VNUHCM Journal of Environment and Earth Sciences 2026, 10(2):1387-1400

Open Access Full Text Article

Research article

Non-revenue water: a predictive application using machine learning

Tran Thi Kim[1,2,*], Huynh Phuoc Sang[3], Nguyen Van Tin[3]

Use your smartphone to scan this QR code and download this article

[1]University of Science, 227 Nguyen Van Cu, Cho Quan Ward, Ho Chi Minh City, Vietnam

[2]Vietnam National University, Dong Hoa Ward, Ho Chi Minh City, Vietnam

[3]Ho Chi Minh City University of Natural Resources and Environment, 236B Le Van Sy, Tan Son Hoa Ward, Ho Chi Minh City, Vietnam

Correspondence

Tran Thi Kim, University of Science, 227 Nguyen Van Cu, Cho Quan Ward, Ho Chi Minh City, Vietnam

Vietnam National University, Dong Hoa Ward, Ho Chi Minh City, Vietnam

Email: ttkim@hcmus.edu.vn

History

- Received: 10-10-2025
- Revised: 04-02-2026
- Accepted: 04-08-2026
- Published Online: 27-08-2026

DOI : https://doi.org/10.32508/vnuhcmj-ees.v10i2.849

ABSTRACT

The shortage and loss of water resources is one of the critical issues in urban areas worldwide, with approximately 32 billion cubic meters of treated water lost annually through leakage from urban distribution systems. Forecasting the future Non-Revenue Water (NRW) rate is essential to mitigate resource depletion and economic waste, contributing to the United Nations' Sustainable Development Goal 6 on Clean Water and Sanitation. This study investigates the application of four machine learning models: Random Forest Regression (RF), Extreme Gradient Boosting Regression (XGBoost), Support Vector Regression (SVR), and K-Nearest Neighbours (KNN) to calculate and forecast the NRW rate of seven joint stock water supply companies under the Saigon Water Corporation (SAWACO) in Ho Chi Minh City, Vietnam. The dataset comprises monthly records from 2019 to 2024, including produced water volumes at treatment plants, distribution meter readings, and customer consumption volumes. The results demonstrate that the RF model provides the best performance for NRW rate prediction, with $R^2 > 0.8$, RMSE < 1.5%, and MAE < 1% during training. The XGBoost and SVR models also produce satisfactory results with R2 > 0.7, RMSE and MAE respectively below 2% and 1.5%, although their test set performance is not as strong as RF. Forecasts through 2030 reveal a seasonal increase in NRW during February, attributed to higher domestic and industrial water demand that elevates pressure on the pipeline network. The results also indicate that water supply companies in inner-city areas record lower NRW (12–17%) compared to those serving suburban areas, reflecting disparities in water supply infrastructure and the impact of urbanization. The findings provide a scientific foundation for SAWACO's sustainable water management strategy and support the strategic development roadmap of Ho Chi Minh City.

Key words: Machine learning, Water supply, Non-Revenue Water

INTRODUCTION

Population growth, urbanization, along with the impact of climate change, will put pressure on access to water sources for domestic, industrial, and agricultural needs. According to EIB (2023), it is projected that more than 240 million people will still not have access to improved drinking water sources, and nearly 1.4 billion people are forecasted to remain without access to basic sanitation services [1]. Urbanization increases the demand for water supply services in the context of global water scarcity as well as the impact of climate change, making NRW management a significant challenge for the sustainability of the water sector. NRW is one of the main causes leading to increased operating costs of water supply companies, thereby increasing water prices [2]. With the approach to the United Nations' Sustainable Development Goal 6 (SDG 6) on Clean Water and Sanitation [3], reducing NRW is highly meaningful in sustainable development [4].

According to the World Bank study in 2023, approximately 32 billion cubic meters of treated water are lost annually due to leakage from urban water distribution systems worldwide [5]. For developing countries, the deterioration of water supply networks increases leakage and causes NRW [6]. According to the American Water Works Association (AWWA) and the International Water Association (IWA), NRW in the supply system is divided into two main categories: apparent losses and real losses [7,8]. Apparent losses, also known as non-physical or management losses, mainly arise from metering inaccuracies, fraud, or deficiencies in consumption data management. In contrast, real losses (or physical losses) refer to the volume of water lost due to physical failures in the system, including leaks in transmission and distribution pipelines, losses in raw water pipes and treatment facilities, leaks and overflows at storage tanks, as well as losses at service connections up to the customer's metering point [6,9].

Globally, many studies have applied AI models to forecast NRW and have achieved reliable results. Models such as XGBoost [10,11], Random Regression [11–13] and Support Vector Machine [14] have shown

Cite this article : Kim T T, Sang H P, Tin N V. **Non-revenue water: a predictive application using machine learning**. *VNUHCM J. Environ. Earth Sci*. 2026; 10(2):1387-1400.

high effectiveness in predicting loss points. The study by Candelieri A et al. (2014)[15] used the EPANET water network simulation model in combination with the SVM method to identify leak locations in pipeline systems. The results showed that this method has very high reliability in identifying and locating leakage areas. Following the same research direction, Soldevila et al. (2019)[16] used the XGBoost model to locate losses in the water supply network. This model was trained on simulation data from EPANET and delivered superior results in terms of accuracy and inference speed compared to traditional methods. Alves Coelho et al.[17] also proposed an accurate NRW monitoring system using wireless sensor networks (WSN), combined with machine learning and real-time data transmission. The system includes water flow sensors that transmit data via LoRa and NB-IoT protocols to a server for processing. The collected data are analyzed by machine learning models (Random Forest, Decision Trees, SVM, Neural Networks, and XGBoost) to detect and locate leaks. During real-world testing, the Random Forest model gave the most accurate results, achieving 85% accuracy during training and 75% under simulated deployment conditions. The system is low-cost, easy to implement, and suitable for use in agricultural irrigation systems and public water networks, contributing to NRW reduction and improved resource management efficiency. For machine learning models applied in leak detection, Gou G et al. (2020)[18] compared the performance of three models—Random Forest, Gradient Boosting, and Neural Networks—in detecting losses based on pressure and flow data. The results indicated that the Random Forest model provided high performance with good accuracy and strong noise resistance. Recently, Liu H (2025) developed a hydraulic model to locate leaks, combining a genetic algorithm with a backpropagation neural network, aiming to predict leakage points in the water supply system of an exhibition area in Eastern China[19].

In Vietnam, some initial studies on NRW using the machine learning approach have achieved certain results. Tuan et al. (2022)[20] evaluated and selected currently advanced machine learning algorithms including the RFR, SVR and XGBoost models, Light Gradient Boosting (LGB) model, and CatBoost (CBR) model to serve the purpose of forecasting the number of NRW points in the water supply network. On that basis, a suitable model is proposed for simulating leakage point forecasting to support effective NRW management in a typical water distribution network in Ho Chi Minh City (HCMC), showing the potential of using machine learning models. Dung HM and Quoc TT (2025) assessed the current status and proposed technical and economic solutions to reduce NRW at Tan Hoa Water Supply Joint Stock Company (belong to SAWACO) using the EPANET model. The study estimated the Economic Level of Leakage (ELL) in order to reduce NRW to the level where the cost of reducing NRW equals the value of the saved water volume, contributing to the protection of water resources and the environment[21]. In general, the identification of leakage points has been widely studied; however, the approach focusing on forecasting the NRW rate still lacks sufficient research.

In an effort to reduce NRW, several companies have begun to make preliminary calculations of the loss rate index based on very limited water audit data. The SAWACO and its member water supply companies currently have an average NRW rate of about 18%. In the objective of reducing NRW, protecting resources, and promoting sustainable development, it is necessary to minimize the level of NRW in the distribution network, especially NRW due to leakage, through the development and application of various techniques[20] for detection. This study investigates the application of machine learning techniques to forecast the Non-Revenue Water (NRW) rate of the joint stock companies under SAWACO and to project future water losses through the year 2030. The objective is to determine long-term NRW trends that can serve as a scientific foundation for SAWACO's sustainable water supply management and support the strategic development roadmap of Ho Chi Minh City. The analysis is based on a multi-source monthly dataset, including: (i) produced water volume recorded at the clean-water production meters of the treatment plants (input), (ii) water volume measured at the main distribution meters (input), (iii) customer consumption volume collected from end-user meters, and (iv) the observed monthly NRW rate.

METHODOLOGY

Study area

The study was conducted in HCMC, the largest economic center of Vietnam. The city has a natural area of approximately 2,061 km^2 with relatively low and flat terrain, influenced by the hydrological regimes of the Dong Nai, Sai Gon, and Vam Co rivers. According to the General Statistics Office (2023), the average population of HCMC reached 9.46 million people; however, if including unregistered residents, the actual population could reach up to 14 million. The urban water supply system in HCMC is managed by the SAWACO, with an extensive distribution network consisting of thousands of kilometers of transmission and distribution pipelines. Currently, the total water supply capacity exceeds 2.4 million m^3/day and night, and the average actual water output is 1,928,000 m^3/day, supplying clean water to nearly 100% of households in urban areas[22]. However, HCMC is facing significant challenges related to the NRW rate, ranging from 15–20% depending on the area, mainly due to leakage in old pipeline systems, inaccurate metering, and unsynchronized data management[23].

This study focuses on forecasting NRW in the areas of water plants Joint Stock Companies (JSCs) under SAWACO in HCMC. These plants play an important role in providing clean water to the city's population. The NRW rate is an important indicator reflecting the efficiency of water resource management and usage.

Materials

The dataset used for training and testing the model consists of monthly records from 2019 to 2024 for seven JSCs, namely Ben Thanh, Cho Lon, Nha Be, Tan Hoa, Phu Hoa Tan, Thu Duc, and Gia Dinh Water Supply JSCs. The input variables include: (i) monthly produced water volumes measured at the clean water production meters of the Corporation's water plants; (ii) monthly water volumes recorded at the main meters of the JSCs; and (iii) monthly customer consumption volumes measured at customer meters of the JSCs. The output variable is the monthly NRW rate of each JSC, which is subsequently compared with the observed NRW data to evaluate model performance.

The Can Gio JSC was excluded from the analysis due to the presence of negative NRW values in 2020 and 2021, primarily resulting from data collection disruptions during the COVID-19 pandemic. Likewise, abnormal negative NRW values observed at several water plants during the same period were removed from the dataset. These anomalies were attributed to measurement errors and meter-reading difficulties under pandemic-related operational constraints. Following data cleaning, only consistent and physically meaningful records were retained for model training and validation.

Research Methodology

The process of forecasting the NRW rate is carried out through five main steps, as described in Figure 1.

Step 1 – Data input: The input variables used include the input production volume from the water plants, the input production volume from the water supply JSCs, the volume of water consumption measured at customer meters, and the NRW rate of the water supply JSCs.

Step 2 – Model execution: The data is split into two sets, in which 80% is used for training and the remaining 20% is used for testing, to ensure the model's accuracy and generalization capability. The machine learning algorithms applied include Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and K-Nearest Neighbors (KNN).

Step 3 – Model performance evaluation: After training and testing the models, the performance of each model is evaluated based on metrics such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). This evaluation helps to compare and determine which model provides the best accuracy and forecasting capability.

Step 4 – Model selection: Based on the evaluation results in the previous step, the model with the highest performance is selected for forecasting.

Step 5 – Forecasting the 2030 NRW rate: The model selected in Step 4 is used to forecast the NRW rate of the water supply companies for the year 2030. The forecast results support regulatory agencies and water companies in formulating policies and solutions to minimize NRW in the future.

Machine Learning Models

Random Forest: Random Forest (RF) is a machine learning method belonging to the ensemble model group, which uses a collection of decision trees independently trained on randomly sampled subsets from the main training set (illustrated in Figure 2a). This method helps reduce the overfitting phenomenon of individual trees, thanks to its ability to decrease variance and improve overall accuracy[27,28]. Specifically, each tree in the RF is built from a bootstrap dataset and uses only a randomly selected subset of features, thereby ensuring independence among the trees. The number of trees is selected through cross-validation or the out-of-bag error method to optimize model performance[29].

Extreme Gradient Boosting: Extreme Gradient Boosting (XGBoost) is a prominent boosting model that works by sequentially constructing decision trees, where each new tree learns from the errors of the previous tree to improve prediction capability (illustrated in Figure 2b). This continuous training process allows the model to adapt well to complex data, reduce overall error, and increase accuracy[28,30,31]. XGBoost uses a loss function to evaluate the contribution of each tree and often requires hyperparameter tuning to achieve optimal performance. Techniques such as grid search

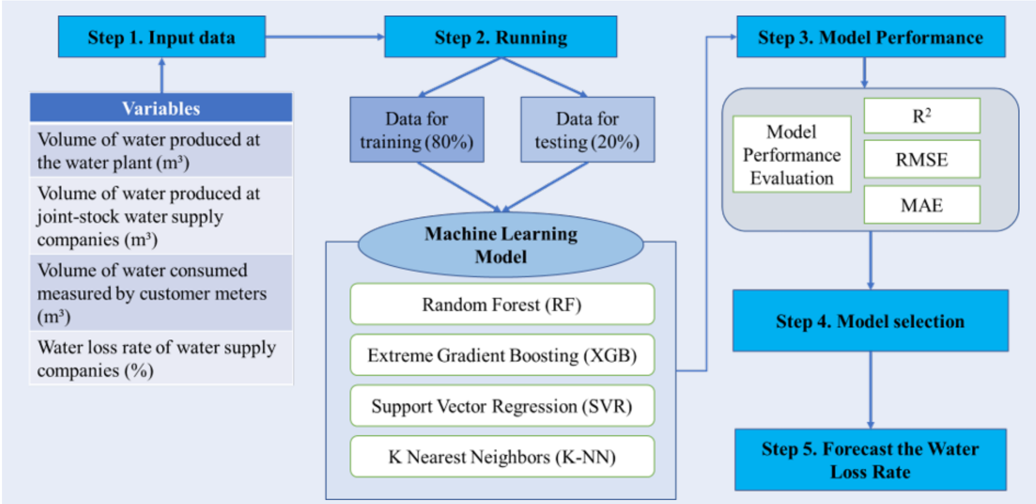

Figure 1: Implementation steps (Source: Author's team)

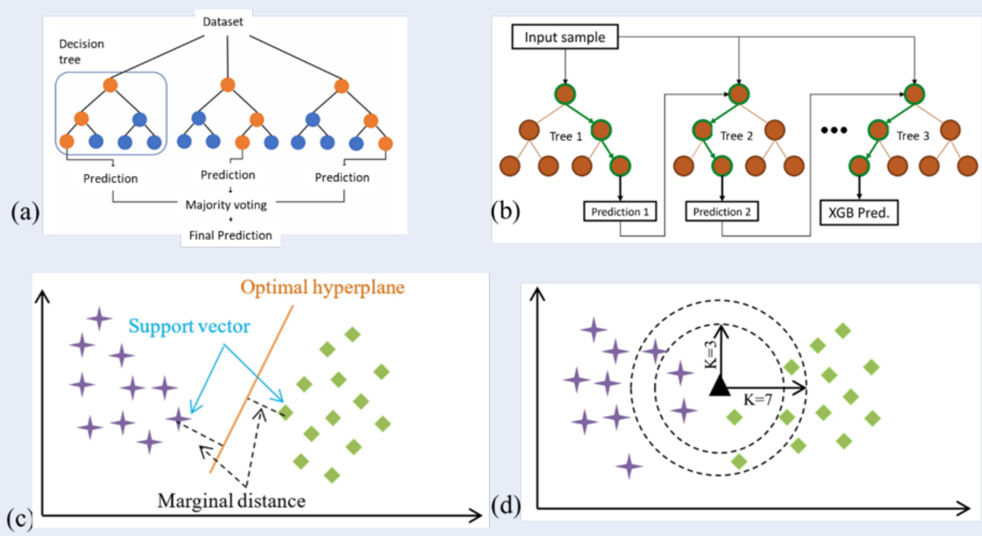

Figure 2: Diagram illustrating the algorithms: Random Forest (a)[24], Extreme Gradient Boosting (b)[25], Support Vector Machine (c) and K-Nearest Neighbors (d)[26]

combined with cross-validation are commonly applied to select appropriate parameters, especially in high-uncertainty problems such as groundwater salinity forecasting or NRW rate prediction[28].

Support vector Machine: Support Vector Machine (SVM)[32] is a machine learning method introduced by Vapnik and colleagues in the 1990s for binary classification problems, and later extended to various applications such as regression, anomaly detection, and clustering. The fundamental problem of SVM is binary classification: given r data points in n-dimensional space, each point is labeled as belonging to one of two classes, denoted +1 or −1. The goal of SVM is to find the optimal separating hyperplane such that this hyperplane divides the dataset into two parts, with all points of the same class lying on one side of the hyperplane (illustrated in Figure 2c). SVR is a variant of the Support Vector Machine algorithm used for regression tasks, i.e., predicting continuous values instead of classification. In Support Vector Regression, this hyperplane is used to predict continuous output values[26].

K-Nearest Neighbors (K-NN): K-NN is a learning algorithm that can handle both classification and regression problems by considering the distance (proximity) between input instances (illustrated in Figure 2d). It is referred to as a non-parametric learning algorithm because, unlike other supervised learning algorithms, it does not learn an explicit model function from the training data. Instead, the algorithm simply memorizes all previous instances and then predicts the output by searching the training set for the *k* nearest neighbors and then: (1) for classification – predicts the majority class among those *k* nearest neighbors, whereas (2) for regression – predicts the output value as the average of the values of its *k* nearest neighbors[26].

Performance Evaluation Metrics

Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

Coefficient of Determination (R^2)[33] – is a measure of how well the model fits, indicating the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The higher the R^2 value, the better the fit of the model. R^2 is calculated using Equation (01):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (01)$$

where: i: number of independent variables; n: number of samples; y_i : actual value; $\hat{y_i}$: predicted value; $\bar{y}$: mean value.

The coefficient R^2 describes the strength of the linear relationship between simulated and observed values and, in several benchmark studies, is classified into four qualitative levels: *Very Good, Good, Satisfactory,* and *Not Satisfactory*. R^2 values $\geq$ 0.80 at the daily scale or $\geq$ 0.85 at the monthly scale are considered *Very Good*; values between 0.70 and 0.80 (daily) or 0.75 and 0.85 (monthly) are rated as *Good*; values between 0.60 and 0.70 (daily) or 0.65 and 0.75 (monthly) are considered *Satisfactory*; and values below these thresholds are regarded as *Not Satisfactory*[34,35].

In statistics, Mean Absolute Error (MAE) is a measure of the error between paired observations expressing the same phenomenon[36]. MAE is calculated using Equation (02):

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \quad (02)$$

where: i: number of independent variables; n: number of samples; y_i: actual value; $\hat{y_i}$: predicted value.

Root Mean Square Error (RMSE)[28,33] is the standard deviation of the residuals, where residuals are the differences between the actual values and the values predicted by the model. RMSE measures the dispersion of these residuals or, in other words, indicates how closely the data points cluster around the regression or prediction line. The formula is given as Equation (03):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \quad (03)$$

where: i: number of independent variables; n: number of samples; y_i: actual value; $\hat{y_i}$: predicted value.

RMSE and MAE are error-based metrics that quantify the magnitude of deviations between simulated and observed values and are expressed in the same units as the variable of interest. Model performance with respect to these metrics is evaluated on a relative basis, where smaller RMSE and MAE values indicate lower simulation errors and better model performance. RMSE is more sensitive to large errors, whereas MAE reflects the average error and is less affected by outliers. A model is considered to perform well when RMSE and MAE are small, stable, and consistent across simulation periods and are reasonable relative to the variability of the observed data[34,35].

RESULTS AND DISCUSSION

Training and Testing the NWR Models

Training results

On the training set (Figure 3 and Figure 4), all three models RF, XGB, and SVR showed very high performance, with determination coefficients R^2 ranging from 0.96 to nearly 1.0 in most companies. Specifically, XGB achieved an absolute R^2 of 1.000 in many companies such as Thu Duc, Ben Thanh, Cho Lon, Nha Be, and Trung An, indicating perfect fitting ability with the training data. Similarly, SVR also reached R^2 > 0.99 in many cases like Cho Lon (0.9996), Nha Be (0.9991), and Gia Dinh (0.9996) (Table 1). RF achieved the highest R^2 at Nha Be (0.9902), showing good learning capability on the training data. However, K-NN had significantly lower performance on the training set, with R^2 only reaching 0.6516 (Thu Duc) and 0.6907 (Ben Thanh), indicating it could not learn the data as well as the other models.

However, significant differences were recorded when evaluating on the test set (Figure 5 and Figure 6). Although XGB trained very well, its performance dropped sharply on the test set, shown by low R^2 values such as 0.4752 (Thu Duc), 0.3236 (Cho Lon, Ben Thanh), indicating overfitting. RMSE and MAE of XGB also increased, with RMSE = 1.819 (Cho Lon) and MAE = 0.6066 (Cho Lon). For SVR, test results varied widely: at Nha Be and Tan Hoa, R^2 were 0.9235 and 0.8577 respectively, but at Ben Thanh and Phu

Hoa Tan, they dropped to 0.0223 and 0.1171 (Table 1). This shows that SVR depends heavily on each company, performing well in some places but poorly in others.

Regarding Random Forest, although it did not reach absolute R^2 on training like XGB, the model showed the most stability when moving to testing. For example, at Ben Thanh, R^2 = 0.8121; at Gia Dinh, 0.8594; and notably at Cho Lon, it reached 0.8199. Although in some places R^2 was not as high as SVR (such as at Thu Duc), the RMSE and MAE of RF were more stable, indicating better generalization ability, avoiding overfitting like XGB and large fluctuations like SVR. Therefore, considering all three criteria R^2, RMSE, and MAE, Random Forest was selected as the optimal model to forecast NRW rates for the water supply companies in this study.

Testing results

NRW Forecasts until 2030

The forecast results until 2030 at the JSCs by RF show that the NRW rates are quite variable (Figure 7 and Figure 8). Specifically, Phu Hoa Tan JSC has the lowest loss rate among the calculated units, ranging from 4.6% to 8.1%, indicating that its water supply system has low leakage levels and relatively effective loss control. In contrast, Trung An JSC maintains a high loss rate throughout the year, with forecast values ranging from 15.2% to 18.0%, peaking in February 2030 (18.0%).

Other companies such as Tan Hoa, Cho Lon, and Ben Thanh have medium to high loss rates, often exceeding 15% in many months, reflecting pressure on the water supply system or limitations in leak control. Notably, Tan Hoa reaches its highest loss rate in November 2030 at 17.9%, while Cho Lon records its peak at 17.3% in the same month. Meanwhile, Thu Duc and Gia Dinh maintain relatively stable loss rates, fluctuating from 10.9% to 16.1%, falling into the lower group across the entire system.

Regarding time, a sudden increase in loss rates is recorded in February (Table 2). HCMC has a tropical monsoon climate, with water demand typically rising during the dry season (January–April) due to high temperatures, low rainfall, and increased domestic and industrial needs. Forecast data show many water supply companies experience high loss rates during these months, for example, Trung An (February: 18.0%), Gia Dinh (February: 12.1%), and Thu Duc (February: 13.8%). The sudden increase in demand causes higher pressure on the pipeline network, increasing the likelihood of leaks and losses [37].

Besides weather factors, regional demand also affects loss fluctuations. Central areas such as Ben Thanh and Cho Lon have high population density and commercial activities, with stable water demand year-round, so loss rates fluctuate but not drastically. Meanwhile, suburban areas like Trung An and Nha Be experience rapid demand growth due to urbanization, leading to higher system pressure and greater losses in some months

SAWACO has set a roadmap to reduce NRW in the HCMC water supply system for the period 2022–2025, with a vision toward 2030, aiming to reduce the NRW rate to below 17.5% by 2025 and below 15% by 2030 [38]. Based on the forecast results from the Random Forest machine learning model for NRW rates until 2030, it can be seen that the trend of reducing NRW continues to be maintained positively, aligning with SAWACO's planned roadmap. Specifically, according to the model results, three companies exceed the 15% threshold: Trung An, Tan Hoa, and Cho Lon, indicating that these areas need focused improvement and investment in better measurement technology and loss management. Conversely, three companies have loss rates below 15% but still show some months exceeding 15%, namely Ben Thanh (February and November), Nha Be (February and August), and Thu Duc. This indicates local fluctuations over time that require close monitoring to avoid surpassing the target threshold. Additionally, two companies consistently maintain loss rates below 15% throughout the year: Phu Hoa Tan and Gia Dinh, reflecting effective management and operation of the water supply network in these areas.

Average loss rates between 12–17% are classified as moderate according to international standards from the IWA (International Water Association). According to this organization, NRW below 10% is considered advanced, 10–20% is acceptable, and above 25% requires improvement [39].

According to World Bank studies, NRW rates vary significantly between countries [40]. n developed countries, water supply companies often maintain NRW below 15%, with the most efficient units like Tokyo and Singapore achieving around 3–5% [41]. Singapore recorded the lowest NRW rate globally, about 4.7% in 2012, thanks to applying pressure management technology, automatic leak monitoring, and modern data management systems [42]. Conversely, in many developing countries, NRW commonly ranges from 30–60% or higher. Typical cases include Manila, with NRW exceeding 35–40%, while many cities in Africa report levels above 45%. In Malaysia, NRW remains high at 36.6%, ranging from 18% to 62% between

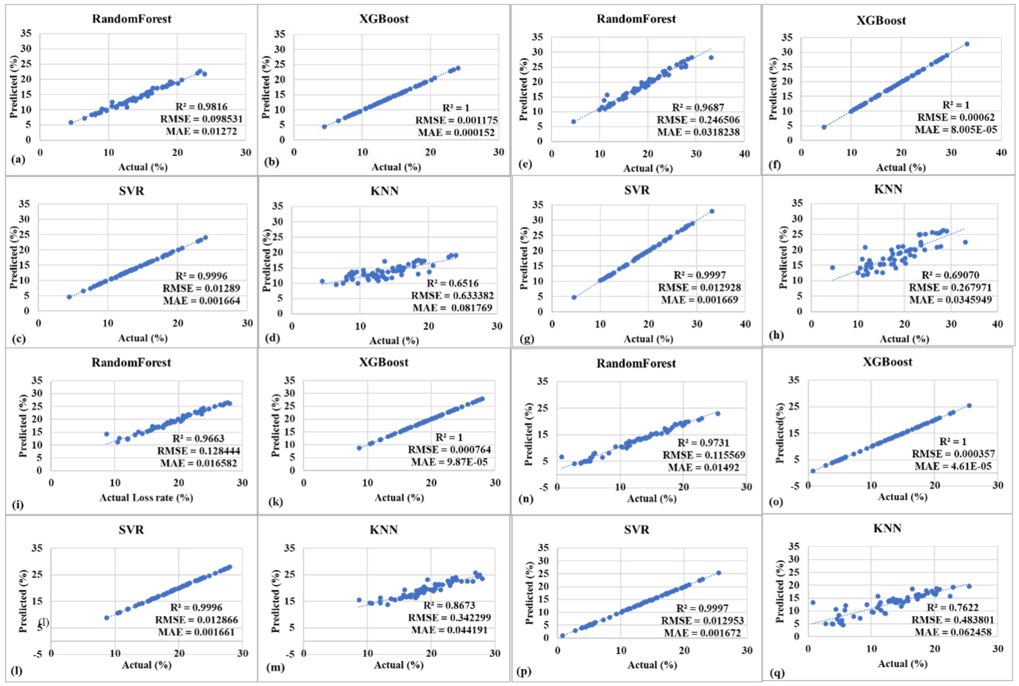

Figure 3: Training results at Thu Duc (with RF (a), XGBoost (b), SVR (c) and (KNN (d)), Ben Thanh (with RF (e), XGBoost (f), SVR (g) and (KNN (h)), Cho Lon (with RF (i), XGBoost (k), SVR (l) and (KNN (m)), and Phu Hoa Tan (with RF (n), XGBoost (o), SVR (p) and (KNN (q)) (Source: Author'steam)

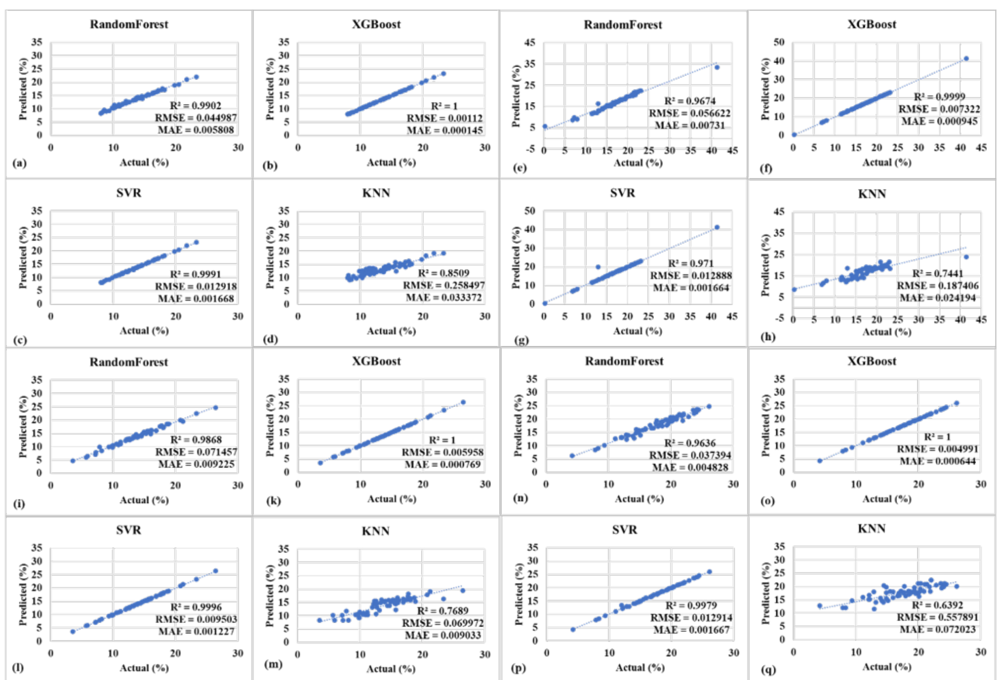

Figure 4: Training results at Nha Be (with RF (a), XGBoost (b), SVR (c) and (KNN (d)), Tan Hoa (with RF (e), XGBoost (f), SVR (g) and (KNN (h)), Gia Dinh (with RF (i), XGBoost (k), SVR (l) and (KNN (m)), and Trung An (with RF (n), XGBoost (o), SVR (p) and (KNN (q)) (Source: Author's team)

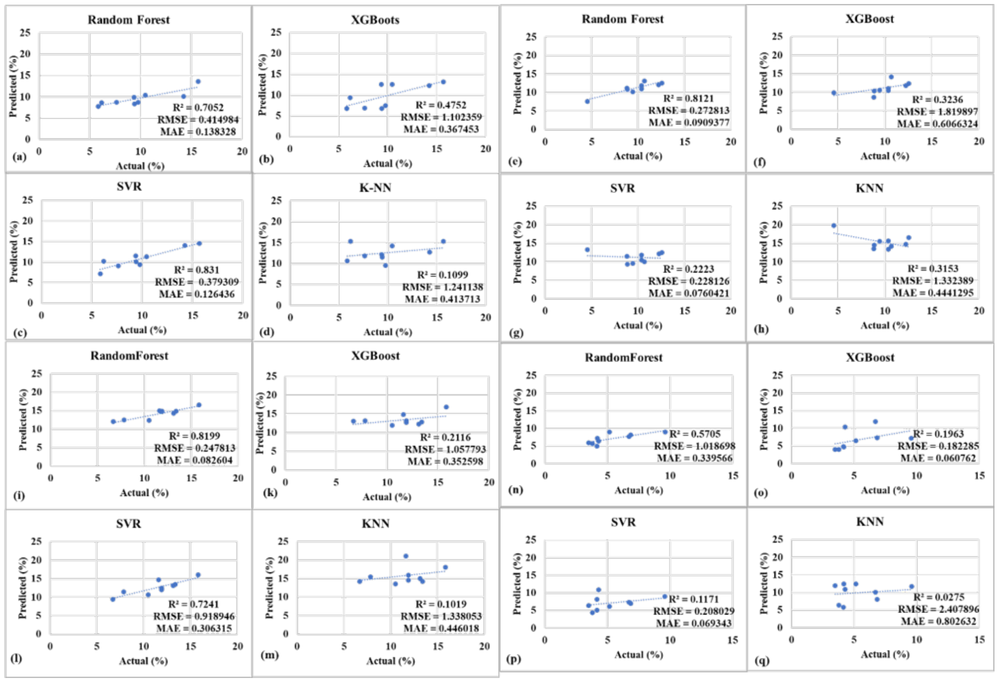

Figure 5: Testing results at Thu Duc (with RF (a), XGBoost (b), SVR (c) and (KNN (d)), Ben Thanh (with RF (e), XGBoost (f), SVR (g) and (KNN (h)), Cho Lon (with RF (i), XGBoost (k), SVR (l) and (KNN (m)), and Phu Hoa Tan (with RF (n), XGBoost (o), SVR (p) and (KNN (q)) (Source: Author's team)

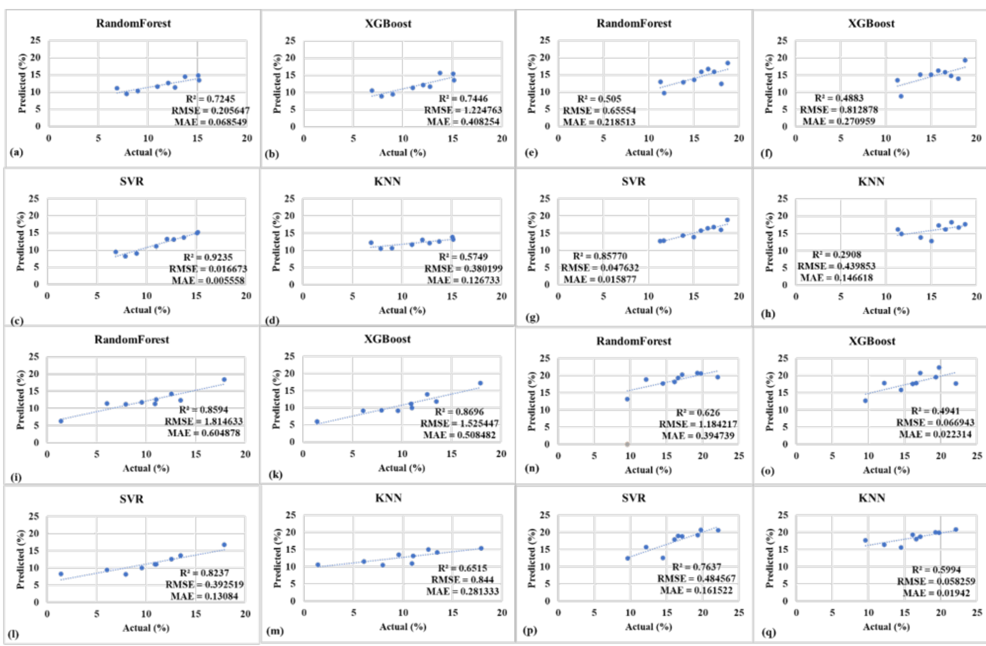

Figure 6: Testing results at Nha Be (with RF (a), XGBoost (b), SVR (c) and (KNN (d)), Tan Hoa (with RF (e), XGBoost (f), SVR (g) and (KNN (h)), Gia Dinh (with RF (i), XGBoost (k), SVR (l) and (KNN (m)), and Trung An (with RF (n), XGBoost (o), SVR (p) and (KNN (q)) (Source: Author's team)

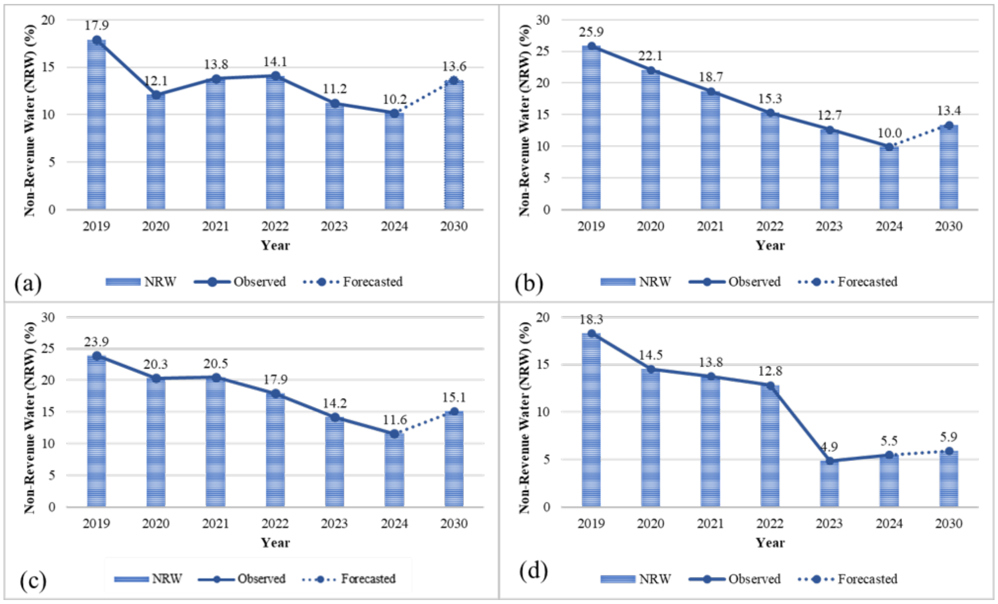

Figure 7: Forecast results at Thu Duc (a), Ben Thanh (b), Cho Lon (c) and Phu Hoa Tan (d) (Source: Author's team)

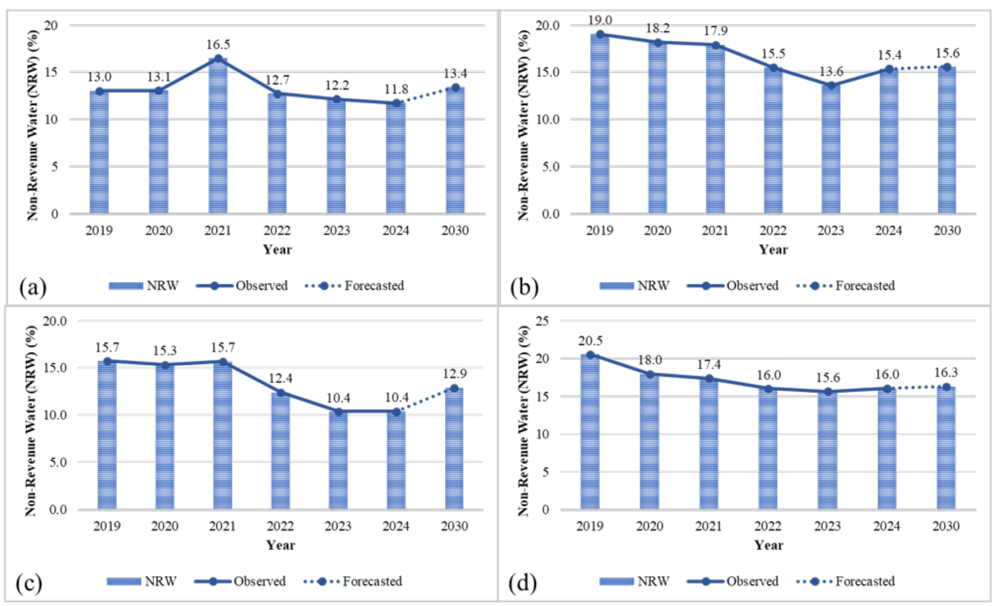

Figure 8: Forecast results at Nha Be (a), Tan Hoa (b), Gia Dinh (c) and Trung An (d) (Source: Author's team)

Table 1: Summary of R^2, RMSE, and MAE indicators of the JSCs models (Source: Author's team)

No.	Water Supply JSCs	Evaluation Metric	RF		XGB		SVR		K-NN	
			Train	Test	Train	Test	Train	Test	Train	Test
1	Thu Duc	R^2	0.9816	0.7052	1.0000	0.4752	0.9996	0.8310	0.6516	0.1099
		RMSE	0.0985	0.4150	0.0012	1.1024	0.0129	0.3793	0.6334	1.2411
		MAE	0.0127	0.1383	0.0002	0.3675	0.0017	0.1246	0.0818	0.4137
2	Ben Thanh	R^2	0.9687	0.8121	1.0000	0.3236	0.9997	0.2223	0.6907	0.3153
		RMSE	0.2465	0.2728	0.0006	1.8199	0.0129	0.2281	0.2680	1.3324
		MAE	0.0318	0.0909	0.0001	0.6066	0.0017	0.0760	0.0346	0.4441
3	Cho Lon	R^2	0.9663	0.8199	1.0000	0.3236	0.9996	0.7241	0.8673	0.1019
		RMSE	0.1284	0.2478	0.0008	1.8199	0.0129	0.9189	0.3423	1.3381
		MAE	0.0166	0.0826	0.0001	0.6066	0.0017	0.3063	0.0442	0.4460
4	Phu Haa Tan	R^2	0.9731	0.5705	1.0000	0.1963	0.9997	0.1171	0.7622	0.0275
		RMSE	0.1156	1.0187	0.0004	0.1823	0.0130	0.2080	0.4838	2.4079
		MAE	0.0149	0.3396	0.0000	0.0608	0.0149	0.0693	0.0625	0.0826
5	Nha Be	R^2	0.9902	0.7245	1.0000	0.7446	0.9991	0.9235	0.8509	0.5749
		RMSE	0.0450	0.2056	0.0011	1.2248	0.0129	0.0167	0.2585	0.3802
		MAE	0.0058	0.0685	0.0001	0.4083	0.0017	0.0056	0.0334	0.1267
6	Tan Hoa	R^2	0.9674	0.5050	0.9999	0.4883	0.9710	0.8577	0.7441	0.2908
		RMSE	0.0566	0.6555	0.0073	0.8129	0.0129	0.0476	0.1874	0.4399
		MAE	0.0073	0.2185	0.0009	0.2710	0.0017	0.0159	0.0242	0.1466
7	Gia Dinh	R^2	0.9868	0.8594	1.0000	0.8696	0.9996	0.8237	0.7689	0.6515
		RMSE	0.0715	1.8146	0.0060	1.5254	0.0095	0.3925	0.0700	0.8440
		MAE	0.0092	0.6049	0.0008	0.5085	0.0012	0.1308	0.0090	0.2813
8	Trung An	R^2	0.9636	0.6260	1.0000	0.4941	0.9979	0.7637	0.6392	0.5994
		RMSE	0.0374	1.1842	0.0050	0.0669	0.0129	0.4846	0.5579	0.0583
		MAE	0.0048	0.3947	0.0006	0.0223	0.0017	0.1615	0.0720	0.0194

states, with 5 out of 13 states having NRW above 50%[43]. According to a World Bank (2006) study, the average NRW in developing countries is often very high, with many water supply systems recording losses over 40% of produced water, sometimes exceeding 50–60%[40] causing significant economic losses to water supply companies.

Compared to other cities worldwide, HCMC shows better NRW levels than many urban areas in developing countries as well as in Southeast Asia, but still remains higher than developed cities. These differences mostly arise from pipeline infrastructure and monitoring technology[44,45]. Developed countries like Japan[46] and Singapore[47] have implemented IoT systems, AI, pressure monitoring, and centralized data management, while Southeast Asian cities are still in the process of upgrading. HCMC currently ranks at a good average level but still has the potential to reduce NRW below 10% if technical and management solutions are deployed comprehensively. This aligns with the recommendations of Liemberger & Wyatt (2018) on improving sustainable water supply efficiency[39].

CONCLUSION

The results show that the Random Forest model provides the best training performance for loss rates on the network, with determination coefficient R^2 >

Table 2: Summary of the 2030 NRW rate forecasts of the water supply JSCs (Source: Author's team)

Unit (%)	Water Supply Joint-Stock Company							
Time	Thu Duc	Ben Thanh	Cho Lon	Phu Hoa Tan	Nha Be	Tan Hoa	Gia Dinh	Trung An
1/2030	12.2	12.1	12.1	4.6	14.5	14.6	13.4	15.8
2/2030	13.8	15.0	16.9	6.6	14.7	16.6	12.1	18.0
3/2030	13.2	13.5	15.3	7.9	12.6	15.5	11.6	15.3
4/2030	15.3	12.6	14.8	5.3	14.6	17.0	12.3	16.2
5/2030	11.2	13.9	14.6	5.8	10.8	13.2	11.7	11.4
6/2030	13.7	13.8	15.4	5.0	14.2	16.3	12.0	17.0
7/2030	12.4	12.8	15.6	6.2	13.0	14.4	11.5	16.4
8/2030	15.6	13.9	15.2	5.8	15.3	17.6	12.7	14.9
9/2030	16.1	13.5	16.7	6.0	14.5	16.1	12.7	16.6
10/2030	13.7	13.2	15.7	6.0	12.5	15.1	10.6	16.1
11/2030	17.0	13.4	17.3	5.8	13.1	16.3	12.6	16.5
12/2030	10.9	13.0	16.2	6.1	10.7	17.9	12.1	16.7

0.8, root mean squared error (RMSE) < 1.5%, and mean absolute error (MAE) < 1%. Extreme Gradient Boosting and Support Vector Regression models also deliver fairly good metrics with R^2 > 0.7, RMSE and MAE below 2% and 1.5% respectively, but their test set performance is not as good as Random Forest. However, the K-Nearest Neighbour model is not highly rated for the collected dataset.

Forecasts until 2030 indicate a sudden increase in NRW recorded in February, due to higher domestic and industrial water demand during this time, which increases pressure on the pipeline network and consequently the risk of leaks and losses. Results also show that central areas have lower NRW compared to suburban areas, due to differences in infrastructure and the impact of urbanization. The predicted loss levels for plants in HCMC range from 12–17%, which is considered acceptable according to international IWA standards. A limitation of this study is that factors such as population fluctuations and changes in water-using households have not yet been incorporated.

CONFLICT OF INTEREST

The authors declare no conflict of interest in the publication of this paper.

AUTHORS' CONTRIBUTIONS

Research idea development: H.P.S., T.T.K. Methodology: H.P.S., N.V.T., T.T.K. Data processing: H.P.S., N.V.T., T.T.K. Result analysis: H.P.S., T.T.K. Writting manuscript: H.P.S., T.T.K.

LIST OF ABBREVIATIONS

AWWA: American Water Works Association
CBR: CatBoost Regression
EIB: European Investment Bank
ELL: Economic Level of Leakage
HCMC: Ho Chi Minh City
IWA: International Water Association
JSC: Joint Stock Company
KNN: K-Nearest Neighbours
LGB: Light Gradient Boosting
MAE: Mean Absolute Error
NRW: Non-Revenue Water
R^2: Coefficient of Determination
RF: Random Forest (Regression)
RMSE: Root Mean Square Error
SAWACO: Saigon Water Corporation
SDG 6: Sustainable Development Goal 6
SVM: Support Vector Machine
SVR: Support Vector Regression
WSN: Wireless Sensor Networks
XGBoost: Extreme Gradient Boosting

REFERENCES

1. EIB. EIB water sector orientation: Building climate-resilient water systems; 2023. Available from: https://wwweiborg/en/publications/20230016-eib-water-sector-orientation.
2. Liemberger R, Marin P. The challenge of reducing non-revenue water in developing countries–how the private sector can help: A look at performance-based service contracting. 2006;.

3. Ortigara AR, Kay M, Uhlenbrook S. A review of the SDG 6 synthesis report 2018 from an education, training, and research perspective. Water (Basel). 2018;10(10):1353. Available from: https://doi.org/10.3390/w10101353.
4. Johnston R. Arsenic and the 2030 Agenda for sustainable development. Arsen Res Glob Sustain-Proc 6th Int Congr Arsen Environ AS; 2016.
5. Mutikanga HE, Sharma SK, Vairavamoorthy K. Multi-criteria decision analysis: a strategic planning tool for water loss management. Water Resources Management. 2011;25(14):3947–69. Available from: https://doi.org/10.1007/s11269-011-9896-9.
6. Alegre H, Baptista JM, Cabrera E, Cubillo F, Duarte P, Hirner W, et al.. Performance indicators for water supply services: IWA publishing; 2016.
7. American Water Works Association. Best Practice in Water Loss Control: Improved Concepts for 21st Century Water Management. Denver, CO, USA: American Water Works Association; 2003; 2003. Available from: https://www.awwa.org/Portals/0/AWWA/ETS/Resources/WLCFlyerFinal.pdf?ver=2015-02-10-083650-287.
8. Lai CH, Chan NW, Roy R. Understanding public perception of and participation in non-revenue water management in Malaysia to support urban water policy. Water (Basel). 2017;9(1):26. Available from: https://doi.org/10.3390/w9010026.
9. Kizilöz B. Prediction model for the leakage rate in a water distribution system. Water Science and Technology: Water Supply. 2021;21(8):4481–92. Available from: https://doi.org/10.2166/ws.2021.194.
10. Wu J, Ma D, Wang W. Leakage identification in water distribution networks based on XGBoost algorithm. Journal of Water Resources Planning and Management. 2022;148(3). Available from: https://doi.org/10.1061/(ASCE)WR.1943-5452.0001523.
11. GM K, TE M, PA O. Automation and Optimization of Pipeline Leak Detection Using ANN-XGBoost Algorithms and PLC Integration. In: 2025 The 16th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT); 2025.
12. Lučin I, Lučin B, Čarija Z, Sikirica A. Data-driven leak localization in urban water distribution networks using big data for random forest classifier. Mathematics. 2021;9(6):672. Available from: https://doi.org/10.3390/math9060672.
13. Ma J, Yang J, Hao X, Cui B, Yang S. Dynamic Simulation Model of Channel Leakage Based on Multiple Regression. Sustainability (Basel). 2023;15(20):14904. Available from: https://doi.org/10.3390/su152014904.
14. GM K, TE M, PA O. Optimizing leak detection and location in water distribution networks using SVM-RF algorithm. In: 2024 15th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT); 2024: IEEE.
15. Candelieri A, Soldi D, Conti D, Archetti F. Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. The icewater approach. Procedia Engineering. 2014;89:1080–8. Available from: https://doi.org/10.1016/j.proeng.2014.11.228.
16. Soldevila A, Sanz G, García R. Leak localization in water networks: A data-driven approach using XGBoost. Procedia Engineering. 2019;186:282–9.
17. Coelho JA, Glória A, Sebastião P. Precise water leak detection using machine learning and real-time sensor data. IoT. 2020;1(2):474–93. Available from: https://doi.org/10.3390/iot1020026.
18. Guo G, Yu X, Liu S, Ma Z, Wu Y, Xu X, et al. Leakage detection in water distribution systems based on timeconvolutional neural network. Journal of Water Resources Planning and Management. 2021;147(2). Available from: https://doi.org/10.1061/(ASCE)WR.1943-5452.0001317.
19. Liu H, Zhang J, An W, Chen Y, Yuan X, Zhang G, et al. Operational efficiency improvement in a water supply network: machine learning-enhanced leakage identification and water resource conservation. Journal of Water Process Engineering. 2025;75. Available from: https://doi.org/10.1016/j.jwpe.2025.107924.
20. Tuan NH, An TD, Ngoc TA, Linh HD. Forecasting Leakage Potential in Water Distribution Networks Using Machine Learning Techniques: A Case Study of the Trung An Water Supply System – Ho Chi Minh City. Journal of Water Resources and Environmental Engineering. 2022;78:44–52.
21. Dung HM, Quoc TT. Assessing the current situation and proposing economic technical solutions to reduce water loss at Tan hoa water supply Joint Stock company (belong to SAWACO). VN J Of Hydro-Meteor. 2025;773:50–64.
22. People's Committee of Ho Chi Minh City. Explanatory Report on the Adjustment of the General Planning of Ho Chi Minh City to 2040, with a Vision to 2060; 2024:(in Vietnamese).
23. SAWACO. Report on Water Production at Joint-Stock Water Plants and Non-Revenue Water Status for the Period 2019–2024; 2024.
24. Ibarra-Vazquez G, Ramírez-Montoya MS, Terashima H. Gender prediction based on University students' complex thinking competency: an analysis from machine learning approaches. Education and Information Technologies. 2024;29(3):2721–39. Available from: https://doi.org/10.1007/s10639-023-11831-4.
25. Tausendschoen J, Stoeckl G, Radl S. Machine Learning for heat radiation modeling of bi-and polydisperse particle systems including walls. Particuology. 2023;74:119–40. Available from: https://doi.org/10.1016/j.partic.2022.05.011.
26. Mrabet HE, Moussa AA. A framework for predicting academic orientation using supervised machine learning. Journal of Ambient Intelligence and Humanized Computing. 2022;14(12):1–11.
27. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. Available from: https://doi.org/10.1023/A:1010933404324.
28. Tran DA, Tsujimura M, Ha NT, Nguyen VT, Van Binh D, Dang TD, et al. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. Ecological Indicators. 2021;127. Available from: https://doi.org/10.1016/j.ecolind.2021.107790.
29. Criminisi A, Shotton J, Konukoglu E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations trends® in computer graphics vision. 2012;7((2-3)):81–227. Available from: https://doi.org/10.1561/0600000035.
30. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001;29(5):1189–232. Available from: https://doi.org/10.1214/aos/1013203451.
31. T C, C G. Xgboost: A scalable tree boosting system.; 2016.
32. Mammone A, Turchi M, Cristianini N. Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics. 2009;1(3):283–9. Available from: https://doi.org/10.1002/wics.49.
33. Tuấn NH, Đ An T, Ngọc TA, Linh HD. Dự báo khả năng rò rỉ trên mạng lưới cấp nước bằng một số kỹ thuật học máy: Nghiên cứu điển hình cho hệ thống cấp nước Trung An-Thành phố Hồ Chí Minh. Tạp chí khoa học kỹ thuật Thủy lợi và Môi trường. 2022;78.
34. Moriasi DN, Gitau MW, Pai N, Daggupati P. Hydrologic and water quality models: performance measures and evaluation criteria. Transactions of the ASABE. 2015;58(6):1763–85. Available from: https://doi.org/10.13031/trans.58.10715.
35. Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE. 2007;50(3):885–900. Available from: https://doi.org/10.13031/2013.23153.
36. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP conference series: materials science and engineering. IOP Publishing; 2018.

37. Zaman D, Tiwari MK, Gupta AK, Sen D. A review of leakage detection strategies for pressurised pipeline in steady-state. Engineering Failure Analysis. 2020;109. Available from: https://doi.org/10.1016/j.engfailanal.2019.104264.
38. Trung An TAWACO. The roadmap for reducing water losses of SAWACO; 2022. Available from: https://capnuoctrunganvn/Main/Home/BaiViet/10705.
39. Liemberger R, Wyatt A. Quantifying the global non-revenue water problem. Water Science and Technology: Water Supply. 2019;19(3):831–7. Available from: https://doi.org/10.2166/ws.2018.129.
40. World Bank. The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries; 2006. Available from: https://documents1worldbankorg/curated/en/643741468330976864/pdf/447220BR0Box321IC10Water0PNOTE41NRWpdf.
41. Marques L, Carvalho RS, MOMd S, Malheiros TF. Benchmarking as a management tool to reduce non-revenue water. Ambiente Sociedade. 2021;24:e0025. Available from: https://doi.org/10.1590/1809-4422asoc20200025vu2021l4de.
42. Ministry of Environment and Water Resources (MEWR). Key Environment Statistic: Water Resource Management. (accessed on 18 March 2025). Available from: http://appmewrgovsg/web/Contents/contentsaspx?ContId=682.
43. Malaysian Water Association. Malaysia water industry guide 2003; 2003. (Acccessed on 22 August 2025). Available from: https://wwwmwaorgmy/water-industry-guide/.
44. Adegboye MA, Fung WK, Karnik A. Recent advances in pipeline monitoring and oil leakage detection technologies: principles and approaches. Sensors (Basel). 2019;19(11):2548. Available from: https://doi.org/10.3390/s19112548.
45. Ayadi A, Ghorbel O, BenSalah M, Abid M. A framework of monitoring water pipeline techniques based on sensors technologies. Journal of King Saud University Computer and Information Sciences. 2022;34(2):47–57. Available from: https://doi.org/10.1016/j.jksuci.2019.12.003.
46. Towards an IoT-based water quality monitoring system with brokerless pub/sub architecture. IEEE.; 2017. Available from: https://doi.org/10.1109/LANMAN.2017.7972166.
47. Singapore PUB. Managing the water distribution network with a Smart Water Grid. Smart Water. 2016;1(1):4. Available from: https://doi.org/10.1186/s40713-016-0004-4.

VNUHCM Journal of Environment and Earth Sciences 2026, 10(2):1387-1400
Open Access Full Text Article

Bài nghiên cứu

Dự báo tỷ lệ lượng nước thất thoát bằng mô hình học máy

Trần Thị Kim[1,2,*], Huỳnh Phước Sang[3], Nguyễn Văn Tín[3]

TÓM TẮT
Tình trạng thiếu hụt và thất thoát tài nguyên nước là một trong những vấn đề nghiêm trọng tại các khu vực đô thị trên toàn thế giới, với khoảng 32 tỷ mét khối nước sạch đã qua xử lý bị thất thoát mỗi năm do rò rỉ từ các hệ thống phân phối nước đô thị. Việc dự báo tỷ lệ thất thoát nước (Non-Revenue Water – NRW) trong tương lai đóng vai trò quan trọng trong việc giảm thiểu sự suy giảm tài nguyên và lãng phí kinh tế, đồng thời góp phần thực hiện Mục tiêu Phát triển Bền vững số 6 của Liên Hợp Quốc về Nước sạch và Vệ sinh. Nghiên cứu này đánh giá khả năng ứng dụng bốn mô hình học máy gồm Random Forest Regression (RF), Extreme Gradient Boosting Regression (XGBoost), Support Vector Regression (SVR) và K-Nearest Neighbours (KNN) để tính toán và dự báo tỷ lệ NRW của bảy công ty cổ phần cấp nước trực thuộc Tổng Công ty Cấp nước Sài Gòn (SAWACO) tại Thành phố Hồ Chí Minh, Việt Nam. Bộ dữ liệu bao gồm các số liệu theo tháng trong giai đoạn 2019–2024, gồm sản lượng nước sản xuất tại các nhà máy xử lý, số liệu từ đồng hồ tổng trên mạng lưới phân phối và lượng nước tiêu thụ của khách hàng. Kết quả cho thấy mô hình RF đạt hiệu quả dự báo tỷ lệ NRW tốt nhất với $R^2 > 0,8$, RMSE < 1,5% và MAE < 1% trong quá trình huấn luyện. Hai mô hình XGBoost và SVR cũng cho kết quả khả quan với $R^2 > 0,7$, RMSE và MAE lần lượt nhỏ hơn 2% và 1,5%, mặc dù hiệu quả trên tập kiểm định chưa đạt mức như RF. Dự báo đến năm 2030 cho thấy tỷ lệ NRW có xu hướng tăng theo mùa vào tháng Hai, chủ yếu do nhu cầu sử dụng nước sinh hoạt và công nghiệp gia tăng, làm tăng áp lực lên mạng lưới đường ống cấp nước. Kết quả cũng chỉ ra rằng các công ty cấp nước tại khu vực nội thành có tỷ lệ NRW thấp hơn (12–17%) so với các công ty phục vụ khu vực ngoại thành, cho thấy sự khác biệt về hạ tầng cấp nước cũng như tác động của quá trình đô thị hóa. Những kết quả thu được cung cấp cơ sở khoa học cho chiến lược quản lý cấp nước bền vững của SAWACO và hỗ trợ lộ trình phát triển chiến lược của Thành phố Hồ Chí Minh.

Từ khoá: Học máy, Cấp nước, Thất thoát nước

Use your smartphone to scan this QR code and download this article

[1]*Trường Đại học Khoa học Tự nhiên, số 227 Nguyễn Văn Cừ, phường Chợ Quán, Thành phố Hồ Chí Minh, Việt Nam*

[2]*Đại học Quốc gia Thành phố Hồ Chí Minh, phường Đông Hòa, Thành phố Hồ Chí Minh, Việt Nam*

[3]*Trường Đại học Tài nguyên và Môi trường Thành phố Hồ Chí Minh, số 236B Lê Văn Sỹ, phường Tân Sơn Hòa, Thành phố Hồ Chí Minh, Việt Nam*

Liên hệ

Trần Thị Kim, Trường Đại học Khoa học Tự nhiên, số 227 Nguyễn Văn Cừ, phường Chợ Quán, Thành phố Hồ Chí Minh, Việt Nam

Đại học Quốc gia Thành phố Hồ Chí Minh, phường Đông Hòa, Thành phố Hồ Chí Minh, Việt Nam

Email: ttkim@hcmus.edu.vn

Lịch sử
- Ngày nhận: 10-10-2025
- Ngày sửa đổi: 04-02-2026
- Ngày chấp nhận: 04-08-2026
- Ngày đăng: 27-08-2026

DOI : https://doi.org/10.32508/vnuhcmj-ees.v10i2.849

Check for updates

Trích dẫn bài báo này: Kim T T, Sang H P, Tín N V. **Dự báo tỷ lệ lượng nước thất thoát bằng mô hình học máy**. *VNUHCM J. Environ. Earth Sci.* 2026; 10(2):1387-1400.