International Journal of Advanced Engineering and Technology
www.allengineeringjournal.com
ISSN: 2456-7655
Received: 14-04-2024, Accepted: 08-05-2024, Published: 31-05-2024
Volume 8, Issue 1, 2024, Page No. 16-18

Predictive analytics for chlorine residual management: A comparative study of machine learning algorithms

Pierpaolo Di Nosotti
Department of Computer Science, Università degli Studi di Milano, Italy

Abstract
This study evaluates the efficacy of various machine learning algorithms in predicting residual chlorine levels in drinking water distribution systems. By comparing models such as Random Forest, Support Vector Machine, and Artificial Neural Networks, the research aims to identify the most accurate and reliable method for maintaining optimal chlorine levels, thereby ensuring water safety and quality.

Keywords: Machine learning algorithms, water safety, artificial neural networks

Introduction
Residual chlorine is critical for disinfecting drinking water and preventing microbial contamination. However, maintaining optimal chlorine levels throughout the distribution system is challenging due to factors like temperature, pH, and flow rate. Traditional methods of chlorine management often fail to account for these variables adequately. This study explores the application of machine learning algorithms to predict residual chlorine levels, offering a data-driven approach to enhance water quality management.

Objective of study
The main objective of this study is to compare the performance of various machine learning algorithms in predicting residual chlorine levels in drinking water distribution systems to identify the most accurate and reliable method for maintaining optimal chlorine concentrations and ensuring water quality.

Methods
Water quality data were collected from multiple sensors within a municipal water distribution system over a period of one year. The dataset includes measurements of residual chlorine, temperature, pH, flow rate, and other relevant parameters. Data were recorded at hourly intervals to capture detailed temporal variations.

The study compared three machine learning algorithms
- **Random Forest (RF):** An ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions.
- **Support Vector Machine (SVM):** A supervised learning model that analyzes data for classification and regression analysis.
- **Artificial Neural Network (ANN):** A computational model inspired by the human brain, capable of identifying complex patterns and relationships in data.

The dataset was split into training (70%) and testing (30%) subsets. Each model was trained using the training data and evaluated on the testing data. Performance metrics included Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to assess the accuracy and reliability of the predictions.

Results
Model Performance
The performance of each machine learning model was evaluated based on the specified metrics:

Metric	Random Forest	Support Vector Machine	Artificial Neural Network
MAE	0.15 mg/L	0.20 mg/L	0.12 mg/L
MSE	0.03 mg/L²	0.05 mg/L²	0.02 mg/L²
R^2	0.89	0.82	0.92

Table 1: Random Forest Model's Performance

Metric	Value
MAE	0.15 mg/L
MSE	0.03 mg/L²
R^2	0.89

Table 2: Support Vector Machine Model's Performance

Metric	Value
MAE	0.20 mg/L
MSE	0.05 mg/L²
R^2	0.82

Table 3: Artificial Neural Network Model's Performance

Metric	Value
MAE	0.12 mg/L
MSE	0.02 mg/L²
R^2	0.92

Predictive Accuracy
The ANN model's superior performance is attributed to its ability to capture nonlinear relationships and interactions between variables. The model successfully identified complex patterns in the data, leading to more accurate predictions of residual chlorine levels. The RF model's ensemble approach also provided robust predictions but was slightly less accurate than the ANN. The SVM, while useful, struggled with the complexity of the data, resulting in lower predictive accuracy.

This comparative analysis demonstrates that the Artificial Neural Network (ANN) model has the highest accuracy and reliability for predicting residual chlorine levels in drinking water distribution systems, followed by the Random Forest (RF) and Support Vector Machine (SVM) models.

Discussion and Analysis of Results

The study evaluated the performance of three machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN)—in predicting residual chlorine levels in drinking water distribution systems. The key performance metrics were Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

The Artificial Neural Network (ANN) model demonstrated the highest accuracy among the three models, with the lowest MAE and MSE and the highest R^2 value. This indicates that the ANN model is highly effective in capturing the complex, nonlinear relationships between the variables influencing residual chlorine levels. The model's architecture, inspired by the human brain, allows it to identify and learn intricate patterns in the data, leading to superior predictive performance. The Random Forest (RF) model also showed strong performance, with relatively low error rates and a high R^2 value. As an ensemble learning method, RF constructs multiple decision trees during training and aggregates their predictions, enhancing robustness and reducing overfitting. Although slightly less accurate than the ANN, the RF model's ability to handle diverse and complex datasets makes it a reliable choice for chlorine residual management. The Support Vector Machine (SVM) model, while useful, was less accurate compared to the ANN and RF models. SVM's effectiveness in classification and regression tasks is well-known, but in this study, it struggled to cope with the complexity of the data. The higher error rates and lower R^2 value suggest that SVM may not be as well-suited for predicting residual chlorine levels in the context of this study. The superior performance of the ANN model can be attributed to its ability to capture and model nonlinear relationships and interactions between multiple variables, such as temperature, pH, and flow rate, which significantly influence chlorine residual levels. The flexibility and adaptability of ANNs make them highly effective in dynamic and complex environments. The RF model's ensemble approach also provided robust predictions. Its strength lies in reducing variance and improving the stability of predictions, making it a reliable option for chlorine residual management, albeit slightly less accurate than the ANN. The SVM model's lower predictive accuracy highlights its limitations in handling highly complex and nonlinear data patterns. While SVMs are powerful for certain tasks, they may not be the optimal choice for predicting residual chlorine levels in this specific application. The findings of this study have several practical implications for water utilities and public health management. By using machine learning models like ANN and RF, water utilities can predict future chlorine levels and make timely adjustments to dosing strategies, ensuring consistent and safe chlorine residuals throughout the distribution system. Continuous real-time data analysis enabled by machine learning models can help detect anomalies and potential issues before they impact water quality, thereby ensuring a higher level of water safety. Accurate predictions of chlorine levels allow for more efficient use of chlorine, reducing waste and operational costs. This optimization can lead to significant cost savings for water utilities while maintaining high water quality standards. Despite the promising results, several challenges need to be addressed for practical implementation. High-quality, continuous data are essential for training accurate models. Utilities must invest in robust data collection and management systems to ensure the availability of reliable data. Integrating machine learning models into existing water management systems requires technical expertise and infrastructure upgrades. Collaboration between data scientists and water management professionals is crucial for successful integration. Models must be scalable to handle large datasets and varying conditions across different regions and distribution networks. Ensuring scalability and adaptability of the models is essential for broader application.

Conclusion

This comparative study highlights the potential of machine learning algorithms, particularly Artificial Neural Networks, in predicting residual chlorine levels in drinking water systems. The findings suggest that ANN models provide the highest accuracy and reliability, making them a valuable tool for water quality management. By leveraging predictive analytics, water utilities can enhance their ability to maintain safe and consistent chlorine levels, ultimately improving public health outcomes. Further research and investment in data infrastructure and integration strategies will be essential for the successful implementation of these models in real-world settings

References

1. Soares TSM, Latorre M, Laporta G, Buzzar M. Spatial and seasonal analysis on leptospirosis in the municipality of São Paulo, Southeastern Brazil, 1998 to 2006. Rev Saude Publica. 2010;44(2):283-91.
2. Benacer D, Thong K, Min N, Verasahib KB, Galloway R, Hartskeerl R, *et al*. Epidemiology of human leptospirosis in Malaysia, 2004-2012. Acta Trop. 2016;157:162-8.
3. Zounemat-Kermani M, Ramezani-Charmahineh A, Adamowski J, Kisi O. Investigating the management performance of disinfection analysis of water distribution networks using data mining approaches. Environ Monit Assess. 2018 Jul;190:1-5.
4. Rajabova ND, Mambetullaeva SM. Ecological assessment of drinking water resources using the residual chlorine and analysis by probabilistic mathematical methods: On the example of Nukus city and Amudaryo district. Int J Geogr Geol Environ. 2020;2(2):01-3.
5. Park J, Lee CH, Cho KH, Hong S, Kim YM, Park Y. Modeling trihalomethanes concentrations in water treatment plants using machine learning techniques. Desalination Water Treat. 2018 Apr 1;111:125-33.
6. Gibbs MS, Morgan N, Maier HR, Dandy GC, Nixon JB, Holmes M. Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. Math Comput Model. 2006 Sep 1;44(5-6):485-98.
7. Kang G, Gao JZ, Xie G. Data-driven water quality analysis and prediction: A survey. In: 2017 IEEE third international conference on big data computing service

and applications (BigDataService). IEEE; c2017 Apr 6. p. 224-32.
8. Tran TT, Choi JW, Le TT, Kim JW. A comparative study of deep CNN in forecasting and classifying the macronutrient deficiencies on development of tomato plant. Appl Sci. 2019 Apr 17;9(8):1601.