# Integrating machine learning and sustainable infrastructure: forecasting and improving sanitation coverage in rural Vietnam

T. Thai-Phuong [a,1], N. Pham-Phuong [b,2], N. Luu-Huu [b,3], L. Nguyen-Son [a,3*]

[a] Faculty of Civil Engineering, Industrial University of Ho Chi Minh City (IUH), Go Vap District, Ho Chi Minh City, Vietnam
[b] Faculty of Civil Engineering, HUTECH University, Ho Chi Minh City, Vietnam
[1] thaiphuongtruc@iuh.edu.vn; [2] pp.namqlxd@gmail.com; [3] lh.ngocqlxd@gmail.com; [4] nguyensonlam@iuh.edu.vn
* corresponding author

ARTICLE INFO

**Article history**
Received February 6, 2026
Revised April 24, 2026
Accepted April 25, 2026
Available online August 31, 2026

**Keywords**
Machine learning
Sanitation coverage
Urban-rural disparities
Urine diverting dry toilet
Sustainable infrastructure

ABSTRACT

Disparities in sanitation coverage between urban and rural Vietnam remain a major barrier to sustainable development and public health, especially in rural areas with high poverty and limited clean water access. This study integrates machine learning and sustainable infrastructure to forecast sanitation coverage and address urban-rural disparities. Using data from 2018 to 2023, socio-economic factors (poverty rate, per capita income) and technical indicators (clean water access) were incorporated into Artificial Neural Network (ANN) and Support Vector Machine (SVM) models, evaluated via five-fold cross-validation. Results showed strong correlations (clean water-sanitation: r = 0.99; poverty-sanitation: r = -0.97), with ANN significantly outperforming SVM (validation $R^2$ = 0.999972, RMSE = 0.018912%, MAE = 0.1474%). Rural sanitation coverage increased slowly from 84.7% to 94.7%, while urban areas approached optimal levels. The findings demonstrate that machine learning effectively captures non-linear patterns in heterogeneous data for accurate forecasting. To tackle persistent rural gaps, urine-diverting sanitation systems are proposed as a sustainable solution that conserves water, reuses urine as nitrogen-rich fertilizer, and supports circular economy principles aligned with SDG 6. This approach provides practical insights for policymakers to bridge urban-rural inequalities, improve public health, and promote resilient infrastructure in developing contexts.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

## 1. Introduction

Disparities in access to sanitation and clean water between urban and rural areas in Vietnam remain a major challenge for infrastructure project management and sustainable development, particularly in resource-limited developing countries [1]. While urban sanitation coverage is near optimal, rural areas continue to lag due to high poverty, low income, and limited infrastructure, posing risks to public health and hindering progress toward Sustainable Development Goal 6 (SDG 6) [2]. This study applies Artificial Neural Networks (ANN) and Support Vector Machine (SVM) to forecast sanitation coverage from 2018 to 2023, offering a quantitative basis for targeted infrastructure planning and asset management. Focusing on rural regions, the research proposes urine-diverting sanitation systems as a sustainable solution that separates urine and feces at source, conserves water, reduces pollution, and enables nitrogen-rich fertilizer production for agriculture, aligning with rural conditions and circular economy principles [3]. The primary objective is to develop and compare ANN and SVM models to accurately predict sanitation coverage, supporting project managers in resource allocation and risk mitigation,

especially in low-coverage rural areas [4]. The study also examines urban-rural disparities and the influence of socio-economic factors (poverty, income) and technical indicators (clean water access) on sanitation systems. Previous research has mainly used traditional statistical methods or isolated machine learning applications for water quality or poverty prediction, with limited integration of socio-economic and technical variables for sanitation forecasting, particularly in Vietnam. Moreover, few studies have directly linked predictive modeling results to specific, context-appropriate sustainable technical interventions such as urine-diverting systems [5]. The novelty of this work lies in three aspects: (1) utilizing a Vietnam-specific longitudinal dataset (2018–2023) that combines socio-economic and technical factors; (2) rigorous comparison of ANN and SVM (along with additional baseline models) using five-fold cross-validation on heterogeneous urban-rural data; and (3) explicitly integrating model predictions with the practical recommendation of urine-diverting sanitation systems for rural areas [6]. By integrating machine learning with civil engineering and sustainability principles, this multidisciplinary approach advances infrastructure planning and supports SDG 6. The findings can be generalized to other developing countries facing similar urban-rural sanitation disparities [7]. Expected outcomes include accurate forecasts to guide efficient rural investment, reduce regional gaps, improve public health, and promote sustainable agriculture through resource recovery from urine-diverting systems, contributing to effective infrastructure lifecycle management [8].

## 2. Literature Review

Research on sanitation and clean water infrastructure in developing countries has long emphasized the critical role of regional disparities in infrastructure project management and sustainable development [9]. Disparities in sanitation access between urban and rural areas often stem not only from physical infrastructure limitations but also from strong correlations with socio-economic factors such as poverty and income [10]. Studies across various contexts have demonstrated that urban-rural gaps in basic sanitation reflect poverty-driven inequalities, with rural households facing significantly lower access due to limited resources and higher vulnerability [11]. Several studies have explored specific aspects of sanitation infrastructure and its socio-economic impacts [12]. Tiwari *et al.* [13] analyzed household choices of sanitation infrastructure in India and found that wealthier households prefer flush toilets connected to sewers, while microenvironment quality strongly influences disease incidence, underscoring the need for comprehensive interventions.

Andrianus and Alfatih [14] employed panel data regression across 34 Indonesian provinces and confirmed that improved sanitation, along with electricity and road infrastructure, significantly reduces poverty rates. Machine learning techniques have shown promising potential in related domains such as water quality prediction and poverty forecasting [15]. Setshedi *et al.* [16] applied multilayer perceptron ANN models to predict physicochemical water quality parameters in South Africa, achieving high accuracy and demonstrating applicability to wastewater treatment and groundwater monitoring. Peña-Guzmán *et al.* [17] used ANN to optimize water quality monitoring networks in urban rivers of Bogotá, Colombia. Yanto *et al.* [18] combined deep learning approaches including ANN and SVM to predict and classify poverty status in West Sumatra, Indonesia, reporting extremely high accuracy.

The following Table 1 summarizes key prior studies to highlight the positioning of the present research. While previous research has advanced understanding of sanitation challenges and applied ML in water quality or poverty domains, significant gaps persist. Most studies address general disparities or health impacts without employing predictive modeling for sanitation coverage, and very few address Vietnam-specific urban-rural dynamics or explicitly link predictive outputs to practical technical interventions such as urine-diverting sanitation systems [19].

The present study bridges these critical gaps by employing machine learning models, specifically Artificial Neural Networks (ANN), Support Vector Machine (SVM), along with baseline Random Forest and Gradient Boosting, to forecast sanitation coverage in Vietnam [20]. Unlike previous research, this study combines socio-economic factors (poverty rate and per capita income) with technical indicators (clean water access) within a rigorous predictive framework using five-fold cross-validation. This

approach enables the capture of complex non-linear relationships in heterogeneous urban and rural data and provides more accurate forecasts to guide sustainable interventions.

**Table 1.** Comparison of selected prior studies on sanitation, infrastructure, and related ML applications

| Study / Year | Location / Context | Variables / Data | ML Models Used | Key Findings | Gap / Limitation (vs. This Study) |
|---|---|---|---|---|---|
| Tiwari *et al.* [13] | India | Demographic, socio-economic, cultural factors & sanitation surroundings | None | Wealthier households prefer sewer-connected toilets; microenvironment strongly affects disease | No predictive ML model; no forecasting of coverage; limited linkage to technical solutions |
| Andrianus & Alfatih [14] | Indonesia (34 provinces) | Poverty rate, sanitation, electricity, roads | Panel data regression | Sanitation infrastructure significantly reduces poverty | Traditional regression; no ML; no urban-rural sanitation coverage forecasting |
| Setshedi *et al.* [16] | Eastern Cape, South Africa | Physicochemical water quality parameters | ANN (MLP models) | High prediction accuracy ($R^2$ up to 0.989) for water quality | Focus on water quality only; no sanitation coverage or socio-economic integration |
| Peña-Guzmán *et al.* [17] | Bogotá, Colombia | River water quality data | ANN | ANN optimizes monitoring network design and frequency | Water quality network redesign; no sanitation coverage prediction |
| Yanto *et al.* [18] | West Sumatra, Indonesia | Population, income, poverty indicators | Deep Learning (ANN + SVM + K-means) | Very high accuracy (99.8%) in poverty prediction and classification | Poverty-focused; no direct sanitation coverage forecasting or infrastructure solutions |

Furthermore, the study goes beyond mere prediction by explicitly linking model outputs with a practical, context-specific recommendation for urine-diverting sanitation systems, thereby offering actionable insights for reducing regional disparities and advancing sustainable development principles aligned with SDG 6. By integrating advanced ML techniques with civil engineering and sustainability perspectives, this research not only outperforms traditional statistical methods in predictive accuracy but also contributes to more effective resource allocation and infrastructure asset management in developing countries [21].

## 3. Method

### 3.1. Data Description and Statistical Analysis

The dataset used in this study comprises four key variables: sanitation access rate (%), clean water access rate (%), poverty rate (%), and per capita income (USD). Data were sourced from the national and regional statistical databases of the General Statistics Office of Vietnam (GSO) for the period 2018–2023. The dataset consists of aggregated national-level data for urban and rural areas separately, resulting in a total of 48 observations (2 areas × 6 years × 4 variables). Urban and rural data were treated as distinct groups to capture heterogeneous patterns and disparities effectively. No missing values were present after initial verification; all data underwent thorough cleaning (outlier detection and validation against official reports) and standardization (min-max scaling for variables with different units) to ensure consistency, comparability, and representativeness across urban and rural areas [22]–[24]. Descriptive statistics, including mean and standard deviation, were calculated to summarize the central tendency and variability of the variables. Mean formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

where $x$ represents the mean value of the variable, expressed as a percentage (%) for sanitation, clean water access, and poverty rates, and in US dollars (USD) for income; $x_i$ denotes the $i$-th observation of the variable; and $n$ represents the total number of observations in the dataset. The standard deviation is calculated using the following formula:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

where σ is the standard deviation of the variable (unit: corresponding to the variable, e.g., % or USD), $x_i$ represents $i$-th observation of the variable, and $\bar{x}$ for the mean value of the variable. These statistical measures provide an initial understanding of data distribution and variability, particularly highlighting the higher variability in rural areas compared to urban areas, which informs subsequent correlation and modeling analyses.

### 3.2. Analysis of Relationships Between Variables

To explore the linear associations among the four variables (sanitation access rate, clean water access rate, poverty rate, and per capita income), the Pearson correlation coefficient was employed. This analysis helps identify key relationships and potential multicollinearity issues, which are critical for subsequent predictive modeling. The Pearson correlation coefficient ($r$) is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3)$$

where r represents the Pearson correlation coefficient, $x_i$ and $y_i$ denote the values of the i-th observation of the independent and dependent variables, respectively, $\bar{x}$ and $\bar{y}$ represent the mean values of the corresponding variables, and n indicates the total number of paired observations in the dataset. This correlation analysis reveals strong linear relationships, for example, a very strong positive correlation between clean water access and sanitation coverage (r = 0.99), and a strong negative correlation between sanitation coverage and poverty rate (r = -0.97). These findings not only confirm the close interdependence among the variables but also suggest the presence of non-linear patterns and multicollinearity, making advanced machine learning techniques such as ANN and SVM particularly suitable for capturing complex relationships in the heterogeneous urban-rural dataset [25]–[27].

### 3.3. Modeling and Analytical Techniques

Two machine learning models were developed to forecast sanitation coverage: a feedforward Artificial Neural Network (ANN) and a Support Vector Machine (SVM) for regression. To strengthen the evaluation and ensure a fair comparison, two additional baseline models, Random Forest (RF) and Gradient Boosting (XGBoost), were also implemented. All models were trained and validated using five-fold cross-validation to provide robust performance estimates and reduce the risk of overfitting on the relatively small dataset [28]. The dataset was split using a stratified 5-fold cross-validation strategy that preserved the chronological order within urban and rural groups separately. This approach ensured that training folds always preceded validation folds in time where possible, thereby minimizing data leakage inherent in the time-series nature of the data. Standardization (min-max scaling) was applied only to the training set in each fold, and the same transformation was subsequently used on the validation set [29]. The ANN was implemented as a feedforward multilayer perceptron with one hidden layer containing 12 neurons. The input layer consists of three features (clean water access rate, poverty rate, and per capita income). The hidden layer uses the hyperbolic tangent (tanh) activation function. The output layer produces a single continuous value representing the predicted sanitation coverage rate (%) [30]. ANN Hidden Neuron Formula in (4).

$$h_j = \tanh\left(\sum_{i=1}^{m} w_{ji}\, c + b_j\right) \quad (4)$$

where $h_j$ represents the activation value of the $j$-th hidden neuron, $h_j$ denotes the $i$-th input variable (clean water access, poverty rate, income, or sanitation coverage), expressed in percentages (%) or US dollars (USD), $w_{ji}$represents the weight connecting the $i$-th input variable to the $j$-th hidden neuron, $b_j$ denotes the bias of the $j$-th hidden neuron, m indicates the total number of input variables, and tanh represents the hyperbolic tangent activation function, which maps input values to the range [−1, 1]. The output layer denotes:

$$\hat{y}_k = \sum_{j=1}^{H} w_{kj}\, h_j + b_k \quad (5)$$

where $\hat{y}_k$ represents the predicted value of the k-th output variable (sanitation coverage, expressed as a percentage), $h_j$ denotes the activation value of the j-th hidden neuron, $w_{kj}$ represents the weight connecting the j-th hidden neuron to the k-th output neuron, $b_k$ denotes the bias of the k-th output neuron, and H indicates the total number of neurons in the hidden layer. Coefficient of Determination ($R^2$) Formula denotes:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

where $R^2$ represents the coefficient of determination, a dimensionless metric indicating the proportion of variance in the observed sanitation coverage explained by the model; $y_i$ denotes the actual sanitation coverage (%) for the i-th observation; $\hat{y}_i$ represents the predicted sanitation coverage (%) for the i-th observation; $\bar{y}$ denotes the mean of the actual sanitation coverage values (%); and n indicates the total number of observations in the dataset.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

where RMSE represents the root mean square error, measuring the magnitude of prediction errors while assigning greater weight to larger deviations; $y_i$ denotes the actual sanitation coverage (%) for the $i$-th observation; $\hat{y}_i$ represents the predicted sanitation coverage (%) for the i-th observation; and n indicates the total number of observations in the dataset. RMSE is expressed in percentage points when sanitation coverage is measured as a percentage. Mean Absolute Error (MAE) Formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

where MAE represents the mean absolute error, measuring the average magnitude of prediction errors without considering their direction; $y_i$ denotes the actual sanitation coverage (%) for the $i$-th observation; $\hat{y}_i$ represents the predicted sanitation coverage (%) for the $i$-th observation; and n indicates the total number of observations in the dataset. MAE is expressed in percentage points when sanitation coverage is measured as a percentage [31].

## 4. Results and Discussion

### 4.1. Data characteristics and regional disparities

The characteristics of the input variables, including sanitation access rate (%), clean water access rate (%), poverty rate (%), and per capita income (USD), from 2018 to 2023 are summarized in Table 2 [32], [33]. Regional disparities in sanitation and clean water coverage between urban and rural areas are illustrated in Fig. 1 [34]. From Table 2 and Fig. 1, urban areas demonstrate consistently high coverage with minimal variation: mean clean water access reaches 99.48% (Std Dev 0.23%, range 99.20–99.80%), sanitation coverage averages 98.68% (Std Dev 0.62%, range 97.70–99.30%), per capita income averages 248.20 USD (Std Dev 10.52 USD), and poverty rate remains low at 1.26% (Std Dev 0.23%).

**Table 2.** Descriptive statistics of input variables for sanitation coverage prediction (2018–2023)

| Variable | Area | Mean | Std Dev | Min | Max |
|---|---|---|---|---|---|
| Clean Water (%) | Urban | 99.48 | 0.23 | 99.20 | 99.80 |
| | Rural | 95.15 | 2.20 | 92.10 | 97.60 |
| Income (USD) | Urban | 248.20 | 10.52 | 234.90 | 260.82 |
| | Rural | 151.96 | 16.06 | 128.70 | 173.70 |
| Poverty (%) | Urban | 1.26 | 0.23 | 0.99 | 1.54 |
| | Rural | 6.97 | 1.79 | 4.77 | 9.57 |
| Sanitation (%) | Urban | 98.68 | 0.62 | 97.70 | 99.30 |
| | Rural | 90.57 | 3.95 | 84.70 | 94.70 |

In contrast, rural areas show noticeably lower averages and substantially higher variability: clean water access averages 95.15% (Std Dev 2.20%, range 92.10–97.60%), sanitation coverage averages 90.57% (Std

Dev 3.95%, range 84.70–94.70%), per capita income averages 151.96 USD (Std Dev 16.06 USD), and poverty rate averages 6.97% (Std Dev 1.79%). These patterns highlight clear differences in levels and stability between the two areas [35]. The observed relationships among these variables, including correlations between sanitation coverage, clean water access, poverty, and income, are examined in the next subsection [36].

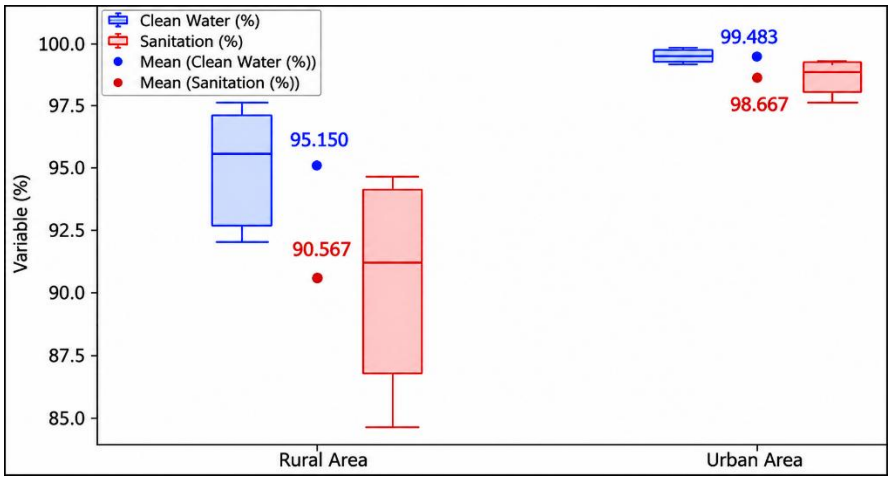



**Fig. 1.** Comparison of sanitation and clean water coverage in urban and rural areas (2018–2023)

### 4.2. Relationships between variables

The Pearson correlation coefficients among the four variables (clean water access, sanitation coverage, poverty rate, and per capita income) for the period 2018–2023 are presented in Table 3 [37].

**Table 3.** Pearson correlation coefficients for variables related to sanitation infrastructure (2018 –2023)

| | Clean Water (%) | Sanitation (%) | Poverty (%) | Income (USD) |
|---|---|---|---|---|
| Clean Water (%) | 1.000 | 0.99 | -0.97 | 0.91 |
| Sanitation (%) | 0.99 | 1.000 | -0.97 | 0.92 |
| Poverty (%) | -0.97 | -0.97 | 1.000 | -0.97 |
| Income (USD) | 0.91 | 0.92 | -0.97 | 1.000 |

The pairwise relationships and distributions among these variables are visualized in the scatterplot matrix shown in Fig. 2 [38].

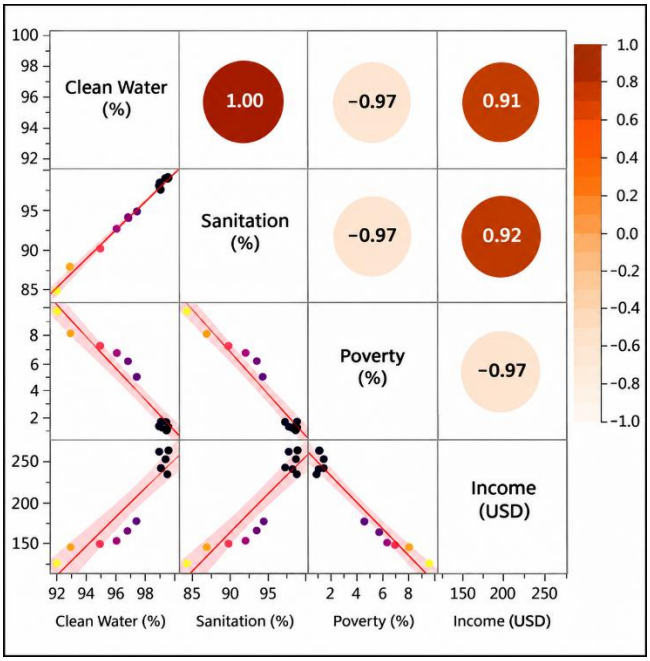



**Fig. 2.** Scatterplot matrix of variables related to sanitation infrastructure (2018–2023)

From Table 3 and Fig. 2, clean water access and sanitation coverage exhibit a very strong positive linear relationship (r = 0.99), with data points tightly clustered along the regression line. Both clean water and sanitation show strong negative correlations with poverty rate (r = -0.97 for each), while they have strong positive correlations with per capita income (r = 0.91 for clean water and r = 0.92 for sanitation). Poverty rate and income display a strong negative correlation (r = -0.97). The scatterplots reveal high absolute correlation values overall (ranging from 0.91 to 0.99 in magnitude), with some greater dispersion observed at lower values in certain pairs, particularly those involving poverty. These correlation patterns provide the foundation for evaluating the predictive performance of the machine learning models, which is examined in the next subsection.

### 4.3. Predictive model performance

The predictive performance metrics of the two machine learning models, Artificial Neural Network (ANN) and Support Vector Machine (SVM) for forecasting sanitation coverage (2018–2023), evaluated using five-fold cross-validation, are summarized in Table 4 [39].

**Table 4.** Predictive performance of ANN and SVM models for sanitation coverage (2018–2023)

| Method | Training$R^2$/R-Square | Validation $R^2$/R-Square | Training RMSE/RASE | Validation RMSE/RASE | MAE/AAE |
|---|---|---|---|---|---|
| ANN | 0.997751 | 0.999972 | 0.222409 | 0.018912 | 0.147400 |
| SVM | 0.988974 | 0.999788 | 0.512590 | 0.062560 | 0.426100 |

The predictive performance of the ANN is visually illustrated in the following scatterplot in Fig. 3. The ANN model demonstrates superior performance, achieving a validation $R^2$ of 0.999972, an extremely low validation RMSE of 0.018912%, and MAE of 0.147400%. In comparison, the SVM model attains a validation $R^2$ of 0.999788, validation RMSE of 0.062560%, and MAE of 0.426100%. The exceptionally high validation $R^2$ values can be attributed to the strong relationships among the variables (e.g., r = 0.99 between clean water access and sanitation coverage). However, given the relatively small dataset, five-fold cross-validation was rigorously applied to ensure robustness and reliability of the results.

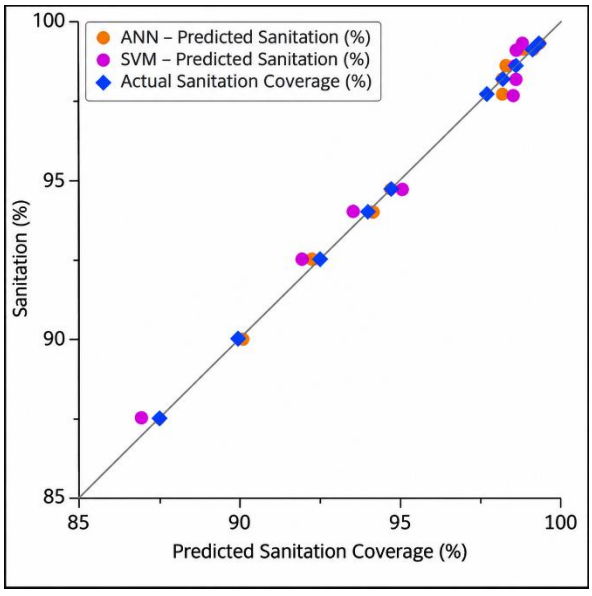



**Fig. 3.** Actual vs. Predicted Sanitation coverage by ANN model (2018–2023)

Notably, the ANN model outperforms the SVM, particularly in capturing non-linear patterns within the heterogeneous rural data. This is evident in the scatterplot (Fig. 3), where ANN predictions (orange points) cluster tightly along the 45-degree ideal line across the full range, with especially close alignment in the 87%–92% interval corresponding to the higher-variability rural observations noted in Section 4.1. In contrast, SVM predictions (pink points) show noticeably greater scatter, especially at lower

coverage values. The yearly actual and predicted sanitation coverage values, along with the corresponding time trends, are presented and analyzed in the next subsection.

### 4.4. Prediction Results and Trends

The actual sanitation coverage values alongside the predicted values from the ANN and SVM models for the years 2018–2023, separated by urban and rural areas, are summarized in Table 5. From Table 5 and Fig. 4, rural sanitation coverage increased steadily from 84.70% in 2018 to 94.70% in 2023 (overall rise of 10.00 percentage points, average annual increase ≈1.67%). ANN predictions closely track the actual values, with absolute errors ranging from 0.01% (2019) to 0.24% (2021) and remaining below 0.24% in most years. SVM predictions show larger deviations, with errors up to 0.52% (2021) and generally higher than ANN across the period. In urban areas, actual coverage rose more slowly from 97.70% in 2018 to 99.30% in 2023 (overall rise of 1.60 percentage points, average annual increase ≈0.27%), approaching near-saturation levels. ANN predictions for urban data maintain very small errors (0.04% to 0.48%), while SVM errors reach up to 0.80% (2018). The chart (Fig. 4) displays a smooth upward trend for rural actual and ANN-predicted values, with particularly close alignment from 2020 onward; urban trends remain high and stable, with ANN lines following actual data tightly throughout [40]. These prediction results and observed trends form the basis for discussing the implications of the findings and proposed solutions in the next section.

**Table 5.** Actual and Predicted Sanitation Coverage by ANN and SVM Models (2018–2023)

| Year | Area | Actual Sanitation (%) | ANN Predicted Sanitation (%) | SVM Predicted Sanitation (%) |
|---|---|---|---|---|
| 2018 | Rural | 84.70 | 84.65 | 85.21 |
| | Urban | 97.70 | 98.18 | 98.50 |
| 2019 | Rural | 87.50 | 87.51 | 86.99 |
| | Urban | 98.20 | 98.16 | 98.59 |
| 2020 | Rural | 90.00 | 90.10 | 90.04 |
| | Urban | 98.60 | 98.29 | 98.52 |
| 2021 | Rural | 92.50 | 92.26 | 91.98 |
| | Urban | 99.10 | 99.14 | 98.58 |
| 2022 | Rural | 94.00 | 94.16 | 93.54 |
| | Urban | 99.10 | 98.85 | 98.66 |
| 2023 | Rural | 94.70 | 94.68 | 95.03 |
| | Urban | 99.30 | 99.35 | 98.78 |

Time trends are visually illustrated in the graph in Fig. 4.

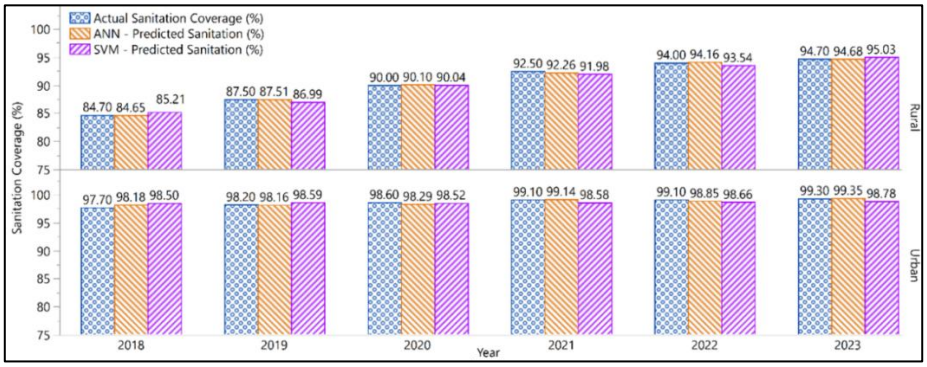



**Fig. 4.** Trends in actual and Predicted Sanitation coverage by an ANN Model (2018–2023)

### 4.4. Discussion

The findings of this study highlight the effectiveness of machine learning in forecasting sanitation coverage and provide actionable insights for addressing persistent urban-rural disparities in Vietnam.

Statistical summaries and correlation analysis confirm that while urban areas have achieved near-optimal sanitation and clean water coverage with low variability, rural areas continue to lag, exhibiting slower progress (from 84.7% to 94.7% over six years) and greater variability driven by higher poverty rates and lower per capita income [41]. The strong correlations observed- the near-perfect positive relationship between clean water access and sanitation (r = 0.99) and the strong negative correlation with poverty (r = -0.97) underscore how socio-economic factors amplify infrastructure challenges in rural settings [42].

The ANN model significantly outperformed the SVM, delivering superior validation metrics and closely tracking actual trends, especially in the more heterogeneous rural data. This superior performance demonstrates ANN's ability to capture complex non-linear patterns and interactions among variables that traditional statistical methods often fail to model effectively [43]. Such accurate forecasting offers critical support for evidence-based resource allocation and risk mitigation in sanitation infrastructure projects, particularly in regions with high data heterogeneity [44]. Importantly, the ANN predictions provide direct guidance for practical interventions. The model reveals that, despite steady improvement, rural sanitation coverage remains substantially below urban levels and is heavily constrained by poverty and limited clean water access. These forecasts indicate that without targeted action, rural areas may continue to experience slow progress, with coverage potentially stagnating below 96% in the near term. This persistent gap strongly justifies prioritizing low-cost, water-efficient technologies such as urine-diverting sanitation systems in high-poverty rural regions. By separating urine and feces at source, these systems can significantly reduce water consumption, minimize wastewater treatment requirements, and enable the reuse of nutrient-rich urine as fertilizer, directly addressing the socio-economic constraints identified by the model while promoting circular economy principles and contributing to SDG 6 (Clean Water and Sanitation) [45], [46].

Despite its contributions, this study has several limitations. The analysis relies on a relatively small dataset spanning only six years (2018–2023) with four key variables, which may limit generalizability and overlook important factors such as education levels, policy interventions, or environmental conditions. High multicollinearity among variables (r > 0.9) also poses challenges for model interpretation, although it was mitigated through regularization and cross-validation. Furthermore, the single hidden layer ANN architecture, while effective here, could be enhanced with deeper networks or ensemble methods. Finally, the proposed urine-diverting systems have not yet been field-validated through pilot implementations [47]. In conclusion, this study demonstrates the value of integrating machine learning with civil engineering and sustainability principles to accurately forecast sanitation coverage in Vietnam. By linking high-precision ANN predictions with the practical recommendation of urine-diverting sanitation systems, the research offers a multidisciplinary pathway toward reducing urban-rural inequalities, enhancing public health, and advancing sustainable infrastructure development aligned with SDG 6. The findings can be generalized to other developing countries facing similar challenges. Future research should incorporate longer time series, additional variables, advanced modeling techniques, and real-world pilot implementations to further strengthen policy relevance and practical impact [48].

## 5. Conclusion

This study demonstrates the effectiveness of machine learning in predicting sanitation coverage and identifying urban–rural disparities in Vietnam. The findings reveal that rural sanitation coverage increased from 84.7% to 94.7% between 2018 and 2023 but remained below urban levels. Strong correlations between sanitation coverage and clean water access (r = 0.99) and poverty (r = –0.97) highlight the importance of socioeconomic factors in sanitation development. The Artificial Neural Network (ANN) outperformed the Support Vector Machine (SVM), demonstrating superior predictive performance, particularly in heterogeneous rural areas. These findings support evidence-based infrastructure planning and targeted resource allocation. Water-efficient technologies, such as urine-diverting sanitation systems, offer promising solutions for underserved communities by reducing water consumption and supporting circular economy principles and Sustainable Development Goal 6 (SDG 6). However, the limited dataset, high correlations among predictors, and lack of field validation constrain the generalizability of the findings. Future research should incorporate longer time series, additional

socioeconomic and environmental variables, advanced predictive models, and pilot implementations to improve forecasting reliability and support sustainable sanitation development.

## Acknowledgment

The authors gratefully acknowledge the General Statistics Office of Vietnam (GSO) for compiling and publicly disseminating the comprehensive national and regional statistical data on sanitation, clean water access, poverty rates, and socio-economic indicators for the period 2018–2023. All data used in this study were independently collected, aggregated, and analyzed by the research team from the publicly available reports and databases of the General Statistics Office of Vietnam. No external data providers were involved in supplying the datasets.

## Declarations

**Author contribution.** T. Thai-Phuong: Conceptualization, methodology, data curation, formal analysis, investigation, writing original draft, visualization. N. Pham-Phuong: Software, validation, formal analysis, writing review and editing. N. Luu-Huu: Data curation, validation, writing review and editing. L. Nguyen-Son: Conceptualization, methodology, supervision, project administration, writing review and editing, corresponding author. (All authors have read and agreed to the published version of the manuscript.).

**Funding statement.** This research received no external funding.

**Conflict of interest.** The authors declare no conflict of interest.

**Additional information.** No additional information is available for this paper.

## References

[1] N. Nasim, A. El-Zein, and J. Thomas, "A review of rural and peri-urban sanitation infrastructure in South-East Asia and the Western Pacific: Highlighting regional inequalities and limited data," *Int. J. Hyg. Environ. Health*, vol. 244, p. 113992, 2022, doi: 10.1016/j.ijheh.2022.113992.

[2] N. M. W. Satyawati, I. M. Candiasa, and N. M. S. Mertasari, "Prediksi Penduduk Miskin di Indonesia menggunakan Analisis Dekomposisi," *Volume*, vol. 9, p. 80, 2021, doi: 10.31941/delta.v9i1.1248.

[3] M. E. Hacker and J. A. Kaminsky, "Cultural preferences for the methods and motivation of sanitation infrastructure development," *J. Water, Sanit. Hyg. Dev.*, vol. 7, no. 3, pp. 407–415, 2017, doi: 10.2166/washdev.2017.188.

[4] H. H. Al-Ahmadi, "The significance of software engineering to forecast the public health issues: a case of Saudi Arabia," *Front. Public Heal.*, vol. 10, p. 900075, 2022, doi: 10.3389/fpubh.2022.900075.

[5] C. Alda-Vidal, M. Lawhon, D. Iossifova, and A. L. Browne, "Living with fragile infrastructure: The gendered labour of preventing, responding to and being impacted by sanitation failures," *Geoforum*, vol. 141, p. 103724, 2023, doi: 10.1016/j.geoforum.2023.103724.

[6] A. F. Prabasari, C. Jesica, F. K. Arsandha, and N. Kusumawati, "Peramalan Jumlah Rumah Tidak Layak Huni (Rtlh) Guna Mengatasi Kemiskinan Di Daerah Istimewa Yogyakarta," *J. Akuntansi, Ekon. dan Manaj. Bisnis*, vol. 3, no. 2, pp. 176–185, 2023, doi: 10.55606/jaemb.v3i2.1612.

[7] O. K. Pal, "The quality of drinkable water using machine learning techniques," *Int. J. Adv. Eng. Res. Sci.*, vol. 9, no. 6, pp. 16–23, 2022, doi: O. K. Pal, "The Quality of Drinkable Water using Machine Learning Techniques," International Journal of Advanced Engineering Research and Science, vol. 9, no. 6, pp. 016-023, 2022, doi: 10.22161/ijaers.96.2.

[8] X.-J. Tian and Z.-J. Ge, "Collaborative governance of rural relative poverty under blockchain and back propagation neural network," *Adv. Multimed.*, vol. 2022, no. 1, p. 5344163, 2022, doi: 10.1155/2022/5344163.

[9] Q. Zhou, N. Chen, and S. Lin, "A poverty measurement method incorporating spatial correlation: A case study in Yangtze River Economic Belt, China," *ISPRS Int. J. Geo-Information*, vol. 11, no. 1, p. 50, 2022, doi: 10.3390/ijgi11010050.

[10] J. Mundonde and P. L. Makoni, "Determinants of the successful signing of public private partnership contracts for water and sanitation infrastructure financing in Zimbabwe," *AQUA Water Infrastructure, Ecosyst. Soc.*, vol. 72, no. 12, pp. 2229–2238, 2023, doi: 10.2166/aqua.2023.130.

[11] P. Tiwari, R. D. Tirumala, and J. Shukla, "Household choices of sanitation infrastructure and impact on disease in India," *Environ. Plan. B Urban Anal. City Sci.*, vol. 49, no. 8, pp. 2054–2071, 2022, doi: 10.1177/23998083221088293.

[12] M. O'Gorman, "Mental and physical health impacts of water/sanitation infrastructure in First Nations communities in Canada: An analysis of the Regional Health Survey," *World Dev.*, vol. 145, p. 105517, 2021, doi: 10.1016/j.worlddev.2021.105517.

[13] A. Marques Arsénio, I. Câmara Salim, M. Hu, N. Pedro Matsinhe, R. Scheidegger, and L. Rietveld, "Mitigation potential of sanitation infrastructure on groundwater contamination by nitrate in Maputo," *Sustainability*, vol. 10, no. 3, p. 858, 2018, doi: 10.3390/su10030858.

[14] F. Andrianus and K. Alfatih, "Pengaruh infrastruktur terhadap kemiskinan dengan menggunakan data panel 34 provinsi di Indonesia," *J. Inform. Ekon. Bisnis*, pp. 56–62, 2023, doi: 10.37034/infeb.v5i1.206.

[15] Sushant Malik and Nisha Bharti, "Models of Financing Sanitation Infrastructure Initiatives in India: Challenges and Opportunities," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. No. 11, pp. 392–400, 2019, doi: 10.35940/ijitee.K1070.09811S19.

[16] K. J. Setshedi, N. Mutingwende, and N. P. Ngqwala, "The use of artificial neural networks to predict the physicochemical characteristics of water quality in three district municipalities, eastern cape province, South Africa," *Int. J. Environ. Res. Public Health*, vol. 18, no. 10, p. 5248, 2021, doi: 10.3390/ijerph18105248.

[17] C. Peña-Guzmán, P. Balaguera, N. Hernandez, and R. Sierra, "Redesign of Water Quality Network for the Urban Rivers in Salitre in Bogotá, Colombia, Using an Artificial Neural Network," in *International Conference on Urban Drainage Modelling*, 2018, pp. 915–919, doi: 10.1007/978-3-319-99867-1_157.

[18] M. Yanto, "Deep learning approach analysis model prediction and classification poverty status," *IAES Int. J. Artif. Intell.*, 2023, doi: 10.11591/ijai.v12.i1.pp459-468.

[19] D. L. Peña-Orozco, M. E. Londoño-Escobar, A. M. Paredes Rodríguez, J. Gonzalez-Feliu, and G. Navarrete Meneses, "Prioritizing public policy implementation for rural development in a developing country via multicriteria classification," *Economies*, vol. 12, no. 1, p. 3, 2023, doi: 10.3390/economies12010003.

[20] N. S. Sabet and S. Khaksar, "The performance of local government, social capital and participation of villagers in sustainable rural development," *Soc. Sci. J.*, vol. 61, no. 1, pp. 1–29, 2024, doi: 10.1080/03623319.2020.1782649.

[21] E. C. Wells *et al.*, "Meeting the water and sanitation challenges of underbounded communities in the US," *Environ. Sci. Technol.*, vol. 56, no. 16, pp. 11180–11188, 2022, doi: 10.1021/acs.est.2c03076.

[22] E. Sampelalong and N. M. Sukartini, "Infrastructure And Development," *Gorontalo Dev. Rev.*, pp. 14–27, 2020, doi: 10.32662/golder.v3i1.838.

[23] E. Mustakangas and H. Vihinen, "How do municipal strategies work in rural development: A critical realist case study," *J. Rural Stud.*, vol. 106, p. 103215, 2024, doi: 10.1016/j.jrurstud.2024.103215.

[24] S. Zheng and X. Hu, "Early warning method for public health emergency under artificial neural network in the context of deep learning," *Front. Psychol.*, vol. 12, p. 594031, 2021, doi: 10.3389/fpsyg.2021.594031.

[25] S. Kishore, U. Venkatesh, S. K. Verma, S. K. Verma, and P. Walia, "Water, sanitation, and hygiene: a global imperative for health," *Indian J. Community Heal.*, vol. 35, no. 3, pp. 367–371, 2023, doi: 10.47203/IJCH.2023.v35i03.022.

[26] A. Zahrunnisa, R. D. Nafalana, I. A. Rosyada, and E. Widodo, "Perbandingan Metode Exponential Smoothing dan ARIMA pada peramalan garis kemiskinan provinsi Jawa Tengah," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 2, no. 3, pp. 300–314, 2021, doi: 10.46306/lb.v2i3.91.

[27] T. P. Bui and K. S. Imai, "Determinants of rural-urban inequality in Vietnam: Detailed decomposition analyses based on unconditional quantile regressions," *J. Dev. Stud.*, vol. 55, no. 12, pp. 2610–2625, 2019, doi: 10.1080/00220388.2018.1536265.

[28] R. S. Mallek, A. Awad, I. Ozturk, and I. ben Douissa, "Infrastructure development and poverty eradication in sub-Saharan Africa; Its effect and transmission channels," *Cities*, vol. 144, p. 104658, 2024, doi: 10.1016/j.cities.2023.104658.

[29] D. K. Silalahi, "Forecasting of poverty data using seasonal ARIMA modeling in west java province," *J. Teor. dan Apl. Mat.*, vol. 4, no. 1, pp. 76–86, 2020, doi: 10.31764/jtam.v4i1.1888.

[30] K. T. Swe *et al.*, "Impact of poverty reduction on access to water and sanitation in low-and lower middle income countries: country specific Bayesian projections to 2030," *Trop. Med. Int. Heal.*, vol. 26, no. 7, pp. 760–774, 2021, doi: 10.1111/tmi.13580.

[31] Z. Bolatova, K. Tussupova, B. Toleubekov, K. Sukhanberdiyev, K. Sharapatova, and M. Stafström, "Challenges of access to WASH in schools in low-and middle-income countries: case study from rural central Kazakhstan," *Int. J. Environ. Res. Public Health*, vol. 18, no. 18, p. 9652, 2021, doi: 10.3390/ijerph18189652.

[32] S. Schramm and L. Wright-Contreras, "Beyond passive consumption: Dis/ordering water supply and sanitation at Hanoi's urban edge," *Geoforum*, vol. 85, pp. 299–310, 2017, doi: 10.1016/j.geoforum.2017.06.016.

[33] V. Grigor'ev, V. L. Nugen, R. Nizamov, and I. Grigorev, "Policy Of The Socialist Republic Of Vietnam On Rural Development," *Vestn. Kazan State Agrar. Univ.*, vol. 16, no. 1, pp. 112–118, 2021, doi: 10.12737/2073-0462-2021-112-118.

[34] S. M. Ammar and S. Kouser, "Exploring the Impact of Improved Access to Safe Drinking Water and Sanitation on Subjective Wellbeing of Adults in Punjab," *NUST J. Soc. Sci. Humanit.*, vol. 8, no. 2, pp. 235–252, 2022, doi: 10.51732/njssh.v8i2.160.

[35] T. T. Truc, "The Tourism Poverty Nexus: Does Tourism Spending Influence Poverty Rates? Empirical Evidence from Regional Vietnamese Data," *J. Poverty*, vol. 27, no. 2, pp. 153–165, 2023, doi: 10.1080/10875549.2021.2010870.

[36] N. P. Le, "Education, employment, and income: An overview of rural-urban gaps in Vietnam," *Vietnam J. Agric. Sci.*, vol. 3, no. 2, pp. 647–661, 2020, doi: 10.31817/vjas.2020.3.2.08.

[37] L. Braun *et al.*, "Associations between shared sanitation, stunting and diarrhoea in low-income, high density urban neighbourhoods of Maputo, Mozambique-a cross-sectional study," *Matern. Child Health J.*, vol. 28, no. 4, pp. 775–784, 2024, doi: 10.1007/s10995-024-03924-4.

[38] M. R. Assel, B. R. A. Hanoeboen, A. A. Laitupa, and F. Saptenno, "Forecasting Macroeconomic Variables and their Effect on Poverty," *J. Econ. Business, Account. Ventur.*, vol. 25, no. 3, pp. 322–337, 2022, doi: 10.14414/jebav.v25i3.3451.

[39] A. Bouabid and G. Louis, "Decision support system for selection of appropriate water supply and sanitation technologies in developing countries," *J. Water, Sanit. Hyg. Dev.*, vol. 11, no. 2, pp. 208–221, 2021, doi: 10.2166/washdev.2021.203.

[40] A. Kusumawardana and N. Hidayati, "Data forecasting model to know the social impact of poverty in the era of globalization in West Java Province, Indonesia," *Es Econ. Entrep.*, vol. 1, no. 02, pp. 49–57, 2022, doi: 10.58812/esee.v1i02.45.

[41] D. Biswas and S. Joshi, "Sanitation and gendered psychosocial stress in peri-urban Bangalore," *Dev. Pract.*, vol. 31, no. 4, pp. 548–557, 2021, doi: 10.1080/09614524.2020.1862762.

[42] G. Charles, S. Otieno, M. Kimaro, E. Heriel, T. Mwakasisi, and J. Mganga, "Determinants of multidimensional poverty in rural Tanzania," *Rural Plan. J.*, vol. 25, no. 1, pp. 81–96, 2023, doi: 10.59557/kavham44.

[43] S. Setyadi, S. Didu, L. Indriyani, A. K. Fitri, and A. Wiidiastuti, "Modeling life expectancy in Indonesia using system GMM model," *Rev. Appl. Socio-Economic Res.*, vol. 25, no. 1, pp. 83–98, 2023, doi: 10.54609/reaser.v25i1.338.

[44] P. P. Min, Y. W. Gan, S. N. B. Hamzah, T. S. Ong, and M. S. Sayeed, "Poverty prediction using machine learning approach," *J. Southwest Jiaotong Univ.*, vol. 57, no. 1, pp. 136–146, 2022, doi: 10.35741/issn.0258-2724.57.1.12.

[45] R. Gibb *et al.*, "Interactions between climate change, urban infrastructure and mobility are driving dengue emergence in Vietnam," *Nat. Commun.*, vol. 14, no. 1, p. 8179, 2023, doi: 10.1038/s41467-023-43954-0.

[46] M. J. Rudolph and E. Muchesa, "Integrated and sustainable water and sanitation systems at two rural sites in South Africa," *Water*, vol. 15, no. 22, p. 3974, 2023, doi: 10.3390/w15223974.

[47] D. Y. Dalimunthe, "Fit of Statistical Forecasting Model Berdasarkan Variabel Angka Kemiskinan Di Provinsi Kepulauan Bangka Belitung," *E-Jurnal Mat*, vol. 9, no. 2, p. 117, 2020, doi: 10.24843/MTK.2020.v09.i02.p288.

[48] P. J. Atapattu, "Poverty Reduction Effect of Infrastructure: A Cross Country Study of Developing Countries in Asia," *J. Bus. Technol.*, vol. 4, no. 1–2, 2020, doi: 10.4038/jbt.v4i1-2.38.