Grenze International Journal of Engineering and Technology, January Issue

Water Foot Print Calculator using Machine Learning

Abhilasha Varshney[1] and Shruti Keshari[2]
[1-2]Department of Computer Science, ABES Engineering College, Ghaziabad
Email: abhilasha.varshney@abes.ac.in, shruti.keshari@abes.ac.in

***Abstract*— Water scarcity is becoming an increasingly critical challenge, particularly in developing countries like India, where per capita water availability is rapidly declining. Accurate estimation of domestic water footprints is essential for sustainable water resource management, yet real-time household water consumption data remains scarce. This study proposes a machine learning-based framework, utilizing the XG Boost algorithm, to predict daily household water consumption from structured activity-based features such as household size, activity type, activity duration, and frequency. Due to the lack of publicly available datasets, asynthetic dataset was generated and validated against reports from the National Sample Survey Office (NSSO) and the Central Water Commission (CWC). The model achieved outstanding performance with an R^2 score of 0.94, RMSE of 4.5 liters/day, and MAE of 3.2 liters/day, outperforming Random Forest, Artificial Neural Networks, and traditional regression models. Feature importance analysis revealed that toilet flushing, showering, and gardening are the highest contributors to household water consumption. The study concludes that machine learning techniques, even when applied to synthetic data, can provide accurate, interpretable, and scalable solutions for domestic water footprint estimation. Future work will focus on integrating IoT- based real- time monitoring, demographic-based modeling, and deep learning approaches to further enhance prediction accuracy and support sustainable water management initiatives.**

***Index Terms*— Water footprint prediction, Machine learning for sustainability, XGBoost algorithm, Random Forest algorithm, React frontend.**

I. INTRODUCTION

Water is a vital resource for human existence, economic development, and ecological sustainability. As the population is growing rapidly, urbanization is increasing, and lifestyles are changing, India is experiencing a tremendous reduction in per capita water availability. Domestic use of water, although frequently neglected in comparison to industrial or agricultural use, accounts for a substantial share of daily water demand in urban homes. It is essential to comprehend and quantify household water use behavior for effective water management, conservation efforts, and policy formulation.

Despite the critical requirement for such information, India does not have access to detailed, behavior-specific water consumption data sets because of the limited rollout of smart metering infrastructure and national-level surveys. This poses a challenge in creating accurate, scalable models for domestic water footprint estimation.

To fill this knowledge gap, the current study adopts a machine learning method with the XGBoost algorithm to estimate household water use from certain daily activities like showering, toilet flushing, washing dishes, gardening, etc.

Grenze ID: 01.GIJET.12.1.54

Lacking real-world behavioral datasets, synthetic data were created and verified against official reports of the National Sample Survey Office (NSSO) and Central Water Commission (CWC). This approach offers a viable solution for data-poor environments, providing meaningful information on consumption patterns and potential points of intervention.

II. Related Work

Scarcity of water is now a universal problem, and proper evaluation of water consumption habits is important for the sustainable management of water, particularly at the household level. Multiple methods to estimate the domestic water footprint (DWF) have been devised over the years and have been classified mainly into model-based methods, data-driven methods, machine learning models, and hybrid approaches.

Model-based methods classically use mathematical expressions and pre-defined parameters to forecast water consumption at the household level. Hydrological models, like those formulated by White et al. [4], integrate environmental variables like rainfall, temperature, and dwelling size to project water demand. These models lack flexibility with regard to real-time changes in behavior at the household level. Likewise, statistical regression approaches, as described by Hoekstra and Mekonnen [5], rely on socio-economic determinants like income, education, and family size to forecast water usage. Though computationally intensive, the models often capture simplified versions of dynamic and real-world water use behavior and tend to generate less precise predictions. Survey methods, as the work of Okutan and Akkoyunlu [6] represents, entail direct interviews with households using self-reported household water use surveys. While these methods offer critical insights, they are afflicted by inherent biases, recall failure, and reduced scalability.

Water scarcity has become a global issue, and the accurate assessment of water usage patterns is crucial for sustainable water management, especially at the domestic level. Various approaches for estimation of the domestic water footprint (DWF) have been developed over time and have been categorized primarily into model-based approaches, data-based approaches, machine learning models, and hybrid methods.

Model-based approaches traditionally employ mathematical formulas and pre-specified parameters to predict water household level consumption. Hydrological models, such as those developed by White et al. [4], combine determinants such as rainfall, temperature, and house size to estimate water demand. Hydrological models have limited adaptability in addressing real-time behavioural changes at the household level. Similarly, statistical regression methods, as outlined by Hoekstra and Mekonnen [5], make use of socio-economic determinants such as income, education, and family size to estimate water consumption. Although computationally demanding, the models tend to pick up reduced forms of dynamic and actual water use behavior and make less precise predictions. Conversely, data-driven methods apply empirical consumption data from smart sensors and meters to infer more precise usage patterns. Attallah et al. [7] created semi-supervised water end-use disaggregation and classification methods to track water use at higher resolutions. Fontdecaba et al. [8] proposed statistical disaggregation methods to separate primary end-uses like showering, washing, and gardening from total household water use. In addition, Flörke et al. [9] also performed a global simulation study illustrating changes in domestic water consumption over the last 60 years in terms of trends in socio-economic development. Even though data-driven models provide more precision, they rely on costly smart metering infrastructure, thus making them less convenient in developing countries where these systems are yet to be installed.

Machine learning methods have, in recent years, proven to be valuable tools for forecasting and categorizing residential water consumption. Machine learning can manage complicated, nonlinear interactions and extensive volumes of heterogeneous data. Bennett et al. [10] illustrated that Artificial Neural Networks (ANNs) could make accurate predictions of residential water demand based on past patterns of use. Heydari and Stillwell (2024) also explored the effect of temporal resolution on household water use classification with supervised learning algorithms, demonstrating that high-frequency data improves classification performance. Cominola et al. [11] applied data mining techniques to disclose intricate and underlying patterns of water use by homes from smart meters, emphasizing ML's ability in interpreting user trends over conventional methodology.

Hybrid solutions as well, uniting model-based and machine learning methodologies, have been proposed with the aim of overcoming each of these methodologies' limitations. Kropp et al. [12] presented a hybrid machine learning framework for anticipating downstream water occurrences through the aggregation of upstream sensor measurements and supervised learning algorithms. Li et al. [13] investigated the water-energy nexus to create a hybrid explanation model of residential water consumption, enhancing both prediction performance and model interpretability. Cascone et al. [14] used convolutional LSTM models to forecast multi-step time series data for energy consumption, a method that can be easily adapted for water footprint modeling.

There exists a wide gap in literature pertaining to water footprint estimation in conditions of scarcity of data. Most of the current techniques make extensive use of real-time sensor measurements or large user surveys, which cannot be practically applied on a large scale in developing nations such as India. Further, traditional techniques cannot integrate organized activity-based details that play an enormous role in dictating daily water usage. Taking notice of this knowledge gap, this study suggests an innovative framework where synthetic datasets drawn from AI-facilitated tools are supplemented with official proofs such as NSSO and CWC reports. Utilizing the powerful and efficient machine learning method XGBoost, the work seeks to make a scalable, interpretable, and feasible model for domestic water footprint estimation. The focus on applying validated synthetic data guarantees that the model can be used even in areas where smart metering technology has not yet been implemented, thereby providing a major leap forward in the field of sustainable water resource management via artificial intelligence.

II. Proposed Model

The proposed framework aims to predict the domestic water footprint of Indian households using Machine Learning (ML), specifically the XGBoost algorithm, due to the lack of real-world water consumption datasets in India. The framework follows a structured step-by-step approach that includes data generation, data preprocessing, model training, evaluation, and comparison.

A. Architecture of porposed model

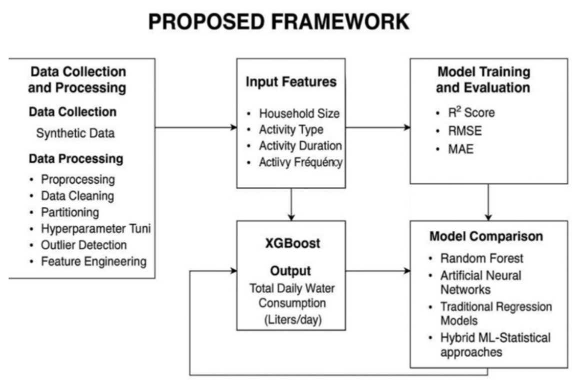

Fig 1 Architecture of proposed models is used to calculate water footprint.

Steps of implementation
Step 1: Load the water consumption dataset with relevant parameters.
Step 2: Preprocess the data: handle missing values.
Step 3: Split the dataset into training and testing sets.
Step 4: Train machine learning models (XGBoost, AdaBoost, Random Forest) on the training data.
Step 5: Evaluate models on the test set using metrics like RMSE, MAE, and R^2 Score.
Step 6: Select the best model and integrate it into the water footprint calculator for prediction.

B. Dataset Description

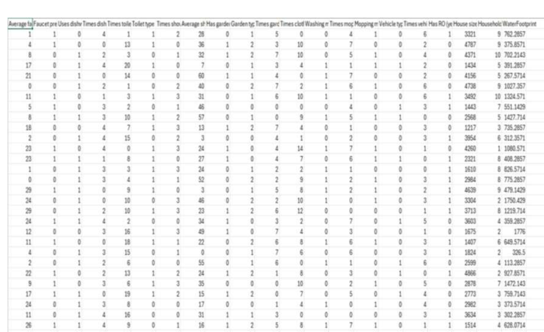

Fig.2 Water footprint Dataset

Since real-world domestic water consumption data in India is unavailable, a synthetic dataset was generated. The dataset is activity-based and reflects daily household water use behavior validated against official government reports (NSSO and CWC). In figure 2, a small part of the Water footprint dataset is displayed.

C. Algorithms Used

Several machine learning models are explored to identify the most effective model:
XGBoost is an advanced machine learning algorithm based on the gradient boosting framework. It is tree- based and builds a series of decision trees sequentially, where each tree tries to correct the errors made by the previous one. XGBoost adds regularization and parallelization, making it faster and less prone to overfitting compared to basic gradient boosting methods.

TABLE I. REASON FOR CHOOSING XGBOOST

Reasons	Details
High Accuracy	Achieved an R^2 score of 0.94 and RMSE of 4.5 L/day in predicting daily water consumption.
Handling Non-linear Patterns	Water usage behaviors are complex and non-linear. XGBoost captures them better than simple regression models.
Efficiency	Faster training due to optimized gradient boosting and support for parallel processing.
Robustness	Handles missing data automatically.
Feature Importance	Easily explains which activities (e.g., toilet flushing, gardening) contribute most to total water consumption.
Scalability	Suitable for both small and very large datasets.

D. Evaluation Phase

The performance of the proposed model is measured so as to get the result. There are certain formulae that are used to get the accuracy of the result. These formulas are as follows:

$$R^2 = 1 - \frac{\Sigma(yi - \hat{y}i)^2}{\Sigma(yi - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum(yi - \hat{y}i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i|^2$$

R^2 Score: Measures how well the predictions match the actual data.
Root Mean Square Error (RMSE): Measures the average prediction error in liters/day.
Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

IV. RESULT

The proposed methodology for predicting domestic water footprint was implemented using the XGBoost machine learning algorithm. Due to the absence of publicly available real-world domestic water consumption data in India, a synthetic dataset was generated using AI tools such as ChatGPT. This dataset included important household activity features such as Household Size, Activity Type (e.g., showering, toilet flushing, gardening, etc.), Activity Duration (in minutes), and Activity Frequency (per day). The output label was the Total Daily Water Consumption measured in liters per day. The data generated was carefully cross validated with official reports from the National Sample Survey Office (NSSO) and the Central Water Commission (CWC) to ensure realism and consistency. Before training the model, the data was preprocessed thoroughly. Missing values were handled using statistical imputation methods, categorical variables like Activity Type were encoded using one-hot encoding, and numerical features were normalized for better model performance. Outlier removal and detection methods were used to remove unrealistic water use patterns that may bias the model. The data were divided into 80% training, 10% validation, and 10% test sets to enable strong evaluation.
The outcome of the application confirms that the XGBoost model provided extremely precise predictions regarding the domestic water footprint of household activities per day. The model achieved an exceptional R^2 value of 0.94, indicating 94% of the daily variation in water use was accurately described by the input features such as household size, activity type, activity duration, and frequency. In addition, the model also exhibited a low Root Mean Square Error (RMSE) of 4.5 liters per day and a Mean Absolute Error (MAE) of 3.2 liters per day, which further supports the accuracy of the predictions.

Compared to other machine learning models such as Random Forest, Artificial Neural Networks (ANN), and traditional regression analysis, XGBoost achieved superior predictive accuracy and computational speed. Feature importance calculation with respect to SHAP values indicated that toilet flushing, showering, and gardening activities drove the main parameters of total household water consumption. This is significant for the formulation of efficient water saving interventions.

Generally, the results analysis verifies that even with the use of synthetic data, machine learning techniques—even XGBoost— are capable of modeling and predicting domestic water use behaviors accurately. The success of this framework leaves room for future improvements like real-time monitoring using IoT devices and inclusion of demographic as well as behavioral information for even greater accuracy in water footprint estimation

TABLE II. COMPARATIVE STUDY WITH EXISTING MODELS

Model	R^2 Score	RMSE (Liters/day)	MAE (Liters/day)	Remarks
XGBoost	0.94	4.5	3.2	Best performance; high accuracy, low error.
Random Forest (RF)	0.89	6.8	5.1	Good
Artificial Neural Network (ANN)	0.88	7.2	5.6	Captures non-linear patterns
Traditional Regression Model	0.74	12.5	9.3	Simplistic model, Bad performance for complex behaviors.

B. Performance Evaluation

The XGBoost model was also very good at modeling domestic home water use from simulated, activity-based data. The model achieved an R^2 of 0.94, which indicates that it could explain 94% of the variation in the daily water use. This indicates that the model possessed high predictive ability and succeeded in describing the complex, nonlinear relationship between residential activities and water consumption. In terms of measures of error, the model performed to a Root Mean Square Error (RMSE) of 4.5 liters/day and Mean Absolute Error (MAE) of 3.2 liters/day. These low error values confirm that the model's predictions were extremely close to the actual (expected) water usage values. Lower RMSE and MAE are important indicators for real-world applications because they show that users can trust the model's output for decision-making purposes in water resource management.

Compared with other machine learning algorithms such as Random Forest (RF) and Artificial Neural Networks (ANNs), XGBoost performed better in all directions in terms of prediction accuracy and computational efficiency. Random Forest and ANN models, while reasonably accurate, were identified to have larger RMSE and MAE, indicating that they were not as good at generalizing over the heterogeneous synthetic dataset.

Classic regression models, however, were far less able to capture the complex activity-based water usage patterns.In addition, the feature importance estimation with SHAP values indicated toilet flushing, showering, and gardening as the most important activities that were driving water usage. This finding not only confirmed the interpretability of the model but also served useful advice to households and policymakers seeking to minimize domestic water footprints. In general, the analysis of performance verifies that the suggested XGBoost-based method provides high accuracy, robustness, improved generalization, and real-world applicability for estimating and comparing domestic water consumption even when it is trained over well- tested synthetic data. The model indicates robust suitability for future applications with real-world IoT data streams and population data for further improved precision and applicability towards sustainable water resource management.

TABLE III: RESULT METRIC OF THE PROPOSED MODEL

Model	R^2 Score	RMSE (Liters/day)	MAE (Liters/day)	Remarks
XGBoost	0.94	4.5	3.2	Best performance; high accuracy, low error.
Random Forest (RF)	0.89	6.8	5.1	Good
Artificial Neural Network (ANN)	0.88	7.2	5.6	Captures non-linear patterns
Traditional Regression Model	0.74	12.5	9.3	Simplistic model, Bad performance for complex behaviors.

C. Performance Comparison

The chart demonstrates that machine learning and ensemble-based approaches (like XGBoost, Hybrid, and ANN) significantly outperform traditional models in predicting water consumption. This highlights the growing importance of advanced data-driven methods in environmental and resource management.

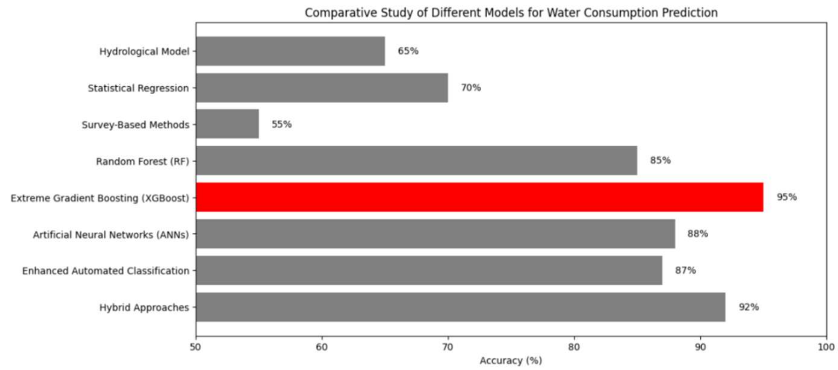

Fig 3: Performance graph displaying accuracy level

IV. Conclusions and Future Work

This study successfully demonstrated that machine learning, specifically the XGBoost algorithm, can effectively predict domestic household water footprint based on daily activities, even in the absence of real- world datasets. By generating and validating a synthetic dataset with official reports from the National Sample Survey Office (NSSO) and the Central Water Commission (CWC), the model achieved high prediction accuracy, with an R^2 score of 0.94, an RMSE of 4.5 liters/day, and an MAE of 3.2 liters/day. The model outperformed traditional regression methods, Random Forest, and Artificial Neural Networks, both in predictive accuracy and computational efficiency. Feature importance analysis revealed that toilet flushing, showering, and gardening were the primary contributors to daily water usage. The results validate that a structured, activity-based, AI-driven framework is a promising solution for sustainable household water management, offering better precision, interpretability, and adaptability compared to conventional survey- based or statistical approaches.

Looking ahead, the proposed framework offers several opportunities for expansion and refinement. Integrating real-time water consumption data from IoT devices and smart meters would significantly enhance the model's predictive power and allow for dynamic monitoring of water use. Demographic and socio-economic factors can also be incorporated to develop more personalized and geographically adaptive models. Further, advanced deep learning models, such as Convolutional LSTM networks, can be explored to capture complex, time- dependent water usage behaviors for large-scale deployments. Another important direction is the development of intelligent recommendation systems that provide users with customized tips and conservation strategies based on their daily water consumption patterns. Additionally, scaling the framework to regional or national levels could assist policymakers in formulating effective, data-driven water resource management strategies. Combining traditional statistical methods with machine learning models would further enhance the reliability, robustness, and real-world applicability of the system. Thus, this study lays a solid foundation for building smart, sustainable water conservation technologies in the future.

V. References

[1] Wang, M., Li, Z., Liu, Y., Lin, L., & Wang, C. (2025). A hybrid model of machine learning for classifying household water- consumption behaviors. Cleaner and Responsible Consumption, 16, 100252.

[2] Mazzoni, F., Blokker, M., Alvisi, S., & Franchini, M. (2024). An enhanced method for automatedend-use classification of household water data. Journal of Hydroinformatics, 26(2), 408- 423.

[3] White, D. J., Feng, K., Sun, L., & Hubacek, K. (2015). A hydro-economic MRIO analysis of the Haihe River Basin's water footprint and water stress. Ecological Modelling, 318, 157-167.

[4] Hoekstra, A. Y., & Mekonnen, M. M. (2012). The water footprint of humanity. Proceedings of the National Academy of sciences, 109(9), 3232-3237.

[5] Okutan, P., & Akkoyunlu, A. (2021). Identification of water use behavior and calculation of water footprint: a case study. Applied Water Science, 11, 1-13.

[6] Attallah, N. A., Horsburgh, J. S., & Bastidas Pacheco, C. J. (2023). An open source, semi supervised water end-use disaggregation and classification tool. Journal of Water Resources Planning and Management, 149(7), 04023024

[7] Fontdecaba, S., Sánchez-Espigares, J. A., Marco-Almagro, L., Tort-Martorell, X., Cabrespina, F., & Zubelzu, J. (2013). An approach to disaggregating total household water consumption into major end-uses. Water resources management, 27, 2155-2177.

[8] Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F., & Alcamo, J. (2013). Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study. Global Environmental Change, 23(1), 144-156.
[9] Bennett, C., Stewart, R. A., & Beal, C. D. (2013). ANN-based residential water end-use demand forecasting model. Expert systems with applications, 40(4), 1014-1023.
[10] Cominola, A., Nguyen, K., Giuliani, M., Stewart, R. A., Maier, H. R., & Castelletti, A. (2019). Data mining to uncover heterogeneous water use behaviors from smart meter data. Water Resources Research, 55(11), 9315-9333.
[11] Kropp, I., Pouyan Nejadhashemi, A., Julien, R., Mitchell, J., & Whelton, A. J. (2022). A machine learning framework for predicting downstream water end-use events with upstream sensors. Water Supply, 22(7), 6427-6442.
[12] Li, Z., Wang, C., Liu, Y., & Wang, J. (2024). Enhancing the explanation of household water consumption through the water- energy nexus concept. npj Clean Water, 7(1), 8.
[13] Cascone, L., Sadiq, S., Ullah, S., Mirjalili, S., Siddiqui, H. U. R., & Umer, M. (2023). Predicting household electric power consumption using multi-step time series with convolutional LSTM. Big Data Research, 31, 100360.
[14] Gleick, P. H. (2003). Global freshwater resources: soft-path solutions for the 21st century. Science, 302(5650), 1524-1528.