an algorithm that generates multiple shallow decision trees and combines all of their high accuracy values [26]. The decision tree created by the XGBoost algorithm is used to reduce the target by considering the loss of function, but also to prevent the tree from overfitting by regularly using the layer standard [11]. Datascientists prefer XGBoost because of the fast output of successful students [13].

$$F1\ Score = 2 * \frac{precision * recall}{precision + recall}$$

IV. RESULT

The result of water quality classification between SVM and XGboost is covered in this section. A total of 1992 row data were split into two parts, one for training the model and the other for testing. A split of 70% was used for the training and 30% for the testing validation

Getting live water parameter in the form of chart along with date similarly we can find varies parameter like turbidity chlorophyll,TSS (Total suspended solids). Below figure shows turbidity of water.

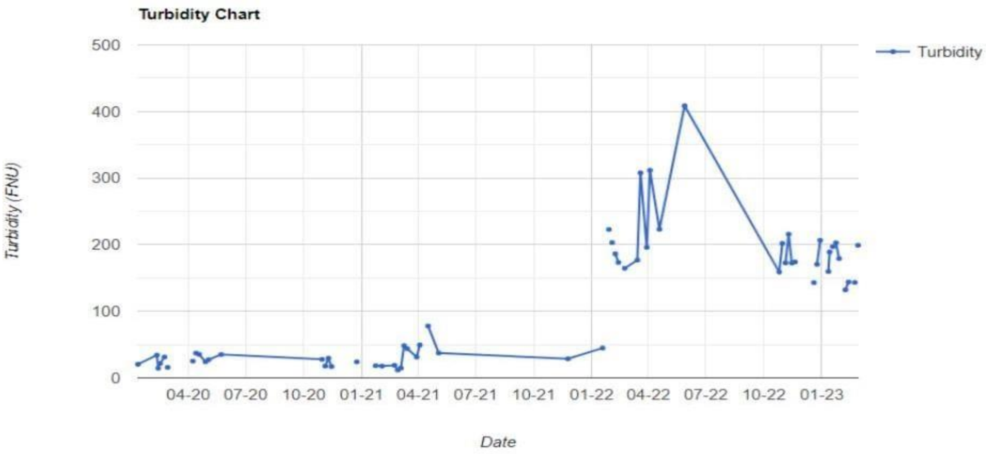

Figure no. 3 . Turbidity Chart

By using XGboost and SVM algorithm we can generate recall value, F1 and precision helps to analysis the purity of water in termsof excellent, good, poor, very poor according to contents of water.

Performance_Analysis

recall,F1 and Precision

	Recall	f1	Precision
Excellent	1.00	1.00	1.00
Good	0.98	0.96	0.94
Poor	0.90	0.89	0.88
Very Poor	0.88	0.92	0.95

Figure no. 4. Performance Analysis

In these module we insert values of water parameter such as DO (Dissolve Oxygen), pH value, Chlorophyll, Turbidity ,TSS(Total Suspended Solids) ,Fecal Coliform and Total Coliform to get the result.

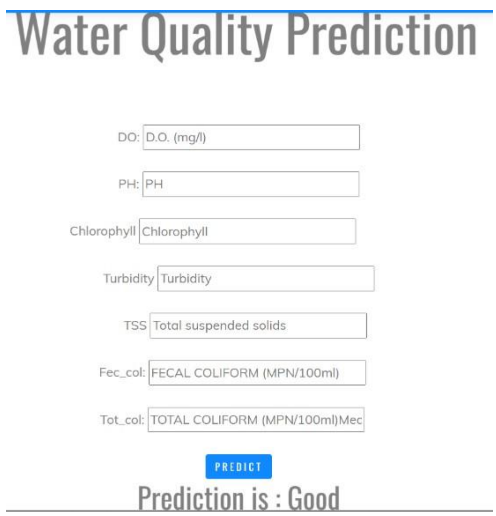

Figure no 5. Final Output

V. CONCLUSION

Water quality is vital in determining whether the water source is qualified for utilization. WQI is fundamental to classify whether the water is secure for utilization. Instead of requiring costly and complex examination to test the water quality, this inquire about employments two machine learning calculations, SVM and XGBoost, to anticipate water quality utilizing promptly available water quality parameters. The parameters utilized for the classification calculation are broken up oxygen, pH, conductivity, organic oxygen request, nitrate, fecal coliform, and add up to coliform. The result appeared that XGBoost outperformed the SVM calculation indeed after the parameters had been tuned. The comes about appeared that SVM brought about in more misclassification of information than XGBoost. In spite of the achievement of this inquire about, enhancements, such as applying more parameters to the demonstrate or utilizing indeed more progressed profound learning calculations, can be connected in future investigate to improve water qualityclassification.

REFERENCES

[1]. World Water Assessment Programme (United Nations), Wastewater : the untapped resource : the United Nations world water development report 2017.

[2]. P. Burek et al., "The Water Futures and Solutions Initiative of IIASA," 2016.

[3]. A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and support vector machine algorithm in classification water quality status," in Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016, Feb. 2017, pp. 137–141. DOI: 10.1109/FIT.2016.7857553.

[4]. K. P. Singh, N. Basant, and S. Gupta, "Support vector machines in water quality management," Analytica Chimica Acta, vol. 703, no. 2, pp. 152–162, Oct. 2011, DOI: 10.1016/j.aca.2011.07.027.

[5]. T. Eitrich and B. Lang, "Efficient optimization of support vector machine learning parameters for unbalanced datasets," Journal of Computational and Applied Mathematics, vol. 196, no. 2, pp. 425–436, Nov. 2006, DOI: 10.1016/j.cam.2005.09.009.

[6]. Z. Pang and K. Jia, "Designing and accomplishing a multiple water quality monitoring system based on SVM," in Proceedings 2013 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2013, 2013, pp.121–124. DOI: 10.1109/IIHMSP.2013.39.

[7]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, vol. 13-17-August-2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[8]. D. N. Myers, "Why monitor water quality?" [Online]. Available: https://www.epa.gov/assessing

[9]. "Artificial Neural Network Modeling of the Water Quality Index Using Land Use Areas as Predictors".

[10]. M. Bouamar and M. Ladjal, "Evaluation of the performances of ANN and SVM techniques used in water quality classification."

[11]. F. Hassanbaki Garabaghi, "Performance Evaluation of Machine Learning Models with Ensemble Learning approach inClassification of Water Quality Indices Based on Different Subset of Features," 2021, DOI: 10.21203/rs.3.rs876980/v1.

[12]. L. Li et al., "Interpretable tree-based ensemble model for predicting beach water quality," Water Research, vol. 211, Mar. 2022,DOI: 10.1016/j.watres.2022.118078.

[13]. N. Nasir et al., "Water quality classification using machine learning algorithms," Journal of Water Process Engineering, vol.48, p. 102920, Aug. 2022, DOI: 10.1016/j.jwpe.2022.102920.

[14]. A.M,Kuthe et al. Prevention of Sucicide Risk and Predicting Suicidal Behaviors by machine learning. Wutan Hutan JisuanJishu 2021, Vol XVII, Issue-I, pp. 563-567.

[15]. A. Kuthe and A.K. Sharma, "Review paper on Desingn and Optimization of Energy Efficient Wireless Sensor Network Model for Complex Network," 2021 5th International Conference on Information System and Computer Newtork (ISCON), 2021 , pp. 1- 3, DOI: 10.1109/ISCON52037.2021.9702421

[16]. D. Dezfooli, S. M. Hosseini-Moghari, K. Ebrahimi, and S. Araghinejad, "Classification of water quality status based on minimum quality parameters: application of machine learning techniques," Modeling Earth Systems and Environment, vol. 4, no. 1, pp. 311–324, Apr. 2018, DOI: 10.1007/s40808-017- 0406-9.

[17]. T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms," Applied Bionics and Biomechanics, vol. 2020, 2020, DOI: 10.1155/2020/6659314.

[18]. "Indian water quality data | Kaggle." https://www.kaggle.com/datasets/anbarivan/indian-water-quality-data (accessed Jul. 01,2022).

[19]. S. C. Dendukuri, D. S. Chandra, M. V. S. Raju, and S. S. Asadi, "Estimation of Water Quality Index By Weighted ArithmeticWater Quality Index Method: A Model Study," International Journal of Civil

Engineering and Technology, vol. 8, no. 4, pp. 1215– 1222, 2017, [Online]. Available: http://www.iaeme.com/IJCIET/index.asp1215
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=4
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=4
http://www.iaeme.com/IJCIE T/index.asp1216

[20]. F. M. Kizar, "A comparison between weighted arithmetic and Canadian methods for a drinking water quality index at selected locations in shatt alkufa," in IOP Conference Series: Materials Science and Engineering, Nov. 2018, vol. 433, no. 1. DOI: 10.1088/1757-899X/433/1/012026.

[21]. M. D. Noori, "Comparative analysis of weighted arithmetic and CCME Water Quality Index estimation methods, accuracy and representation," in IOP Conference Series: Materials Science and Engineering, Mar. 2020, vol. 737, no. 1. DOI: 10.1088/1757- 899X/737/1/012174.

[22]. C. K. Ojukwu, G. O. C. Okeah, and P. C. Mmom, "A Comparative Analysis of the Weighted Arithmetic and Canadian Council of Ministers of the Environment Water Quality Indices for Water Sources in Ohaozara, Ebonyi State, Nigeria," International Journal of Engineering Research & Technology (IJERT) www.ijert.org, vol. 10, 2021, [Online]. Available:

[23]. "Fourth edition incorporating the first and second addenda Guidelines for drinking-water quality."

[24]. S. Bouslah, L. Djemili, and L. Houichi, "Water quality index assessment of Koudiat Medouar Reservoir, northeast Algeria using weighted arithmetic index method," Journal of Water and Land Development, vol. 35, no. 1, pp. 221–228, Dec. 2017, DOI: 10.1515/jwld-2017-0087.

[25]. S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, and X. Guo, "Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel2 Time Series Data," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 9, pp. 3295–3306, Sep. 2019, DOI: 10.1109/JSTARS.2019.2922469.

[26]. S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, "Prediction of groundwater quality using efficient machine learning technique," Chemosphere, vol. 276, Aug. 2021, DOI: 10.1016/j.chemosphere.2021.130265.

[27]. M. I. Shah, M. F. Javed, and T. Abunama, "Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques," Environmental Science and Pollution Research, vol. 28, no. 11, pp. 13202–13220, Mar. 2021, DOI: 10.1007/s11356-020- 11490-9.

[28]. M. Y. Cho and T. T. Hoang, "Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems," Computational Intelligence and Neuroscience, vol. 2017, 2017, DOI: 10.1155/2017/4135465.

