

Real-Time Multi-Class Classification of Water Quality Using MLP and Ensemble Learning

Essa Q. Shahra¹, Shadi Basurra¹, and Wenyan Wu¹

¹ Faculty of Computing, Engineering and Built Environment, Birmingham City University, Birmingham B5 5JU, United Kingdom;
[Essa.Shahra, Shadi.Basurra, Wenyan.Wu]@bcu.ac.uk,

Abstract. The major goal of water management planning and the iterative evaluation of operational policies and procedures is to ensure that good water quality is always maintained. Effective water monitoring requires examining many water samples, which is a time-consuming and labour-intensive process that takes a lot of effort. This paper aims to evaluate the quality of drinking water samples with high accuracy by using multi-class classification models: multilayer perceptron (MLP) and ensemble learning. Real datasets with different sizes that include the essential water quality parameters have been used to train and test the developed models. The results showed the effectiveness of the developed models in detecting water contamination with high accuracy in both datasets used. The results demonstrate that bagging Ensemble learning outperforms the multilayer perceptron with an overall accuracy of 94% for station-A and 92% for station-B compared to MLP, which shows an overall accuracy of 89% for station-A and 87% for station-B.

Keywords: Classification, Ensemble learning, MLP, Water Quality.

1 Introduction

It is difficult to maintain water quality in water distribution system (WDS) because of the impacts of vast distances and long periods spent traveling from the water supply to the consumers [1]. In addition, the nature of WDS renders it susceptible to external disruptions, such as sabotage, because of its design. Therefore, researchers in academia and industry have been looking into creating contamination warning systems due to these considerations [2]. Two essential components to creating these systems are sensor placement techniques in a WDS and data processing methods for analysing the collected data [3]. It is possible to evaluate water quality with the help of water quality sensors installed on a WDS. The optimum sensor placement technique will enhance detection efficiency by gathering all the data. Sensors generate large quantities of data streams, which need to be analysed in real-time to identify anomalous occurrences, particularly those that may result in water contamination in WDS [4]. Furthermore, the procedure of detecting water contamination was characterized as being very difficult [5]. It includes using common water quality indicators

such as turbidity, electrical conductivity, pH, and the concentration of residual chlorine to identify pollution [6]. In light of these measures, it is also necessary to develop a detection model that, based on these data, accurately detects changes in water quality. The development of multi-parameter water quality models is complex due to the massive data and the many factors that must be considered [7]. Numerous classification methods based on statistics and machine learning have been proposed, including linear discriminant analysis (LDA), regularized discriminant analysis (RDA), support vector machines (SVM), quadratic discriminant analysis (QDA), and k-n, and neural networks (NN). This paper aims to investigate how ensemble learning helps enhance the accuracy of predicting water contamination in the water distribution system. This paper's organisation is as follows: section 2 re-views the relevant work conducted in the area, section 3 explains in detail the methodology of the models applied in this work, and section 4 presents the results and the discussion. Finally, section 6 concludes the work

2 Related Work

This study looks into the approaches used to assist tackle water quality challenges. In most cases, traditional lab analysis and statistical analysis are utilised in research to help determine water quality. Still, some investigations use machine learning approaches to help identify an optimum solution to the water quality problem. Authors in [8] used Monte Carlo optimised artificial neural networks to predict water quality in the Danube River (Serbia). They predicted 18 common water quality metrics (WQPs) on active monitoring sites. Shakhari et al. in [9] proposed an algorithm for an efficient analysis of the water quality data. The proposed method not only monitors continuous water quality, but it also develops a superior classification model for other continuous datasets. Instead of reducing the dataset's attributes, they provide three extra reference indications that depend on the real characteristics. The classification accuracy of the proposed approach is compared to others classification models. In comparison to previous approaches, the proposed method provides higher classification accuracy. In [10] they have created a system for monitoring two water quality indices in real-time: chlorophyll-a (Chl-a) and dissolved oxygen (DO). They used CEEMDAN and VMD algorithms with LSSVM and ELM models to create a two-layer decomposition. In [11] have developed a classification system called the Water Quality Classification Index in an effort to create a four-level pollution index on the water quality of the marine environment. Dilmi et al. [12] they proposed a novel strategy to improving water quality classification that is based on the fusion of deep learning and feature extraction techniques. They used an advanced deep learning technology called Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs) to build an intelligent model for predicting drinking water quality. Authors in [13] investigated a set of supervised machine learning methods for predicting the water quality index (WQI), a unique index for describing the general quality of water, and the water quality class (WQC), a separate

class formed based on the WQI. Temperature, turbidity, pH, and total dissolved solids are the four input parameters used in the suggested technique. Several machine learning, statistical, and mathematical models for contamination detection have been reviewed. Accordingly, we aim to investigate and develop two machine learning models: MLP and Ensemble learning in a real dataset to obtain high multi-class classification accuracy.

3 Methodology

Two distinct models are used in this work: Multilayer Perceptron (MLP) and Ensemble learning.

3.1 Multilayer Perceptron

A sequential model comprised of three layers was used to construct MLP: Input layer: this layer takes the seven water quality indicators that we utilise in our dataset, and its dimension is 7, but after feature selection, we utilised just six as inputs: turbidity, PH, TOC, temperature, conductivity, and pressure. This layer's activation function is rectified linear. The used method returns the usual ReLU activation maximum ($x, 0$), the input tensor value, and the maximum of 0. Hidden layers: In this model, two hidden layers were utilised; the first hidden layer had 10-neurons, and the second hidden layer comprised 8-neurons; in order to activate both hidden layers, it was necessary to utilise the rectified linear function as the activation function. Output layer: the output of this layer is the 5-classes of water samples that are generated. With the Softmax activation function, this layer may convert an actual vector into a vector of category likelihood with a single operation. An optimizer, losses, and measurements are all options for MLP implementation in addition to the standard options. We utilised the "Adam" optimizer to build the model. Losses functions are used to determine the quantity of data that should be minimised during training. We used the term "categorical crossentropy" in this instance. Finally, the metrics utilised to assess the model's performance; the accuracy measure is used to evaluate our model.

3.2 Ensemble Learning

Bootstrap aggregation, or bagging for short, is a kind of ensemble machine learning used for classification in this study. A bootstrap sample is a random sample from a dataset that has been replaced. Replacement refers to replacing a sample taken from the dataset, enabling it to be picked again and perhaps many times in the new sample. This implies that the sample may include duplicates of the original dataset's instances. The ensemble bagging algorithm is implemented in Python using the scikit-learn machine learning library. The model is constructed using randomness, implying that a slightly different model will be produced each time the method is performed on the same data. The model is evaluated using repeated stratified k-fold cross-validation with three repetitions and 10-folds.

3.3 Datasets

The US Environmental Protection Agency (EPA) provided the dataset used in this research (Koch and McKenna, 2011). This information was gathered from two stations A, and B. These datasets can be utilised for both training and testing purposes. For three months, time series at two-minute intervals were gathered. Each station’s data include chlorine (Cl), temperature (TEMP), total organic carbon (TOC), PH, pressure (PRES), and conductivity (COND). The datasets for each station are different; station-A has 60,000 time steps, and station-B has 70,000 time steps. See table 1 which includes a sample of these datasets. The dataset is divided into five classes: Dblue, Purple, Red, Orange, and Yellow. These classifications corresponded to the degree of water contamination. The values were distributed in the range 0-4, with 0 (DBLue) representing clean water (no contaminants) and (Yellow) representing highly contaminated water.

Table 1. Dataset sample

TIME- STEP	Cl2	COND	PH	TEMP	TOC	TURB	PRES OP
08/06/2007 00:00	2.06	44.9974	9.52652	14.6669	0.644472	0.18992	35.4681
08/06/2007 00:05	2.07	44.9911	9.52593	14.6564	0.641543	0.191788	35.4726
08/06/2007 00:10	2.07	44.9784	9.53607	14.645	0.650092	0.190603	35.4697
08/06/2007 00:15	2.07	44.9593	9.53467	14.6227	0.625475	0.191484	35.4696
08/06/2007 00:20	2.06	44.9784	9.53114	14.6161	0.64178	0.190641	35.4721

3.4 Feature Selection

The selection of features is one of the most crucial steps in any machine learning task. A feature means a column when a dataset is used. With any dataset, not every column (feature) will necessarily affect the output variable. Feature selection can be made in several ways, but overall, three different categories are available: 1) filtering method, 2) wrapper method, and 3) embedded method [14]. In this work we have applied the embedded method on our data set; more detail about this approach is in the next sub-section.

Embedded Method: The embedded method is another feature selection that has been applied to our data set. This method is iteratively in the sense that the model training process takes care of each iteration and carefully extracts those features which most contribute to the training for a particular iteration

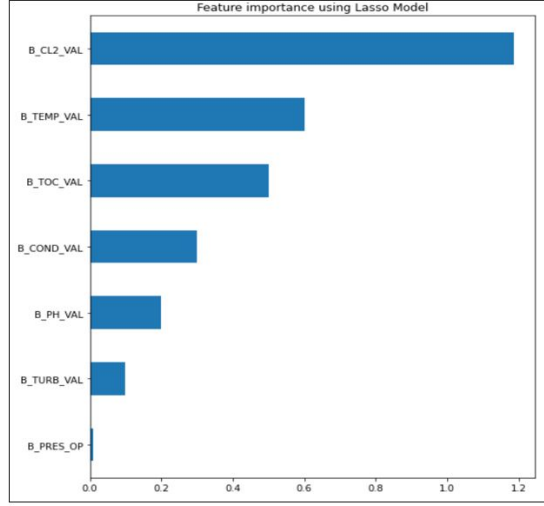


Fig. 1. Feature selection using Lasso regularization

[15]. Regularisation methods are the most widely used embedded methods that penalize a feature provided a threshold of the coefficient. We have applied the Lasso regularization method, and in this method, If the feature is irrelevant, it will also penalize its coefficient and make it 0. Therefore, coefficient = 0 features are removed, and the remaining ones are taken. Figure 1 shows the importance of all features using Las-so model, and from the figure, we notice that the features are arranged based on their significance, and we can see that the pressure has scored close to zero (0.009) that may remove from the data set, and we also notice that chlorine got a very high score above 1 this because the output was made based on the chlorine.

4 Results and Discussion

The effectiveness of MLP is discussed in more depth in this section, and the performance was assessed using a confusion matrix that incorporated all metrics. A Con-fusion Matrix is an $N \times N$ matrix that is utilised to evaluate the outcomes of classification models, where N is the number of desired classes. When comparing actual target values in y-axis to predicted target values in x-axis produced by machine learning models, the matrix may be used to identify differences. This provides us with a comprehensive picture of how well our classification models perform and the kinds of errors that occur. There is a list of rates that are often computed based on the confusion matrix for a classification model called: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

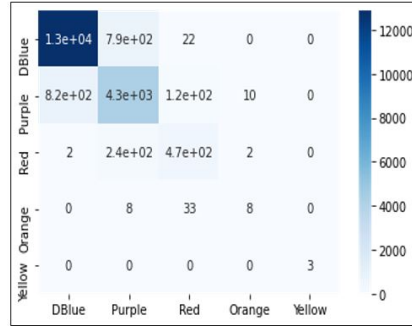


Fig. 2. MLP Confusion matrix (Station-A)

4.1 Multi-Class Classification: MLP

Station-A: The confusion matrix of the MLP model for station-A is shown in Figure 2. This matrix shows the ability of the developed model to correctly predict the samples of the water for each class. From this matrix the model looks a perfect classifier as indicated by the fact that the majority of the values in this matrix are zero. Table 2 presents estimates of the metrics performance (TP, TN, FP, FN, FP) of the MLP-A classifier model, which includes all of the metrics variables. Figures 3 depicts the performance of the MLP in terms of accuracy and loss rate. There is a noticeable reduction in the loss rate as the number of epochs rises, and an improvement in accuracy, until it reaches the SteadyState for both at epochs=100. The accuracy of the MLP model is 94% in total.

Table 2. MLP performance (Station-A)

	FP	FN	TP	TN
DB	826	12864	816	5161
Purple	1042	4261	958	13406
Red	179	472	244	18772
Orange	12	8	41	19606
Yellow	0	3	0	19664

Station-B The B-station is the station that contains dataset with the largest size (72,000 sample). The performance of this station was evaluated in the same way of station-A. The confusion matrix was calculated and displayed in Figure 4, and because the data used in this station is very large, we will notice the numbers contained in this station are also large, whether they are in the correct prediction or the wrong prediction. Other parameters were also calculated and summarized in Table 3. The accuracy rate of this model and the rate of loss are shown in Figures 5. The general evaluation showed that the model of this station is estimated at 89.77% and the loss rate is 30%.

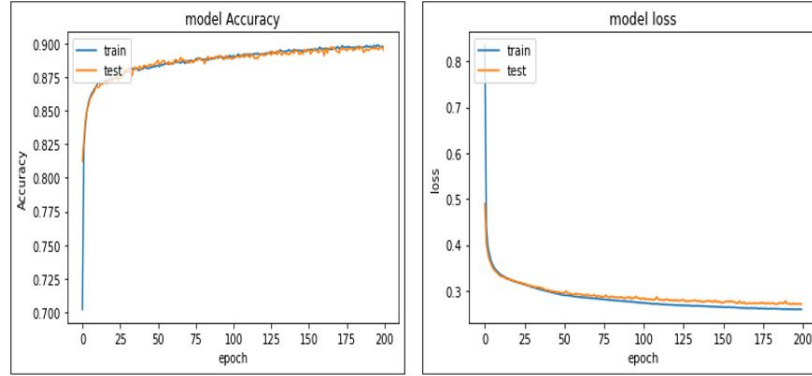


Fig. 3. MLP model accuracy and loss (station-A)

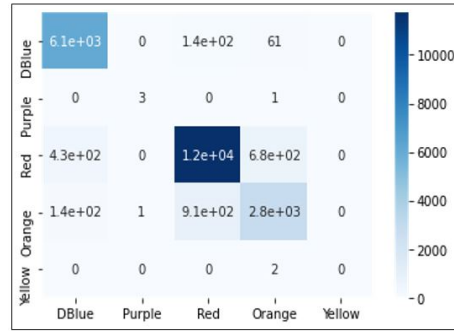


Fig. 4. MLP Confusion matrix (Station-B)

4.2 Multi-Class Classification: Ensemble Learning

Station-A: The accuracy of the model has been compared to the ensemble members as shown in Figure 6. We can notice that after approximately two members, the bagged ensemble outperforms the performance of any individual model. The overall accuracy of the model has been enhanced using bagged ensemble up to 94%.

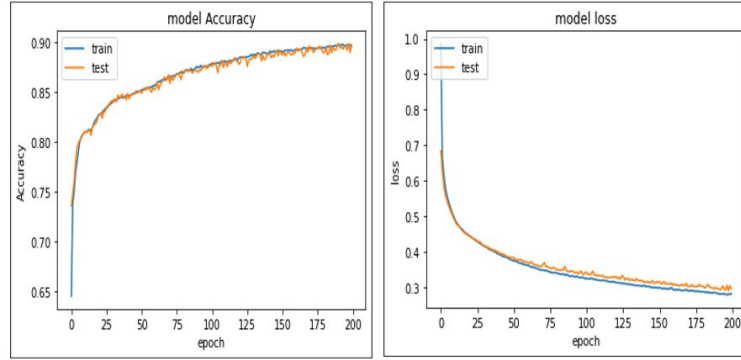
Station-B: The performance accuracy of the model is shown in Figure 7. We can notice that after approximately two members, the bagged ensemble outperforms the performance of any individual model. The overall accuracy of the model has been enhanced using bagged ensemble up to 92%.

4.3 Discussion

We have used machine learning in this paper to accurately detect contaminants in drinking water. Two different methods have been used to accurately test the

Table 3. MLP Performance (Station-B)

	FP	FN	TP	TN
DB	575	6123	196	16151
Purple	1	3	1	23040
Red	1043	11730	1115	9157
Orange	746	2824	1051	18424
Yellow	0	0	2	23043

**Fig. 5.** MLP model accuracy and loss (station-B)

models' abilities to classify the contamination and these models applied to two datasets of different size. In general, the results showed that bagging ensemble learning was able to over-come the MLP models and showed a significant improvement in the accuracy of classification (see Table 4)

Table 4. The performance of our classification models

	Multilayer Perceptron Network	Bagging Ensemble Learning
Station-A	89%	94%
Station-B	87%	92%

5 Conclusion

This paper discusses how to utilise machine learning to classify water quality in a water distribution system. This was achieved by analysing and comparing the performance of two techniques namely MLP, and bagging ensemble learning. The study was conducted using real-world datasets on water quality from two different areas (A, B) with different sizes. The data set includes chlorine, temperature, turbidity, conductivity, PH, pressure, and total organic carbon. The

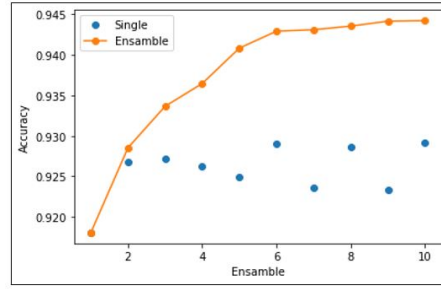


Fig. 6. Ensemble classification, station-A

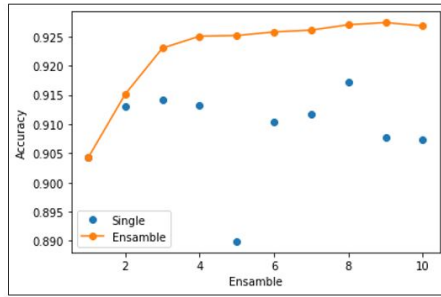


Fig. 7. Ensemble classification, station-B

data has been labeled into five different categories: no contamination, low contamination, medium contamination, above average, and highly contaminated. The results showed that bagging ensemble learning overcame MLP and showed a significant improvement in the accuracy of classification. The results showed that our models used multi-classifications with high accuracy using MLP and ensemble learning (89%, 94%).

6 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. H. Mala-Jetmarova, N. Sultanova, and D. Savic, "Lost in optimisation of water distribution systems? a literature review of system design," *Water*, vol. 10, no. 3, p. 307, 2018.

2. J. Simões, Z. Yang, and T. Dong, "An ultrasensitive fluorimetric sensor for pre-screening of water microbial contamination risk," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 258, p. 119805, 2021.
3. 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.
4. H. Mahmoud, W. Wu, and M. M. Gaber, "A time-series self-supervised learning approach to detection of cyber-physical attacks in water distribution systems," *Energies*, vol. 15, no. 3, p. 914, 2022.
5. E. Q. Shahra, W. Wu, and R. Gomez, "Human health impact analysis of contaminant in iot-enabled water distributed networks," *Applied Sciences*, vol. 11, no. 8, p. 3394, 2021.
6. E. Q. Shahra, W. Wu, S. Basurra, and S. Rizou, "Deep learning for water quality classification in water distribution networks," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2021, pp. 153–164.
7. T. M. Tung, Z. M. Yaseen *et al.*, "A survey on river water quality modelling using artificial intelligence models: 2000–2020," *Journal of Hydrology*, vol. 585, p. 124670, 2020.
8. T. Mitrović, D. Antanasijević, S. Lazović, A. Perić-Grujić, and M. Ristić, "Virtual water quality monitoring at inactive monitoring sites using monte carlo optimized artificial neural networks: a case study of danube river (serbia)," *Science of the Total Environment*, vol. 654, pp. 1000–1009, 2019.
9. S. Shakhari and I. Banerjee, "A multi-class classification system for continuous water quality monitoring," *Heliyon*, vol. 5, no. 5, p. e01822, 2019.
10. E. Fijani, R. Barzegar, R. Deo, E. Tziritis, and K. Skordas, "Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters," *Science of the total environment*, vol. 648, pp. 839–853, 2019.
11. R. Wan, F. Meng, E. Su, W. Fu, and Q. Wang, "Development of a classification scheme for evaluating water quality in marine environment receiving treated municipal effluent by an integrated biomarker approach in meretrix meretrix," *Ecological Indicators*, vol. 93, pp. 697–703, 2018.
12. S. Dilmi and M. Ladjal, "A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques," *Chemometrics and Intelligent Laboratory Systems*, vol. 214, p. 104329, 2021.
13. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, p. 2210, 2019.
14. C.-W. Chen, Y.-H. Tsai, F.-R. Chang, and W.-C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Systems*, vol. 37, no. 5, p. e12553, 2020.
15. X. Zhang, J. Wang, and Y. Gao, "A hybrid short-term electricity price forecasting framework: Cuckoo search-based feature selection with singular spectrum analysis and svm," *Energy Economics*, vol. 81, pp. 899–913, 2019.