# Deep Learning for Water Quality Classification in Water Distribution Networks

Essa Q. Shahra[1], Wenyan Wu[1], Shadi Basurra[1], and Stamatia Rizou[2]

[1] Faculty of Computing, Engineering and Built Environment, Birmingham City University, Birmingham B4 7XG, United Kingdom;
Essa.Shahra@bcu.ac.uk, Wenyan.Wu@bcu.ac.uk, Shadi.Basurra@bcu.ac.uk
[2] Singular Logic S.A., Athens, Greece; Srizou@singularlogic.eu

**Abstract.** Maintaining high water quality is the main goal for water management planning and iterative evaluation of operating policies. For effective water monitoring, it is crucial to test a vast number of drinking water samples that is time-consuming and labour-intensive. The primary objective of this study is to determine, with high accuracy, the quality of drinking water samples by machine learning classification models while keeping computation time low. This paper aims to investigate and evaluate the performance of two supervised classification algorithms, including artificial neural network (ANN) and support vector machine (SVM) for multiclass water classification. The evaluation uses the confusion matrix that includes all metrics ratios, such as true positive, true negative, false positive, and false negative. Moreover, the overall accuracy and f1-score of the models are evaluated. The results demonstrate that ANN outperformed the SVM with an overall accuracy of 94% in comparison to SVM, which shows an overall accuracy of 89%.

**Keywords:** Water Distribution System · Water Quality · Classification · SVM · ANN.

## 1 Introduction

Providing clean drinking water is a critical challenge for water supply companies worldwide due to events that are hard to predict or control, such as physical disruption, biological contamination, and chemical contamination [1]. Water distribution systems (WDS) are vital to cities' water supply, particularly given that their safety directly affects public health. However, the water quality is hard to maintain in WDS due to the effects of the distance and time taken during the long process of delivery from the water source to the end-users [2]. Besides, typically major parts of WDS exist in the open environment; thus, it is vulnerable to external disturbances, such as sabotage. For these reasons, researchers in industry and academia have been investigating the development of real-time contamination warning systems [3]. The two fundamental parts in developing these systems are sensor placement methods in the WDS, and the analysis of the big volume of data generated from these sensors [4] [5]. Typically, sensors generate

large amounts of data streams, which need to be analyzed in real-time to detect abnormal events that cause significant water pollution in WDS [6]. A large number of machine learning and statistical models for classification have been proposed, such as regularized discriminant analysis (RDA) [7], linear discriminant analysis (LDA) [8], quadratic discriminant analysis (QDA) [9], support vector machines (SVM) [10], neural networks (NN) [11], and k-nearest neighbour classifier (KNN) [12]. Nevertheless, the critical question remains: what is the effectiveness of these methods when applied in detecting contamination in the water supply network in terms of accuracy and performance? This paper aims to investigate and evaluate the performance of two supervised classification algorithms, including support vector machine (SVM) and artificial neural network (ANN) for water multi-classification. The evaluation uses the confusion matrix that includes all metrics ratio false positive, false negative, true positive, and true negative. Moreover, the overall accuracy, f1-score of the models are evaluated. The paper is organized as follows: section 2 reviews the most recent selective related work, section 3 describes the dataset used in this study and the contamination scenarios. Section 4 explains the methodology of the proposed work in more detail. Section 5 presents how SVM and ANN models are applied for water evaluation; section 6 shows the results, analysis, and performance of the models. Finally, section 7 concludes the work.

## 2 Related Work

A few studies have focused on water quality measurements for contamination detection events. For example, Chang et al. [13] proposed a water quality evaluation method based on clustering analysis with a self-organizing map ( SOM) and Fuzzy C-Mean. The results from this SOM classification showed higher efficiency in comparison to the traditional clustering method. Hadi et al. [14] developed an adaptive neuro-fuzzy model that classified the condition of drinking water into two classes: safe and unsafe conditions. The authors used a time-series for the real-time dataset, which contains four different water quality parameters: turbidity, color, PH, and bacteria count. The results showed the ability of the proposed model to detect contamination data with an accuracy of 92%. Oliker and Ostfeld [15] generated a model using a support vector machine to improve the ability of the proposed model to detect contamination and multi-dimensional data analysis. The model was trained and evaluated with the randomly simulated events superimposed on actual water distribution system data. The proposed model was highly effective as it managed to draw conclusions based on a few measured water quality parameters. Moreover, Arad et al. [16] developed a framework that consists of online and offline Phases. A genetic algorithm (GA) is used during the offline stage to change five decision variables: window size, positive and negative dynamic thresholds, and positive and negative filters. During the online process, a regular rule of Bayes is invoked for online event detection in real-time. The proposed model outperformed benchmark models, and the results showed that the detection capacity significantly improved. Yu et al. [2] proposed a method for

detecting water pollution using multi-station spatial and temporal data. WDS consists of multiple stations with large-scale characteristics and high complexity. The proposed method is evaluated over two water networks, and its detection efficiency is measured using the time analysis. The results showed high accuracy when using a massive amount of data from spatial and temporal dimensions. Despite the above-mentioned work, there is still a research gap in the prediction of water contamination with high accuracy. Therefore, this work aims to address this by applying deep learning for water quality multiclass classification.

### 2.1 Dataset

The dataset used in this study is provided by the US Environment Protection Agency (EPA) [17]. This data was collected from three stations named B, C, and F. These stations were selected to provide three distinctly different sets of water quality data that can be used for training and testing. Time series were collected for three months at two minutes interval. The data from each station includes chlorine (CL), conductivity (COND), PH, total organic carbon (TOC), temperature (TEMP), and pressure (PRES). In this study, we use the data collected only from station B, and a size of 7201 time-steps.

### 2.2 Contamination Scenario

Since most real water contamination incidents are rare in WDS, we used the known simulation tool EPANET to simulate water contamination [18]. Contamination modeled while ensuring the water conditions and network structure are identical to the WDS condition from which the dataset is generated. Since chlorine is one of the most used water treatment disinfectants, and it is used as the standard index for evaluation in this study. Chlorine is added as an indicator of the selected station to simulate water quality events. The contamination can be described by EPANET using a set of properties, for example, event time, duration, concentration, and the station. Based on the Surface Water Treatment Rule (SWTR) rules, the minimum and maximum residual rates can be set for the consumer's tap. Under the SWTR, water entering the distribution system cannot fall less than a 0.2 mg/L disinfectant residual benchmark for more than four hours without failure to comply. Low disinfectant residuals in the distributed system would increase opportunities for biological regeneration and lead to complaints about taste and odor from consumers. On the other hand, high disinfectant concentrations can lead to potentially harmful disinfection byproducts (DBPs). Hence, the maximum residual disinfectant concentrations cannot exceed 4.0 mg/L in chlorine [19].

## 3 Methodology

Based on the studies reviewed in the literature section, two supervised algorithms have been selected based on their performance while focusing on multiclassification accuracy and computational performance. These algorithms are

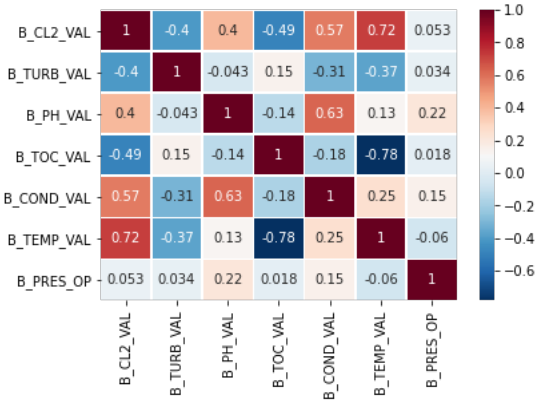


**Fig. 1.** Correlation coefficient among the water quality parameters (features)

ANN and SVM. These two algorithms used in this study to classify drinking water quality. In order to be able to achieve this, we need to explore the dataset and analyse the relationship between the water quality parameters/features. Then data labeling is done based on what was explained in sub-section 2.2. Finally, the selective models are implemented and applied for performance evaluation and analysis. More details of these steps are described in the below sub-sections: (3.1 and 3.2).

### 3.1 Data Exploration

**Pearson Correlation Coefficient (PCC)** PCC helps in determining the relationship between two parameters/features in the dataset. It is a metric for determining the degree to which two variables are related [20]. PCC has a range of values from -1 to 1. Fig. 1 represents the correlation between the seven water quality parameters used on our dataset. Generally, Authors notice a strong correlation among all parameters in the positive and negative directions. Based on the filtering method for feature selection, all features with medium and high correlation will be used for building the model i.e., during the learning phase. The pressure feature has been removed during the learning phase as it depicts a low correlation among all other features.

**Boxplot Distribution** Box Plot is an excellent method for visualizing how data features are distributed [21]. In our case, the dataset is represented using this distribution, and the distribution of the seven water quality parameters including in our dataset are shown in Fig. 2. From Fig. 2, authors notice the lowest mean relative values appeared significantly in both chlorine and turbidity, while the highest mean relative values appeared in the conductivity. We also see high variability in the data in general, but this variability is almost small in PH and turbidity, while the variability appeared more clearly in the pressure.

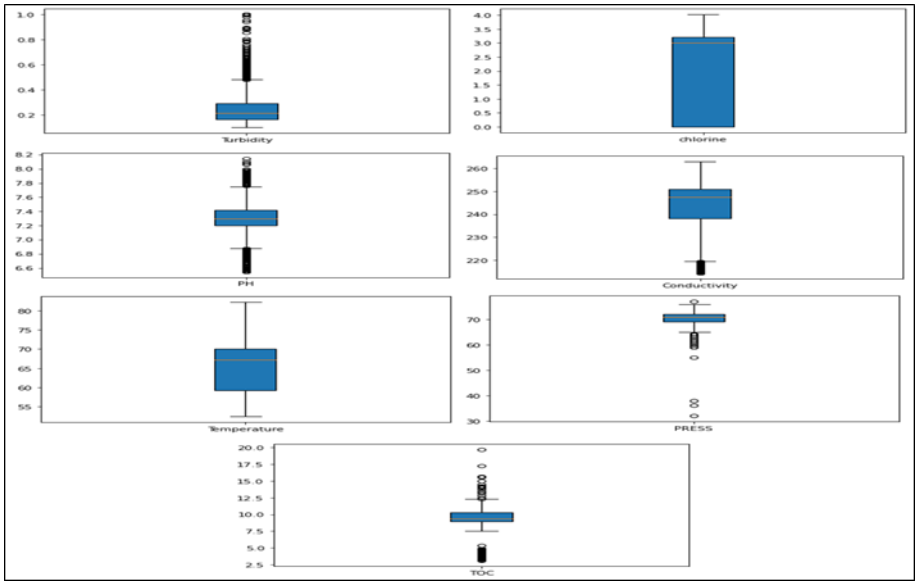


**Fig. 2.** Statistical descriptive of water quality data using Boxplot

Temperature and conductivity seem highly skewed, and the outlier data can be noticed from the distribution of PH, turbidity, and TOC.

### 3.2 Data Representation Concerning Water Conditions (Data Labeling)

The dataset has been categorized into five classes: Dark_Blue (DBlue), Purple, Red, Orange, and Yellow. These five classes represented the level of water contamination. The levels are distributed in the range 0-4, where the 0 level (DBlue) represented the pure water (no contamination), while level 4 (Yellow) shows the high contamination water. This has been conducted based on chlorine values in which SWT is used to specify the minimum and maximum contamination rates. Fig. 3 shows the 3-D scatter plot of the water samples, which depicts the relationship between the water quality parameters and the condition of the water samples. As shown in Fig. 3 (a), the water quality is better if they have low turbidity with high pressure. However, the water samples are more contaminated when they have low TOC with low temperatures, as shown in Fig. 3 (b). PH shows a marginal impact on the water condition, as the data is accumulated more with the low value of PH (see Fig. 3 (c)). The temperature shows the impact of the water quality among the pressure and the chlorine—the water quality increases when the temperature is low, and the pressure is high. In contrast, water is likely to be highly contaminated when the temperature is low and the pressure is low. (see Fig. 3. (d))

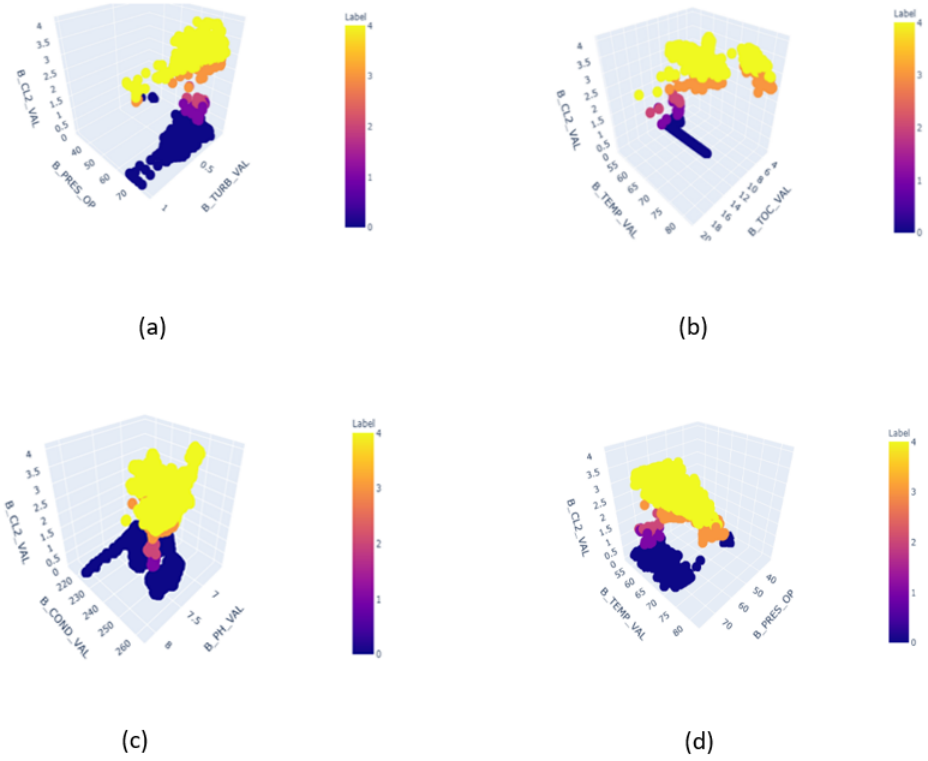


**Fig. 3.** The relation between the water quality parameters and contamination level

## 4 Modeling Implementation

### 4.1 Support Vector Machine (SVM)

SVM was implemented using the following parameters C=1, *cache_size*= 200, degree=3, gamma=aut, kernel='rbf'. Moreover, the performance of SVM was validated using $K-Fold\ cross-validation$ which is described in the subsection 5.1.

### 4.2 Artificial Neural Network (ANN)

ANN was implemented using a sequential model that comprises three layers. Input layer: this layer accepts the six water quality parameters used in our case. The activation function used with this layer is rectified linear. This function returns the standard ReLU activation max(x, 0), the maximum of 0, and the input tensor. Hidden layers: two hidden layers have been used in this model; the number of neurons in the first hidden layer was ten and in the second hidden

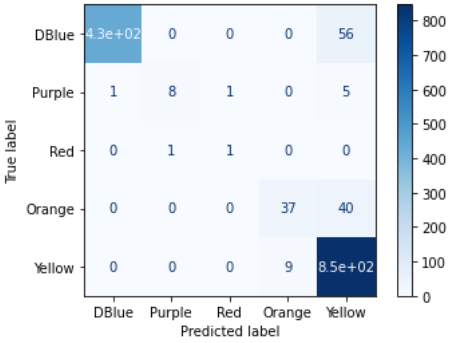


**Fig. 4.** Confusion matrix of the SVM classifier

layer was eight; the rectified linear function was used as an activation function for both hidden layers. Output layer: this layer produces the five different classes of water samples. Softmax is an activation function used by this layer; this function converts a real vector to a vector of unconditional probability. Other options used with ANN implementation are optimizer, losses, and metrics. For the optimizer that use to compile the model, we used the "Adam" optimizer. Losses functions are intended to calculate the error a model should attempt to minimize during training, and we used "categorical_crossentropy." The last thing, we have used accuracy as a metric to evaluate the model.

## 5 Results

In this section, the performance of SVM and ANN are presented in more detail. A confusion matrix was used to evaluate the performance of these models with all metrics included with this matrix. A confusion matrix is an N x N matrix used to assess a multiclass classification models' results, where N is the target class numbers. The matrix compares the actual target values to those that the machine learning model predicts. This gives us a holistic view of how well our classification model works and what types of mistakes.

### 5.1 SVM Performance

The confusion matrix is calculated for the SVM model and represented in Fig. 4 in which the x-axis represents the predicted features (labels). The y-axis indicates the actual features (labels). For instance, in our case from Fig. 4, looking to the top left box, it has a value of 434 (4.3e+02), and the three consecutive values are zero while the last value is 56. This means that the SVM model correctly predicted 434 samples of DBlue class out of a total of 482 samples used in the test dataset. On the other hand, looking at the second row of the confusion matrix that refers to the Purple class, we notice that the model was able to

**Table 1.** SVM performance (a)

| | FP | FN | TP | TN |
|---|---|---|---|---|
| DBlue | 1 | 56 | 434 | 949 |
| Purple | 1 | 7 | 8 | 1424 |
| Red | 1 | 1 | 1 | 1437 |
| Orange | 9 | 40 | 37 | 1354 |
| Yellow | 101 | 9 | 847 | 483 |

**Table 2.** SVM performance (b)

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| DBlue | 0.99 | 0.89 | 0.94 | 490 |
| Purple | 0.88 | 0.53 | 0.67 | 15 |
| Red | 0.50 | 0.50 | 0.50 | 2 |
| Orange | 0.80 | 0.48 | 0.60 | 77 |
| Yellow | 0.89 | 0.99 | 0.94 | 856 |

correctly classify eight samples out of the total of 15 samples and missed seven samples that were predicted as Dblue, Red, and Yellow consecutively. Using the same way for the rest of the classes, we can notice the correct and incorrect predictions for the other classes (Red, Orange, and Yellow).
The overall effectiveness of the classification model can be described using the resulting scores of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). These performance metrics in Table 1. are usually understood and calculated based on the confusion matrix described and presented in Fig 4. Other performance metrics (precision, recall, and f1-score) used to evaluate the model are computed and presented in Table 2. The results in Table 2 shows that the precision for DBlue class is 0.99 which means that the model produce no or very low false positive. In other hand, the recall of the Orange is very low that means that the model predicts high false positive for the same class. Moreover, Table 2 shows the precision and the F1-score, where the precision often indicates the models are correctly predicted, and the F1-score measures the average of the true positive rate (recall) and precision. Finally, the accuracy of SVM in the classification process is 89%.

**K- Fold Cross-Validation** k-fold cross-validation uses to evaluate the performance of SVM. K-fold is less biased as it ensures that every observation in the original dataset appears in the training and test set. Hence, the dataset sliced into $10-folds$. It was tested using a dataset with a size equal 20% from the size of the training dataset that includes 80% from the original size of the dataset. The score values of the 10fold crossvalidation are [0.90, 0.87, 0.90, 0.88, 0.88, 0.88, 0.90, 0.90, 0.90, and 0.88] with 10-Fold cross-validation Accuracy equal 0.89 (+/- 0.02).

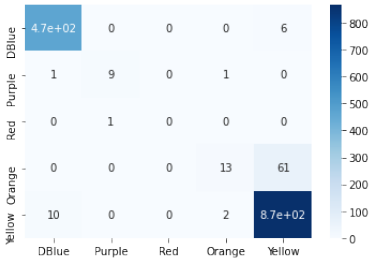


**Fig. 5.** ANN confusion matrix

**Table 3.** ANN Performance (a)

| | FP | FN | TP | TN |
|---|---|---|---|---|
| DBlue | 11 | 10 | 466 | 953 |
| Purple | 0 | 8 | 14 | 1418 |
| Red | 0 | 2 | 0 | 1438 |
| Orange | 7 | 58 | 28 | 1347 |
| Yellow | 71 | 11 | 843 | 515 |

### 5.2 ANN Performance

The confusion matrix of ANN is represented in Fig. 5, and as most of the values are zero, this means that the ANN model shows a perfect classifier. From Fig 5. looking to the top left that indicates the first class DBlue, we can see that the model was correctly classified 469 samples out of a total of 475 that are available in the test samples, and six sample were classified incorrectly as Yellow class. In the second line the ANN classifier was able to correctly classified 9 samples out of 11 samples as Purple. Using the same way, we can notice the correct and incorrect prediction of all the classes. Based on the confusion matrix computed in Fig. 5, all the mercies performance (TP, TN, FP, FN,) of the ANN classified model are estimated and presented in Table 3. Furthermore, the precision and the f1-score have been computed and summarized in Table 4. The results in Table 4 shows precision for DBlue class is 0.98 which means that the model produces very low false positive of DBlue in Table 3. In other hand, the recall of the DBlue class is 0.99 that means that the model predicts very low false positive for the same class in Table 3. Finally, Fig. 6 and Fig. 7 show the performance of the ANN in terms of loss rate and accuracy. The loss rate summarizes the errors made by the model for each sample in the training and validation dataset. The accuracy made by the model is reported, compared to the actual values and the percentage of misclassification is then determined. We notice that the loss rate is reduced as the number of epochs increases while the accuracy increased till it arrived at the steady-state for both at epochs=100. The overall accuracy for the ANN model is 94%.

From the above analysis, it is clear that the ANN has achieved more accuracy than SVM and has outperformed most of the machine learning approaches

**Table 4.** ANN Performance (b)

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| DBlue | 0.98 | 0.99 | 0.98 | 475 |
| Purple | 0.90 | 0.82 | 0.86 | 11 |
| Red | 0.0 | 0 | 0 | 1 |
| Orange | 0.81 | 0.18 | 0.29 | 74 |
| Yellow | 0.93 | 0.99 | 0.96 | 879 |

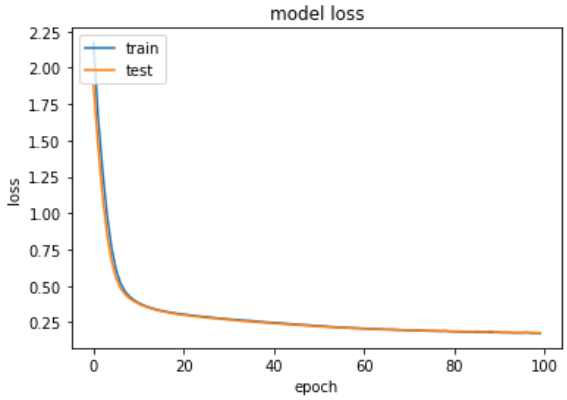


**Fig. 6.** ANN model loss

that were reviewed in section 2. However, the downside of ANN is the computational power and time required during the learning phase for the construction of deep neural network models. Despite this, once the ANN model is created, classification of water contamination can be achieved in real-time, similar to other machine learning algorithms.

## 6 Conclusion

In this paper, we applied deep learning classification algorithms for water quality in a water distribution system. The performance of two supervised learning techniques, namely SVM and ANN, has been investigated and evaluated. The analysis used a real dataset of water quality from the US EPA institute. The data consists of the following water quality features chlorine, temperatures, turbidity, conductivity, total organic carbon, PH, and pressure. The contamination scenarios were generated using the EPANET tool, the most known tool used in the water domain. The results show that ANN achieved the best performance when compared to the reviewed ML algorithms including SVM with an overall accuracy of 94% for multiclass classification.

For future work, ANN models are applied at the edge for real-time water contamination detection. Here, the learning is performed at the cloud to utilise its

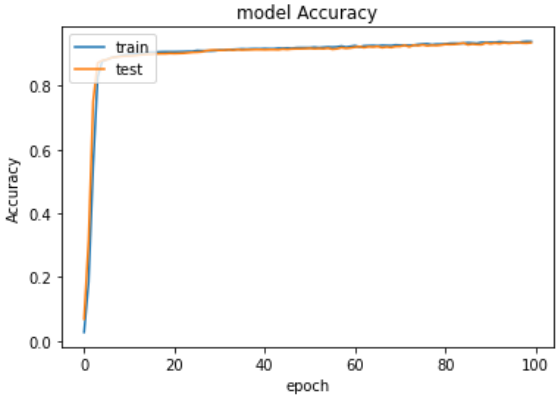


**Fig. 7.** ANN model accuracy

substantial computational power and stable power supply. Moreover, to improve the classification, an ensemble-based system is applied at the cloud to combining diverse models representing readings from various water stations.

## 7 Acknowledgment

This research is supported by European Union's Horizon 2020 research and innovation program Under the Marie Skłodowska-Curie–Innovative Training Networks (ITN)- IoT4Win-Internet of Things for Smart Water Innovative Network (765921)

## References

1. 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*, vol. 3, no. 3, pp. 294–307, 2019.
2. J. Yu, L. Xu, X. Xie, D. Hou, P. Huang, and G. Zhang, "Contamination event detection method using multi-stations temporal-spatial information based on bayesian network in water distribution systems," *Water*, vol. 9, no. 11, p. 894, 2017.
3. M. S. Khorshidi, M. R. Nikoo, E. Ebrahimi, and M. Sadegh, "A robust decision support leader-follower framework for design of contamination warning system in water distribution network," *Journal of Cleaner Production*, vol. 214, pp. 666–673, 2019.
4. Y. Jiang, X. Yang, P. Liang, P. Liu, and X. Huang, "Microbial fuel cell sensors for water quality early warning systems: Fundamentals, signal resolution, optimization and future challenges," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 292–305, 2018.
5. E. Q. Shahra and W. Wu, "Water contaminants detection using sensor placement approach in smart water networks," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2020.

6. E. Q. Shahra, W. Wu, and M. Romano, "Considerations on the deployment of heterogeneous iot devices for smart water networks," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Cloud & Big Data Computing, Internet of People and Smart City Innovation*. IEEE, 2019, pp. 791–796.
7. K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2464–2479, 2020.
8. S. Negi, Y. Kumar, and V. Mishra, "Feature extraction and classification for emg signals using linear discriminant analysis," in *2016 2nd International Conference on Advances in Computing, & Automation (ICACCA)*. IEEE, 2016, pp. 1–6.
9. A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 145–180, 2016.
10. S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (svm) learning in cancer genomics," *Cancer Genomics-Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.
11. F. Shaukat, G. Raja, R. Ashraf, S. Khalid, M. Ahmad, and A. Ali, "Artificial neural network based classification of lung nodules in ct images using intensity, shape and texture features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 10, pp. 4135–4149, 2019.
12. B. Jeong, H. Cho, J. Kim, S. K. Kwon, S. Hong, C. Lee, T. Kim, M. S. Park, S. Hong, and T.-Y. Heo, "Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data," *Diagnostics*, vol. 10, no. 6, p. 415, 2020.
13. K. Chang, J. L. Gao, W. Y. Wu, and Y. X. Yuan, "Water quality comprehensive evaluation method for large water distribution network based on clustering analysis," *Journal of Hydroinformatics*, vol. 13, no. 3, pp. 390–400, 2011.
14. H. Mohammed, I. A. Hameed, and R. Seidu, "Machine learning: based detection of water contamination in water distribution systems," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2018, pp. 1664–1671.
15. N. Oliker and A. Ostfeld, "A coupled classification–evolutionary optimization model for contamination event detection in water distribution systems," *Water research*, vol. 51, pp. 234–245, 2014.
16. J. Arad, M. Housh, L. Perelman, and A. Ostfeld, "A dynamic thresholds scheme for contaminant event detection in water distribution systems," *Water research*, vol. 47, no. 5, pp. 1899–1908, 2013.
17. R. Murray, T. Haxton, S. McKenna, D. Hart, K. Klise, M. Koch, E. Vugrin, S. Martin, M. Wilson, V. Cruze *et al.*, "Water quality event detection systems for drinking water contamination warning systems—development, testing, and application of canary," *EPAI600IR-lOI036, US*, 2010.
18. B. KOWALSKA, E. HOŁOTA, and D. KOWALSKI, "Simulation of chlorine concentration changes in a real water supply network using epanet 2.0 and watergems software packages," *WIT Transactions on The Built Environment*, vol. 184, pp. 39–48, 2018.
19. X.-F. Li and W. A. Mitch, "Drinking water disinfection byproducts (dbps) and human health effects: multidisciplinary challenges and opportunities," 2018.
20. P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.
21. B. I. Babura, M. B. Adam, A. R. A. Samad, A. Fitrianto, and B. Yusif, "Analysis and assessment of boxplot characters for extreme data," in *Journal of Physics: Conference Series*, vol. 1132, no. 1. IOP Publishing, 2018, p. 012078.