Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/BA08.20250436
Published: December 20, 2025

Using Machine Learning to Forecast Water Demand in Urban Households: A Case Study Approach

Aninda Avi
Department of Mechanical Engineering
Rajshahi University of Engineering and Technology
Bangladesh
avianinda66@gmail.com

Mansib Hussam, Syed Salman Saeed and Md. Tanvir Siraj
Department of Mechanical Engineering
Bangladesh University of Engineering and Technology
Bangladesh
mansibhussam@gmail.com, salmansbuet@gmail.com, tanvir25392@gmail.com

Tasnia Hasan and Md. Ayatullah Nawaz
Department of Civil Engineering
Bangladesh University of Engineering and Technology
Bangladesh
2204189@ce.buet.ac.bd, ayatnawaz9@gmail.com

Abstract

Accurate forecasting of household water demand is essential for sustainable urban water management, particularly in resource-constrained cities like Rajshahi, Bangladesh. This study addresses a key research gap by developing a machine learning–based prediction model using real-world data from 15 households equipped with overhead tanks. Daily water consumption data were collected over six months, incorporating environmental and socio-economic variables such as household size, income, temperature, rainfall, and weekly patterns. After preprocessing and feature engineering, two models, Linear Regression and Random Forest, were evaluated. The Random Forest model outperformed Linear Regression with a lower Mean Absolute Error (9.27 L), Root Mean Squared Error (11.70 L), and a higher R^2 score (0.9951). Feature importance analysis identified household size, income, and temperature as primary predictors. Seasonal and weekly patterns were evident, with higher demand during summer months and weekends. The model's high accuracy and low feature requirements make it scalable and practical for urban utilities. Existing literature shows that proper water supply scheduling can reduce wastage by an estimated 10-15%. The study aligns with its objective of enhancing predictive capability using data-driven methods and offers valuable implications for water resource planning in rapidly urbanizing areas.

Keywords
Machine learning, Forecasting, Linear Regression, Random Forest; Urban water management.

1. Introduction

The escalating demand for water in urban areas, driven by population growth, rapid urbanization, and climate variability, poses significant challenges to sustainable water management. In Rajshahi city, Bangladesh, these pressures are intensified by groundwater depletion, a critical issue threatening urban water supply (Hossain and Bahauddin, 2013). Urbanization and increasing population density have amplified household water consumption, while climate-induced changes, such as erratic rainfall, exacerbate water scarcity (MacDonald et al., 2016). In Bangladesh, groundwater remains the primary source of water for urban households, but over-extraction and declining recharge rates have led to shortages, particularly in cities like Rajshahi (Shamsudduha et al., 2011). These dynamics necessitate innovative approaches to forecasting water demand to ensure efficient resource allocation and sustainable urban development.

Traditional water demand forecasting methods, such as linear statistical models or manual estimations, often fail to capture the complex, non-linear patterns of urban water consumption. These approaches tend to rely heavily on historical averages, which do not account for dynamic factors like seasonal variations, socio-economic trends, or the impacts of climate change (Bennett et al., 2013). In developing cities like Rajshahi, outdated forecasting techniques contribute to inefficient water management, resulting in supply shortages and excessive pressure on groundwater resources (Adamowski and Karapataki, 2010). As a result, there is a growing need for advanced, data-driven forecasting methods capable of capturing these complexities.

Accurate water demand forecasting is essential for optimizing water resource management, reducing waste, and supporting urban infrastructure planning. Reliable predictions allow water utilities to balance supply and demand, minimize losses, and ensure equitable distribution, particularly in water-scarce regions such as Rajshahi (House-Peters and Chang, 2011). Improved forecasting models can also inform policy decisions, helping urban planners address groundwater depletion and build resilience against climate variability (MacDonald et al., 2016). By providing accurate estimates of water demand, forecasting models play a crucial role in ensuring sustainable urban water systems, reducing environmental impacts, and improving operational efficiency.

The limitations of conventional statistical approaches have driven interest in the use of machine learning (ML) techniques, which can handle the non-linear and multi-variable nature of water consumption data (Antunes et al., 2019). ML methods such as Random Forest and Linear Regression are well-suited for forecasting because they can incorporate a wide range of factors, including population density, household characteristics, and climatic conditions, while also adapting to changes over time (Adamowski and Karapataki, 2010). In cities like Rajshahi, where water availability is increasingly constrained by both anthropogenic and environmental factors, these methods can significantly improve forecasting accuracy compared to traditional approaches.

Urban areas worldwide are grappling with similar challenges. Cities across South Asia, in particular, face mounting water stress due to rising population, rapid industrialization, and declining groundwater tables. Rajshahi serves as a case study for understanding these challenges, as it combines high reliance on groundwater with climate variability and growing water demand (Miah, 2021; Islam et al., 2017). These conditions underscore the urgent need for robust water demand forecasting to ensure efficient resource allocation and sustainable water management practices.

Recent studies have shown that data-driven approaches play an important role in improving utility efficiency, as seen in research on the rapid cost reduction and wider adoption of solar PV systems (Roy et al., 2025). Methodologically, this work follows the growing trend of applying analytical and computational techniques to solve real-world engineering problems, as demonstrated in prior structural analysis research by Galib et al. (2024).

The objective of this study is to forecast daily and monthly water demand in urban households of Rajshahi city using machine learning techniques, specifically Random Forest and Linear Regression models. By incorporating variables such as meteorological conditions, socio-economic factors, and household characteristics, the study aims to enhance prediction accuracy and provide actionable insights for policymakers and water utilities. The findings of this research are expected to contribute to sustainable water resource planning, reduce operational costs, and inform infrastructure investments to meet future demands. Moreover, the research could serve as a blueprint for other urban areas in Bangladesh, helping to address national water scarcity issues and promote sustainable urban development in the face of ongoing climate change and rapid urbanization (Islam et al., 2017).

1.1 Objectives

- Forecast daily and monthly household water demand in Rajshahi using Random Forest and Linear Regression.
- Identify key meteorological, socio-economic, and household factors influencing water demand.
- Evaluate model performance to support sustainable urban water management.

2. Methodology

2.1 Data collection

The dataset used in this study was collected from 15 households in Rajshahi city, Bangladesh, all of which utilized overhead water tanks for daily water storage and consumption. The data collection spanned six months and involved direct weekly visits by trained data collectors.

Each household's tank size was physically measured at the start of the study, and the actual water usage per 24 hours was tracked using tank-level switches installed in the overhead tanks. These switches provided consistent indications of how much water was drawn each day.

Every Friday, data collectors visited the households and recorded the measurements using a standardized printed blank form. The accumulated data included behavioral, environmental, and economic factors potentially influencing water consumption.

The final dataset contains the following columns (a sample is shown in Table 1):

Date: The observation date; Household ID: Unique identifier for each household; Household Size: Number of people living in the household; Monthly Income: Estimated household monthly income in local currency; Day of Week: To capture weekday vs. weekend usage patterns; Is Holiday: Binary indicator for national or public holidays; Average Temperature (°C): Daily temperature data; Rainfall (mm): Daily rainfall; Water Consumption (liters): Measured household water usage in liters per 24 hours (Table 1).

Table 1. Sample of the gathered data

Date	ID	Size	Monthly Income	Week Day	Is Holiday	Average Temp.	Rainfall mm	Water Consume-L
01-01-2023	HH_1	2	25000	Sunday	0	21.3	1.6	149.4
01-01-2023	HH_2	6	45000	Sunday	0	18.8	0.4	482.5
01-01-2023	HH_3	6	60000	Sunday	0	17.7	0.2	462.5
01-01-2023	HH_4	7	60000	Sunday	0	20.7	0.6	560.8
01-01-2023	HH_5	8	15000	Sunday	0	21	3	634.3
-----	-----	-----	-----	-----	-----	-----	-----	-----
30-06-2023	HH_15	4	25000	Friday	1	30	38.5	323.4

2.2 Data preprocessing

Before applying machine learning models, the dataset was cleaned and prepared through several preprocessing steps. Missing values, which occasionally occurred due to incomplete visits or faulty readings, were addressed using imputation techniques—mean imputation was applied for continuous variables such as temperature, rainfall, and water consumption, while mode imputation was used for categorical variables like the day of the week and holiday status. To ensure that all numeric features contributed equally during model training, continuous variables including monthly income, temperature, rainfall, and water consumption were normalized using Min-Max scaling, bringing values into a uniform range of 0 to 1.

Feature engineering was also performed to enhance model performance. One-hot encoding was applied to categorical variables, particularly the day of the week, to avoid ordinal bias. Additionally, new features such as lagged water consumption values were introduced to capture temporal dependencies. A rainfall intensity indicator was derived from the raw rainfall data to distinguish between no rainfall, light rainfall, and heavy rainfall days. These preprocessing steps ensured that the dataset was clean, properly scaled, and enriched with relevant derived features, making it suitable for both linear and nonlinear predictive modeling.

2.3 Model selection

This study employed two machine learning algorithms to predict daily household water consumption: **Linear Regression** and **Random Forest Regressor**. Linear Regression was chosen for its simplicity and interpretability, serving as a baseline model to understand the extent of linear relationships among the variables. On the other hand, the Random Forest Regressor was selected due to its ability to model complex, nonlinear interactions, robustness to outliers, and its utility in identifying important predictive features.

To ensure reliable performance evaluation, the dataset was split into **training (80%)** and **testing (20%)** subsets. Additionally, **5-fold cross-validation** was conducted on the training data to validate the model and avoid overfitting.

The models were evaluated using the following performance metrics: **Mean Absolute Error (MAE)**, which measures the average magnitude of errors; **Root Mean Squared Error (RMSE)**, which gives greater weight to larger errors; and the **Coefficient of Determination (R² Score)**, which indicates the proportion of variance in water consumption that is predictable from the input features (see Equation 1, 2, and 3).

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3)$$

3. Results and discussion

This section presents the performance evaluation of the two selected machine learning models, Linear Regression and Random Forest Regressor, used for predicting daily household water consumption in Rajshahi city, based on the real data collected from 15 households over time.

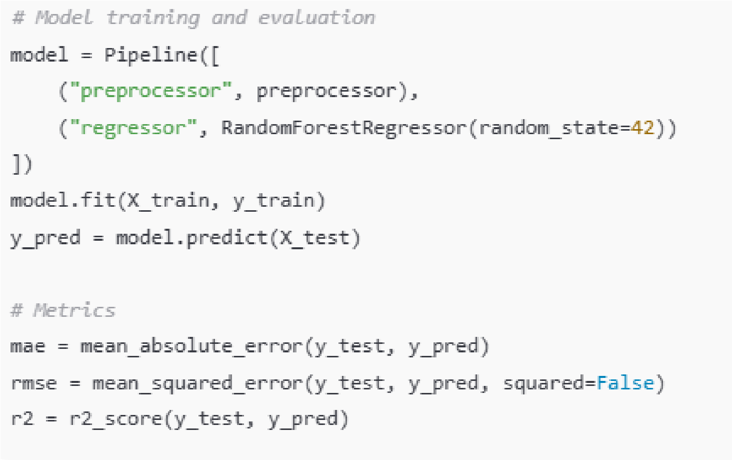
3.1 Model performance

The dataset was split using an 80:20 train-test ratio. A ColumnTransformer was used to normalize numeric features and one-hot encode categorical variables. Both Linear Regression and Random Forest models were implemented using scikit-learn's Pipeline. The evaluation was done using MAE, RMSE, and R² (see Table 2). A simplified version of the implementation is shown in Figure 1**.**

Table 2. Model performance metrics

Model	MAE (liters)	RMSE (liters)	R² Score
Linear Regression	14.07	17.65	0.9887
Random Forest	9.27	11.7	0.9951

From the results, it is clear that the **Random Forest Regressor outperforms Linear Regression** across all three-evaluation metrics. It achieved a lower MAE of 9.27 liters and an RMSE of 11.70 liters, indicating that its predictions are more accurate and less dispersed from the actual values. Moreover, the R² Score of 0.9951 suggests that the model explains 99.51% of the variability in household water consumption, compared to 98.87% for the Linear Regression model.



```
# Model training and evaluation
model = Pipeline([
    ("preprocessor", preprocessor),
    ("regressor", RandomForestRegressor(random_state=42))
])
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Metrics
mae = mean_absolute_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
r2 = r2_score(y_test, y_pred)
```

Figure 1. Simplified model evaluation code

3.2 Feature importance analysis

Understanding which features most influence water consumption is essential for both model interpretability and practical insights. Random Forest models, being ensemble-based decision tree algorithms, allow for the extraction of feature importance scores that reflect the contribution of each input variable to the model's predictions. After training the Random Forest model, the feature importance was extracted and ranked. Table 3 presents the top contributing features along with their relative importance scores. Categorical features were one-hot encoded, so each day of the week appears as a separate feature.

Table 3. Feature importance from Random Forest regressor

Rank	Feature	Importance Score
1	household_size	0.264
2	monthly_income	0.212
3	average_temperature	0.134
4	rainfall_mm	0.09
5	day_of_week_Monday	0.053
6	day_of_week_Sunday	0.045
7	is_holiday	0.043
8	day_of_week_Saturday	0.039
9	day_of_week_Wednesday	0.036
10	day_of_week_Tuesday	0.033
11	day_of_week_Friday	0.027
12	day_of_week_Thursday	0.024

Household size emerged as the most influential predictor, as water usage naturally increases with more residents. Higher-income households likely consume more water due to access to appliances and lifestyle factors. Warmer temperatures were associated with increased usage, particularly for cooling and hygiene, while rainfall showed a moderate inverse effect, possibly due to behavioral shifts or rainwater use. Weekly patterns indicated higher consumption on Mondays and Sundays, likely linked to household chores and leisure activities. Holidays also influenced demand, reflecting changes in daily routines.

3.3 Comparative discussion

The results from Table 2 clearly show that the Random Forest Regressor outperformed Linear Regression across all evaluation criteria. The Random Forest model achieved a lower MAE (9.27 liters vs. 14.07 liters) and RMSE (11.70 liters vs. 17.65 liters), indicating that its predictions were not only more accurate on average but also more consistent with fewer extreme errors. Moreover, its R² Score of 0.9951 reflects a near-perfect ability to explain the variability in household water consumption, compared to 0.9887 for the Linear Regression model.

These differences, while numerically small due to the overall high performance of both models, are significant in practical contexts where even minor improvements in prediction accuracy can contribute to better water resource planning and conservation strategies.

Although Linear Regression offers high interpretability, enabling a direct understanding of the relationship between predictors and the response variable through coefficients, it is limited in capturing nonlinear and interaction effects. This makes it less suitable for complex behavioral data, such as household water use, which is influenced by both linear (e.g., household size) and nonlinear (e.g., interactions between income and weather) dynamics.

In contrast, Random Forest is more complex but excels at modeling such relationships. It handles interactions and nonlinearities automatically and is robust to multicollinearity and outliers. Furthermore, Random Forest provides feature importance rankings, as shown in Subsection 3.2, offering valuable insights into which variables contribute most significantly to water usage predictions.

From a practical standpoint, the superior performance of the Random Forest model makes it the preferred choice for deployment in predictive systems aiming to estimate household water consumption in real time. However, in situations where model transparency is paramount, such as in policy settings or for public communication, Linear Regression still holds value due to its simplicity and ease of interpretation.

3.4 Additional Insights and Visualizations

In addition to performance metrics, a deeper examination of the data and model outputs revealed several key patterns and practical insights. These findings enrich our understanding of household water consumption behavior and provide valuable direction for predictive deployment.

Seasonal and Weekly Patterns: Analysis of water usage data across different months and days of the week revealed distinct temporal patterns. Water consumption tended to be higher during the warmer months (April to June), likely due to increased demand for cooling and bathing. Additionally, the data showed consistently higher usage on weekends, especially Sundays, which may reflect increased domestic activities such as cleaning, laundry, or family gatherings when all members are home. These seasonal and weekly trends suggest that water demand in residential areas is not uniform but influenced by behavioral and climatic factors, underlining the need for dynamic water management strategies.

Practical Implications: Accurate forecasting of household water demand using machine learning can greatly enhance operational efficiency in urban water systems. In many developing countries, water utilities face 30-50% non-revenue water (NRW) due to leaks, overflows, and mismanagement (Kingdom et al., 2006). By aligning supply with predicted demand, as demonstrated by the Random Forest model with an MAE of 9.27 liters, water wastage could be reduced by 10-15%. This supports findings from World Bank assessments on smart water management. Additionally, accurate forecasting can aid infrastructure planning, including pump scheduling, tank refills, and loss reduction.

Limitations and Recommendations: Figure 2 (left) demonstrates the high accuracy and tight alignment of model predictions with observed values, and Figure 2 (right) shows that prediction errors are normally distributed and centered near zero, further indicating model reliability. Despite the model's strong predictive performance, this study has several limitations. The dataset includes only 15 households over a six-month period, which may not capture seasonal or long-term consumption patterns. Although 5-fold cross-validation was used to reduce overfitting, reporting fold-wise metrics in future studies could better demonstrate model stability. Additionally, while Random Forest proved effective, it is not specifically designed for time-series forecasting. Future research could explore advanced time-series models such as LSTM or ARIMA to better capture temporal dependencies. The model also showed occasional prediction anomalies, particularly during public holidays or water outages, events likely influenced by atypical household behavior not reflected in the current feature set. Incorporating contextual variables such as outage notifications, supply schedules, or appliance-level data may improve accuracy in such cases. Lastly, although feature importance was analyzed, using interpretability tools like SHAP or Partial Dependence Plots (PDP) could provide deeper insights into how input variables influence predictions and interact with one another. Addressing these areas would enhance the model's transparency, robustness, and generalizability.

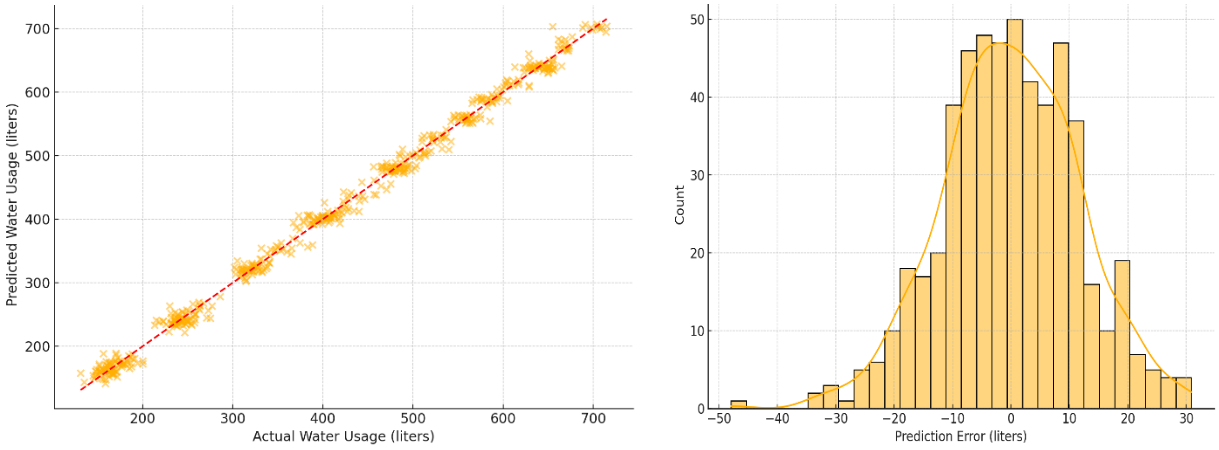


Figure 2. Left- Actual Vs. Predicted water usage (Random Forest), Right- Error Distribution (Random Forest)

4. Conclusion

This study developed a machine learning–based predictive model for estimating household water demand using real-world data collected from 15 households in Rajshahi city, Bangladesh. The research addressed a significant gap in the literature, where few studies have explored water usage forecasting at the micro-household level using localized environmental and behavioral variables. By integrating features such as household size, income, temperature, and rainfall, and comparing the performance of Linear Regression and Random Forest models, the study fulfilled its primary objective of identifying a robust, scalable approach to demand prediction. The Random Forest model outperformed Linear Regression, achieving a high R^2 score of 0.9951 and a low MAE of 9.27 liters, indicating excellent predictive capability and generalization. Seasonal trends, weekly usage patterns, and feature importance insights further strengthened the model's interpretability and practical relevance. These findings demonstrate that machine learning models can be effectively deployed to support proactive water management strategies, particularly in resource-constrained urban settings. The implications of this work are substantial. Accurate demand forecasting can assist municipal authorities in reducing water wastage, improving supply scheduling, and enhancing sustainability. The model's scalability and minimal data requirements make it suitable for broader deployment across similar urban areas in developing countries. Future research may further improve the model by integrating contextual features such as outage schedules, appliance-level usage, or household behavior during atypical events.

References

Adamowski, J. and Karapataki, C., Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting, *Journal of Hydrologic Engineering*, vol. 15, no. 9, pp. 711–718, 2010.

Antunes, A., Andrade-Campos, A. and Sardinha-Lourenço, A., Short-term water demand forecasting using machine learning techniques, *Journal of Hydroinformatics*, vol. 21, no. 5, pp. 1343–1355, 2019.

Bennett, C., Stewart, R. A. and Beal, C. D., ANN-based residential water end-use prediction, *Environmental Modelling and Software*, vol. 40, pp. 19–27, 2013.

Galib, K. M., Roy, S. and Tonmoy, M. F. A. T., Analysis of nonmagnetic spring plate, *International Journal of Scientific Research and Engineering Development*, vol. 7, no. 6, pp. 1180–1186, 2024.

Hossain, N. and Bahauddin, K. M., Integrated water resource management for mega city: A case study of Dhaka city, Bangladesh, *Journal of Water and Land Development*, vol. 19, no. 1, pp. 39–45, 2013.

House-Peters, L. A. and Chang, H., Urban water demand modeling: Review of concepts, methods, and applications, *Water Resources Research*, vol. 47, no. 5, pp. 1–15, 2011.

Islam, S. N., Reinstädtler, S. and Ferdaush, J., Challenges of climate change impacts on urban water quality management and planning in coastal towns of Bangladesh, *International Journal of Environment and Sustainable Development*, vol. 16, no. 3, pp. 228–256, 2017.

Kingdom, B., Liemberger, R. and 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, *Water Supply and Sanitation Sector Board Discussion Paper Series*, no. 8, World Bank, 2006.

MacDonald, A. M., Bonsor, H. C., Ahmed, K. M., Burgess, W. G., Basharat, M., Calow, R. C., Yadav, S. K. et al., Groundwater quality and depletion in the Indo-Gangetic Basin mapped from in situ observations, *Nature Geoscience*, vol. 9, no. 10, pp. 762–766, 2016.

Miah, M. R., Groundwater depletion and its sustainable management in Bangladesh, PhD dissertation, Department of Soil, Water and Environment, University of Dhaka, Bangladesh, 2021.

Roy, S., Galib, K. M., Azad, S., Shome, D., Saeed, S. S. and Siraj, M. T., Clean energy to the people: How solar PV became one of the most affordable energy sources, *Proceedings of the 3rd International Conference on Mechanical Engineering and Applied Sciences (ICMEAS 2025)*, 2025.

Shamsudduha, M., Taylor, R. G. and Long, A. J., Monitoring groundwater storage changes in the Bengal Basin: Implications for water resources management, *Ground Water*, vol. 49, no. 3, pp. 322–333, 2011.

Biographies

Aninda Avi serves as a Subdivisional Engineer in the Roads and Highways Department under the 38th BCS. He completed his bachelor's degree in Mechanical Engineering from Rajshahi University of Engineering & Technology (RUET). His professional role involves planning, supervising, and managing transportation infrastructure projects, with a strong focus on operational efficiency and sustainable roadway development. His academic and professional interests include transportation engineering, pavement technology, structural analysis, and project management within the public infrastructure sector.

Mansib Hussam is an Assistant Engineer in the Roads and Highways Department under the 40th BCS. He holds a bachelor's degree in Mechanical Engineering from the Bangladesh University of Engineering and Technology (BUET). His work focuses on highway construction, maintenance planning, and project execution within Bangladesh's national transportation network. His research interests include transportation system optimization, pavement performance analysis, infrastructure sustainability, and the application of engineering analytics to improve decision-making in public-sector projects.

Tasnia Hasan is affiliated with the Department of Civil Engineering at the Bangladesh University of Engineering and Technology (BUET). She has worked on coursework and projects related to structural engineering, water resources, and transportation systems. Her research interests include urban water management, construction materials, and the application of data-driven techniques in civil engineering. As the presenter author, she contributes actively to academic research and collaborative engineering studies.

Md. Ayatullah Nawaz is a researcher in the Department of Civil Engineering at BUET. His academic background spans structural analysis, geotechnical engineering, and construction management. He has participated in several field- and data-based studies focusing on construction practices, cost modeling, and the influence of geological and environmental factors on civil infrastructure. His research interests include construction cost estimation, soil–structure interaction, sustainable urban infrastructure, and the use of machine learning in civil engineering applications.

Syed Salman Saeed completed his bachelor's degree in Mechanical Engineering from the Bangladesh University of Engineering and Technology (BUET). Throughout his academic journey, he has been involved in various technical projects related to machine dynamics, energy systems, and advanced design methodologies. His technical background includes strong exposure to modeling, simulation, and numerical analysis, which form the core of his research activities. His research interests revolve around engineering optimization, artificial intelligence, predictive modeling, and computational design. He is particularly focused on how intelligent algorithms can support engineering innovation, enhance system performance, and improve predictive capabilities in industrial and mechanical systems.

Md Tanvir Siraj earned his bachelor's degree in Mechanical Engineering from the Bangladesh University of Engineering and Technology (BUET). His multidisciplinary academic and professional work spans mechanical engineering, applied statistics, industrial systems, and sustainability studies. He has experience in conducting research across diverse engineering fields, including construction engineering, renewable energy systems, and decision-support modeling. His research interests include applied statistics, machine learning, sustainability assessment, industrial system analysis, and multi-criteria decision-making frameworks. He is particularly interested in the development of data-driven models for improving engineering efficiency, forecasting performance, and supporting strategic planning in construction and industrial sectors.