

IoT-Based Monitoring and Machine Learning Classification of Urban Water Consumption Patterns

J. M. Luna-Rivera , Carlos A. Hernández-Morales , L. E. Baez-Salazar , J. Rufo , and J. Rabadan 

Abstract—Efficient urban water management requires real-time monitoring and intelligent analytics to support both day-to-day operation and long-term resource optimization. This paper reports the design, deployment, and evaluation of an energy-aware LPWAN monitoring system implemented at a university facility in Mexico. Over five months, the system recorded hourly volumetric flow and transmitted a single daily burst, revealing clear circadian and weekly consumption patterns influenced by academic schedules and external events. To characterize consumption behavior, a percentile-based labeling strategy was employed to categorize daily water usage into four operational states: Low, Normal, High, and Anomalous. Using these labels, supervised machine learning models were evaluated for classification of consumption patterns, while unsupervised methods were used to explore natural groupings in the data. Among the tested models, Random Forest, Decision Tree, and XGBoost achieved the highest classification performance under stratified 10-fold cross-validation. From an energy-efficiency standpoint, once-per-day bursting delivered 3,816 hourly readings through 159 daily bursts (six 12-byte frames each; 954 frames total), reducing the number of daily transmissions by 75% relative to hourly uploads while preserving full hourly resolution. These results show that minimal sensing coupled with lightweight LPWAN telemetry can support scalable, pattern-driven water management and provide a practical foundation for context-aware decision support in smart urban distribution networks.

Link to graphical and video abstracts, and to code:
<https://latam.ieee9.org/index.php/transactions/article/view/10711>

Index Terms—Internet of Things, low-power wide-area networks (LPWAN), machine learning classification, smart metering, urban water management, water consumption monitoring.

I. INTRODUCTION

WATER distribution systems operate under dynamic and often irregular conditions, particularly in urban environments subject to variable demand patterns. Traditional monitoring strategies typically prioritize increasing sensing granularity through multi-sensor deployments to capture pressure, flow, and quality indicators simultaneously. While effective, such approaches increase architectural complexity, cost, and deployment barriers. At the same time, the United

Nations projects that nearly 70% of the global population will reside in urban areas by 2050 [1], intensifying challenges associated with climate change, population growth, and aging infrastructure [2], [3]. These pressures highlight the need for monitoring frameworks that enhance resilience without proportionally increasing system complexity.

Biological homeostasis offers a useful analogy. In living systems, hydration balance emerges from adaptive responses to observable signals rather than exhaustive measurement of all internal variables [4], [5]. Similarly, urban water networks may be interpreted as regulatory systems in which consumption anomalies indicate imbalance within the distribution infrastructure [6].

Inspired by this analogy, the present study evaluates whether reliable pattern recognition and anomaly detection can be achieved using a minimal sensing strategy based on a single volumetric flow sensor. Instead of increasing observability through sensor multiplicity, the proposed architecture emphasizes temporal segmentation and contextual interpretation of consumption behavior. Water usage patterns are characterized through four operational states: Low, Normal, High, and Anomalous using a percentile-based labeling strategy. By reducing sensing complexity while preserving analytical robustness, the proposed framework aims to provide a scalable and flexible monitoring solution suitable for institutional deployments and smart urban distribution environments.

II. RELATED WORK

Recent advances in IoT-based water monitoring systems have demonstrated the feasibility of large-scale telemetry using LPWAN technologies such as Sigfox and LoRaWAN, as well as local networks including Wi-Fi, across residential, campus, and urban infrastructures [7]–[11]. Many of these implementations adopt multi-sensor architectures integrating pressure, flow, and water quality indicators to enable fine-grained supervision and leak detection.

Building upon these infrastructures, artificial intelligence (AI) techniques have been increasingly applied to support forecasting, consumption classification, and anomaly detection in water systems. A comprehensive review by Morain et al [12] highlights the rapid expansion of AI-driven methodologies in this domain. Approaches including neural networks (ANN), support vector machines (SVM), ensemble learning and boosting algorithms have demonstrated strong performance in predictive modeling and behavioral segmentation [13]–[18]. Supervised methods, such as Random Forest (RF), are commonly used for consumption categorization, while clustering

The associate editor coordinating the review of this manuscript and approving it for publication was Javier Vázquez Castillo (*Corresponding author: Carlos Alberto Hernandez-Morales*).

J. M. Luna-Rivera, Carlos Alberto Hernandez-Morales, and L. E. Baez-Salazar are with the Universidad Autónoma de San Luis Potosí, San Luis Potosí, Mexico (e-mails: mlr@uaslp.mx, carlos.morales@uaslp.mx, and a210989@alumnos.uaslp.mx).

J. Rufo is with the Universidad de La Laguna, Tenerife, Spain (e-mail: jrufotor@ull.edu.es).

J. Rabadan is with the Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain (e-mail: jose.rabadan@ulpgc.es).

and unsupervised techniques such as K-Means have been employed to identify irregular usage profiles [19]–[22]. In addition, anomaly detection frameworks incorporating reinforcement learning, statistical thresholding, and explainable AI mechanisms have been proposed to improve interpretability and operational transparency [23]–[25].

On the other hand, prior studies have demonstrated the feasibility of single flow-meter approaches for domestic leak detection using statistical and ANN models. However, these works primarily focus on high-frequency consumption analysis and do not address energy-aware [26], [27].

In contrast, this work presents the design and evaluation of an energy-aware IoT-based water monitoring architecture deployed at the Faculty of Nursing and Nutrition of the Universidad Autónoma de San Luis Potosí (UASLP), Mexico, where five months of operational data were collected under real institutional constraints. The system relies on a single volumetric flow sensor as a deliberate minimal sensing configuration aligned with cost and energy limitations. To express operational water-consumption patterns, multiple supervised and unsupervised machine learning models were comparatively evaluated to classify daily consumption into Low, Normal, High, and Anomalous states. Anomaly detection was further strengthened through time-segmented analysis distinguishing distinct daily operational periods. The proposed framework demonstrates that robust consumption profiling and anomaly detection can be achieved without complex sensing infrastructures.

To further contextualize the proposed framework within the existing literature, Table I summarizes representative IoT-based water monitoring approaches. The comparison highlights differences in deployment scale, communication technology, sensing configuration, and objectives. Although many studies focus on multiple sensing sources, approaches also vary significantly in their goals, ranging from leak detection to consumption prediction or unsupervised pattern discovery.

The remainder of this paper is structured as follows: Section III describes the system architecture and data acquisition methods. Section IV outlines the dataset and preprocessing steps. Section V presents the exploratory data analysis. Section VI reports the evaluation of machine learning models. Finally, Sections VII and VIII discuss the implications of the findings and outline directions for future research.

III. IOT-BASED MONITORING SYSTEM ARCHITECTURE

The proposed system adopts a four-layer IoT architecture: Perception, Network, Middleware, and Application, to continuously monitor water consumption, transmit data over long-range communication networks, and support offline analysis. The system was designed to ensure scalability, low power consumption, and operational reliability in an urban setting. Fig. 1 illustrates the conceptual design of the system.

Perception Layer. An ultrasonic flow meter (MA408 IoT model, Kaifa) was installed at the main intake of the Faculty of Nursing and Nutrition (UASLP), upstream of the facility’s primary cistern, which is automatically refilled as needed. This placement enables tracking of total daily consumption

by measuring incoming flow only, independently of internal cistern dynamics. The meter uses ultrasonic sensing to minimize mechanical wear and is powered by an internal battery, supporting autonomous outdoor operation. Fig. 2 shows the physical installation of the sensor.

To preserve hourly granularity while minimizing radio-on time, the meter is operated in a *daily-burst* profile: the on-board MCU buffers 24 hourly snapshots and, at midnight, transmits them as a *frozen-data* burst composed of six consecutive 12-byte frames (72 bytes in total). Between bursts the transceiver remains in microamp sleep. This policy shortens on-air time and reduces uplink overhead relative to hourly uploads while preserving the full time-stamped record.

Network Layer. Communication takes place over the Sigfox LPWAN, selected for its sub-100 mW power envelope and city-wide coverage. The network infrastructure in the city of San Luis Potosí was made available through a collaboration with the Sigfox operator in Mexico, ensuring stable connectivity for the system deployment. The daily-burst policy keeps the device well below Sigfox’s 140-message limit, while shrinking the radio duty-cycle to 0.002% and preserving the full 24h chronology. This interval reflects a deliberate trade-off between data granularity, energy efficiency, and LPWAN-duty regulations, enabling multi-year battery life without sacrificing hourly resolution.

As summarised in Table II, operating in daily-burst mode reduces LPWAN transmissions from 3,816 to 954 frames over the 159-day window (75% traffic saving) while preserving hourly resolution. Under a conservative radio-energy model (50 mJ per 12-byte uplink), the RF budget drops from 53 mWh to 13.3 mWh.

Middleware Layer. Upon reception in the Sigfox backend, each burst is relayed via HTTPS to Kaifa’s cloud-hosted Advanced Metering Infrastructure (AMI) system. This service unpacks the six frames, reconstructs the 24 hourly registers and appends auxiliary fields (battery voltage, valve status) to a time-series database. Integrity checks are implemented to discard malformed packets, flag anomalies, and log statistical outliers. Centralising data handling in the AMI backend also facilitates future integration of additional Sigfox sensors (e.g., pressure or quality), while ensuring consistent formatting for both real-time dashboards and offline analysis.

Application Layer. The final layer, Kaifa’s AMI dashboard, was used primarily for offline visualization and structured export of measurement logs. Although the platform supports real-time alarms and client-oriented features, this study focuses on historical patterns. Daily, weekly, and monthly records are exported as CSV files for further processing and machine learning analysis.

IV. DATASET DESCRIPTION

The analysis is based on hourly water-consumption data collected between November 23, 2024 and April 30, 2025, covering 159 days across multiple academic periods, including regular semesters and holiday breaks. A total of 3,816 hourly records were reconstructed from 159 successfully received daily transmissions, representing diverse operational conditions within the monitored facility.

TABLE I
COMPARISON OF THE PROPOSED SYSTEM* AGAINST EXISTING WORK IN THE CURRENT STATE OF THE ART

Ref	Scale	comm.	Sensing	Task	Method
[23]	Pipeline	4G LTE	Multiple vibration sensors	Leak detection	SVM, RF, KNN, XGBoost
[27]	Residential	AMI	Multiple smart meters	Leak magnitude classification	ANN
[18]	Building	Wi-Fi	Single pressure sensor	Consumption pattern clustering	k-means, Random Forest
[11]	Urban	LoRaWAN	Multiple sensor	Water system monitoring	Three-tier IoT architecture
[26]	Residential	AMI	Multiple Smart Meters	Anomaly detection	Data analytics
[9]	Residential	Wi-Fi	Single Smart meter	Consumption pattern clustering	SVM, RF, KNN, ANN
*	Institutional building	Sigfox	Single volumetric flow sensor	Multi-level consumption classification and anomaly detection	Statistical labeling + PCA + Comparative ML

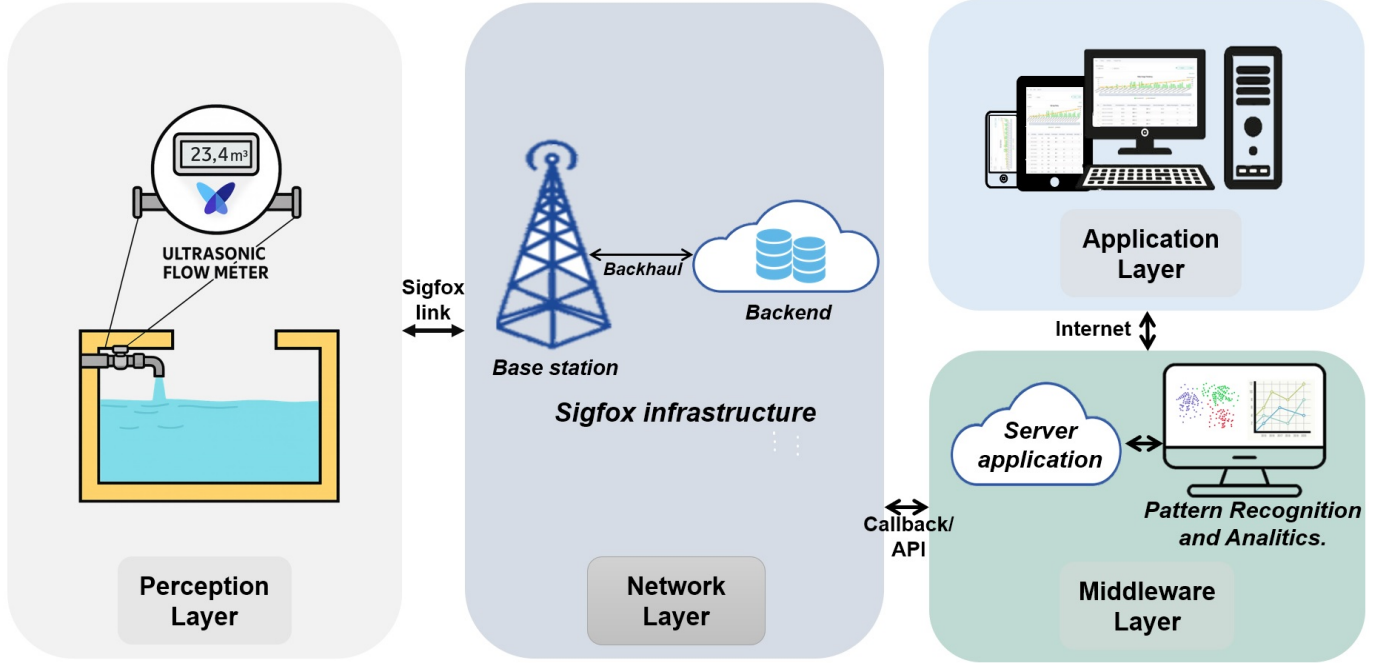


Fig. 1. Four-layer architecture for the IoT-based water monitoring system.



Fig. 2. Installed ultrasonic IoT flow meter at the facility intake.

As described in Section III, the flow meter transmitted hourly measurements through the Sigfox LPWAN using

TABLE II
LPWAN TRAFFIC AND RADIO ENERGY OVER THE 159-DAY CAMPAIGN

Policy	Frames (12 B)	Energy (Wh) [†]	Saved
Hourly upload	3 816	0.0530	—
Daily burst	954	0.0133	75.0%

a daily buffered transmission strategy. Each hourly record includes volumetric consumption (m³), running daily total, lifetime counter, valve state, battery voltage, and timestamps. Backend validation preserved data integrity: fewer than 4% of transmissions were incomplete and were subsequently recovered from the device's internal log to maintain chronological consistency.

Prior to analysis, timestamps were verified for chronological order, duplicate retransmissions were removed, and isolated missing values were flagged without imputation. Hourly consumption was aggregated into three operational shifts: Morning (06:00–15:00), Afternoon (15:00–23:00), and Night (23:00–06:00). This segmentation was inspired by circadian

analysis principles, not in a biological sense, but as an analytical framework for understanding intraday rhythms shaped by institutional routines.

The resulting dataset offers a context-aware representation of institutional water usage. It serves as a robust foundation for the classification and anomaly detection techniques explored in this study. Although the dataset is not publicly available for policy reasons, research access can be granted upon formal request to UASLP.

V. EXPLORATORY DATA ANALYSIS

To support the application of machine learning techniques for identifying consumption patterns and detecting anomalies in water usage, a comprehensive exploratory data analysis (EDA) was performed. This process was essential for understanding the temporal dynamics and operational structure of the monitored academic facility, as well as for guiding the definition of features and labels used in the subsequent supervised and unsupervised learning tasks.

A. Contextual Basis for Pattern Discovery

Effective detection and classification of water consumption patterns require understanding the operational context and supply regime. The monitored faculty operates under intermittent water delivery governed by regional distribution constraints rather than demand-responsive supply. This condition produces alternating inactivity periods and refill-driven demand spikes, introducing non-stationary behavior that challenges conventional modeling approaches and motivates context-aware analysis.

B. Statistical Summary of Daily Water Consumption

Daily consumption was computed by summing hourly values across each 24-hour window (00:00–23:59). Table III summarizes descriptive statistics for the 159-day period. The distribution exhibits high variability, with mean (3.58 m^3) comparable to standard deviation (3.65 m^3), indicating substantial dispersion. Zero-consumption days correspond to shutdowns or supply interruptions, while peak values (up to 12.96 m^3) reflect refill or exceptional operational events. The right-skewed profile and percentile thresholds ($Q1 = 0.36 \text{ m}^3$, $Q3 = 6.29 \text{ m}^3$) provide a non-parametric basis for distinguishing regular from atypical behavior and support the labeling strategy described later.

TABLE III
SUMMARY STATISTICS OF DAILY WATER CONSUMPTION
(IN m^3).

Statistic	Value (m^3)
Total Consumption	569.70
Mean Daily Consumption	3.58
Standard Deviation	3.65
Minimum Consumption	0.00
Maximum Consumption	12.96
25th Percentile (Q1)	0.36
Median (Q2)	2.21
75th Percentile (Q3)	6.29

C. Time-Series Analysis and Seasonal Trends

To visualize the temporal dynamics of water consumption, a time-series analysis was conducted using daily aggregated values. Fig. 3 presents daily usage from November 23, 2024, to April 30, 2025, with shaded areas denoting the winter and spring academic breaks. The percentile thresholds derived in the previous subsection are also included to support the visual interpretation of the Low, Normal, High, and Anomalous consumption categories.

The time series reveals non-stationary behavior shaped by academic schedules and intermittent supply conditions. Inactivity periods during academic breaks are followed by refill-driven spikes, while recurring weekday operational cycles remain visible despite supply variability. Several high-usage observations exceed typical levels and may be associated with maintenance activities or exceptional refilling operations.

Overall, the series exhibits regime shifts associated with academic closures and supply events. The correspondence between these temporal variations and the percentile thresholds provides visual support for the adopted labeling strategy. These results further reinforce the need for a context-aware approach when interpreting water consumption under intermittent supply conditions.

D. Distribution Analysis of Consumption Patterns

Fig. 4 presents the histogram of daily water consumption over the monitoring period. The distribution is strongly right-skewed, with most observations concentrated between 0.5 and 6 m^3 , representing typical operational activity. A long high-consumption tail reflects refill or maintenance events, while zero-consumption days correspond to closures or supply interruptions. This skewed profile further supports the use of non-parametric percentile thresholds for consumption categorization.

E. Circadian Shift Analysis

Building upon the shift-based aggregation defined in Section IV, daily cumulative consumption was computed for each operational shift. The dominant shift — defined as the interval with the highest consumption — was identified for every day, regardless of total daily volume, to capture operational behavior across varying activity levels. Fig. 5 shows the distribution of dominant shifts, highlighting recurring intraday demand patterns.

Key Observations: Morning consumption dominates most days, reflecting regular academic and administrative activity. Afternoon dominance appears during extended operations such as evening classes, whereas night dominance is typically associated with refill events, external supply operations, or low-activity days where nighttime usage exceeds inactive daytime periods. This shift-based representation provides a compact and operationally meaningful description of daily behavior for classification and pattern detection tasks.

To analyze interday variability, daily consumption was grouped by weekday. Fig. 6 presents boxplots illustrating the central tendencies and variability observed across the weekly

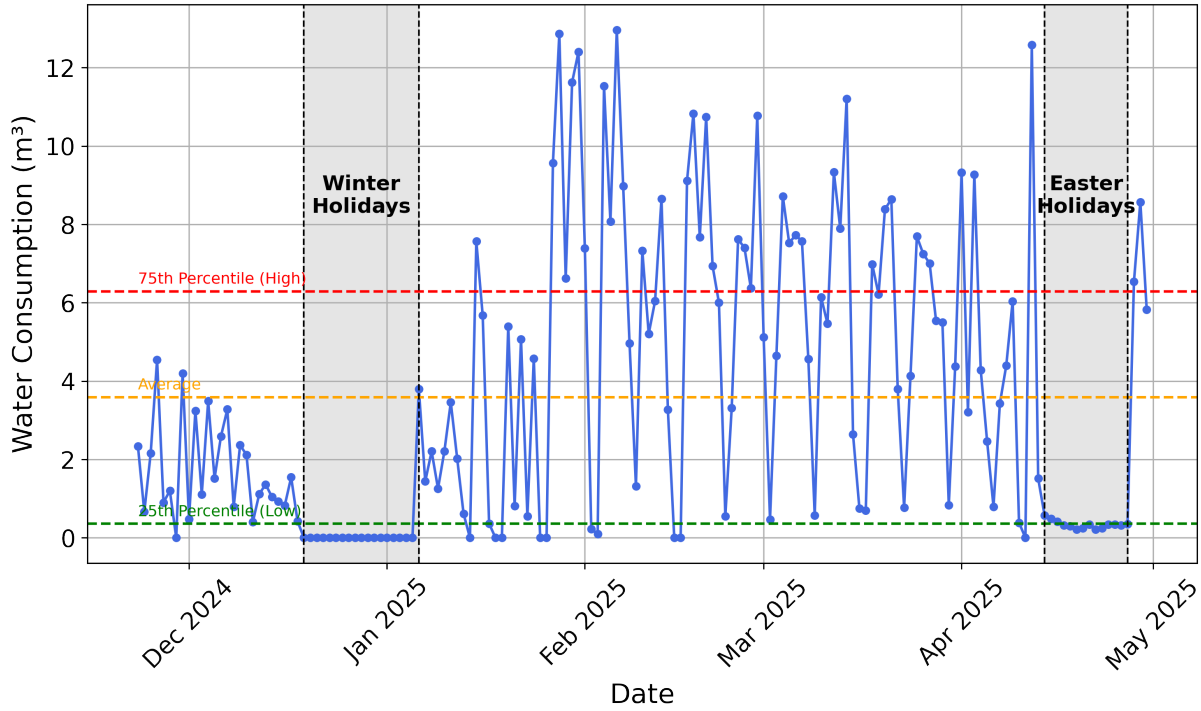


Fig. 3. Daily water consumption time series.

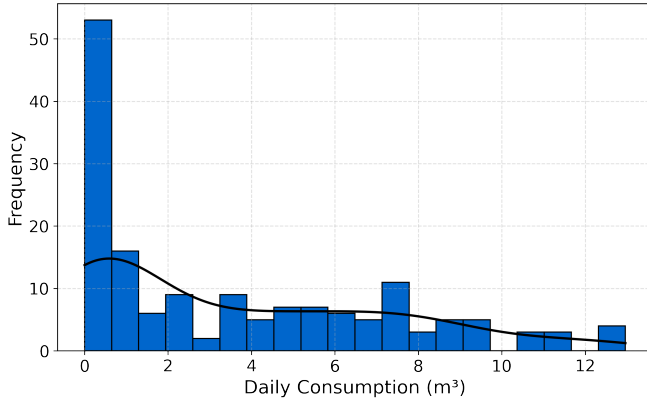


Fig. 4. Histogram of daily water consumption.

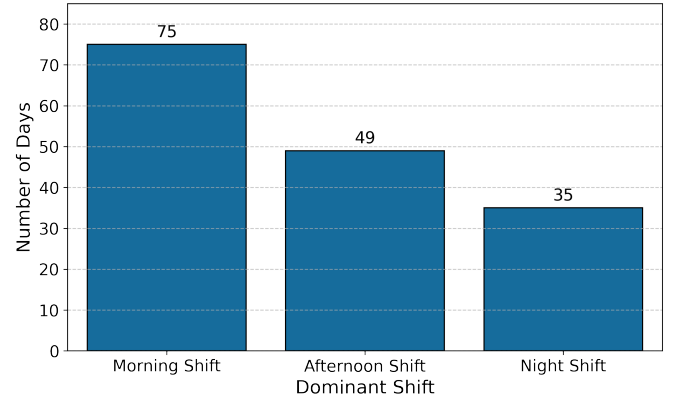


Fig. 5. Distribution of dominant consumption shifts across all recorded days.

cycle. This representation helps identify differences in both typical consumption levels and dispersion among operational days.

Key Observations: Wednesday exhibits the highest median consumption, indicating sustained midweek activity. Tuesday, Thursday, and Friday show the greatest variability, as reflected by their wider interquartile ranges. Saturday and Monday present moderate consumption with narrower distributions, whereas Sunday consistently records the lowest consumption and minimal variability, confirming limited operational activity.

Together, the circadian and weekly analyses reveal structured intraday and interday variability shaped by operational routines and intermittent supply conditions. These temporal patterns support the inclusion of contextual features in the

classification process. They also improve the interpretation of unusual consumption events by distinguishing expected operational variation from potential anomalies.

F. Label Generation and Static Thresholding

To support supervised classification, each day was assigned a semantic label based on total daily consumption and operational characteristics, using fixed non-parametric (percentile-based) thresholds for the entire analysis window; the thresholds are summarized in Table IV.

The anomaly threshold (8.0 m^3) was defined using both statistical dispersion and observed operational limits. Prior to January 20, 2025, the maximum recorded daily consumption under regular conditions was 7.57 m^3 ; setting the cutoff at

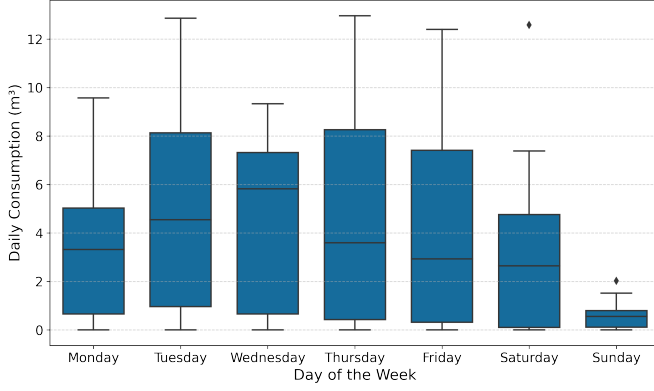


Fig. 6. Boxplot of daily water consumption by weekday, highlighting variability and operational rhythms.

TABLE IV
DAILY CONSUMPTION LABELS AND FIXED
NON-PARAMETRIC THRESHOLDS

Label	Threshold (m ³ /day)	Basis
Low	$x < 0.36$	Below 25th percentile (Q1).
Normal	$0.36 \leq x \leq 6.29$	Between 25th and 75th percentile (Q1–Q3).
High	$6.29 < x \leq 8.0$	Above typical usage range but within expected operational variability.
Anomalous	$x > 8.0$	Exceeds contextual and statistical norm.

8.0 m³ introduces a tolerance margin to capture structurally atypical events such as large-scale refilling or infrastructure testing without misclassifying natural fluctuations.

This study focuses exclusively on high-volume anomalies (≥ 8.0 m³). Pattern-based or shift-level deviations were not included in the labeling process.

Fig. 7 illustrates the distribution of daily consumption across the four categories, showing clear separation between consumption levels. The labeled dataset comprises 79 Normal, 40 Low, 18 High, and 22 Anomalous days, providing adequate representation for supervised modeling. Zero or minimal-activity days were grouped under Low to preserve class robustness while maintaining operational interpretability. These labeled patterns establish the basis for the classification and anomaly detection models evaluated in the following section.

VI. MACHINE LEARNING ALGORITHM EVALUATION

This section evaluates the classification of daily water consumption within the proposed framework. Each day is assigned to one of four operational categories (Low, Normal, High, Anomalous) using the percentile-based thresholds defined in Table IV. The input features correspond to cumulative consumption in the three operational shifts (Morning, Afternoon, Night).

The objective is not to introduce a new learning algorithm but to assess the ability of standard machine learning models to identify operational consumption patterns under realistic monitoring constraints. To this end, both unsupervised and

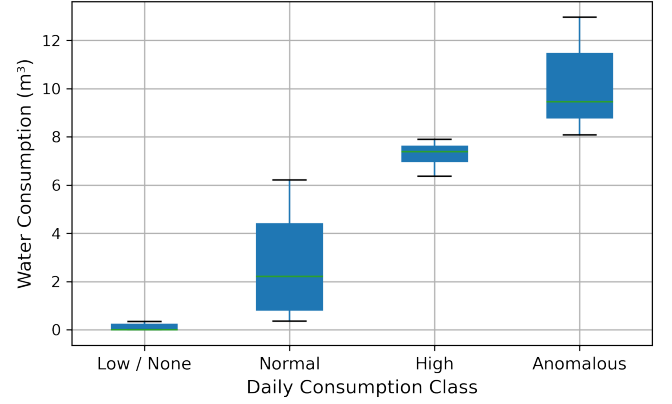


Fig. 7. Distribution of daily consumption across assigned labels.

supervised approaches were evaluated. Unsupervised clustering provides insight into latent behavioral structure, while supervised models perform explicit classification based on the labeled dataset.

We proceed in two steps. First, an unsupervised analysis explores latent structure in the data and its alignment with the operational labels, acknowledging that the percentile-based thresholds represent a first-order approximation of a non-stationary process. Second, a supervised multiclass formulation performs classification using shift-level consumption features.

For each day, four features were extracted: total daily consumption (m³) and cumulative consumption during the morning, afternoon, and night operational shifts.

All features are continuous and were normalized to the [0,1] range using Min-Max scaling to ensure consistent input magnitudes across models. No categorical variables were present, so encoding was not required. This preprocessing yields an interpretable feature set suitable for both clustering and classification tasks.

A. Unsupervised Learning Analysis

To explore the intrinsic structure of daily water consumption patterns, unsupervised clustering algorithms were applied to the normalized feature set. The objective was to assess whether meaningful groupings emerge without predefined labels, providing insight into natural separability and latent behavioral profiles.

Three representative clustering paradigms were evaluated: centroid-based (K-Means, $k = 4$), density-based (DBSCAN), and hierarchical (Agglomerative Clustering, $k = 4$). Prior to clustering, principal component analysis (PCA) was used to project the feature space into two dimensions for visualization. Fig. 8 shows the resulting cluster assignments, where DBSCAN additionally identifies isolated observations as noise points.

Cluster cohesion and separation were evaluated using the Silhouette Score (Table V). K-Means and Agglomerative Clustering achieved scores above 0.5, indicating moderately well-separated groupings, while DBSCAN produced lower scores but effectively isolated potential outliers.

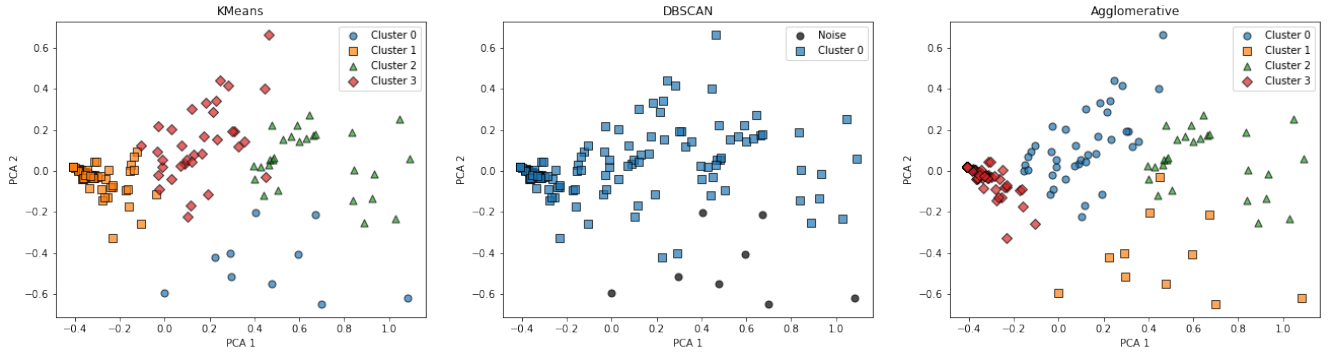


Fig. 8. PCA-based visualization of clustering results obtained using K-Means with $k = 4$ (left), DBSCAN (center), and Agglomerative Clustering with $k = 4$ (right).

TABLE V
SILHOUETTE SCORES FOR CLUSTERING ALGORITHMS

Algorithm	Silhouette Score
K-Means ($k=4$)	0.534
DBSCAN	0.436
Agglomerative ($k=4$)	0.519

While the clustering algorithms did not fully align with the operational labels used in supervised classification, distinct groupings emerged, particularly for low-consumption and high-demand days. Both K-Means and Agglomerative Clustering produced clusters that partially reflect expected usage categories such as Low, Normal, High, and Anomalous, indicating that unsupervised methods can approximate consumption profiles without predefined labels.

In contrast, DBSCAN produced a dominant cluster together with several isolated noise points, which frequently corresponded to exceptional events such as external supply deliveries or maintenance operations. This behavior highlights DBSCAN's potential as an unsupervised anomaly indicator.

Although clustering alone cannot match the precision of supervised classification, its ability to reveal latent behavioral patterns and detect outliers makes it a valuable complementary tool for exploratory monitoring and early detection of atypical consumption events.

B. Supervised Models Analysis

Building on the clustering analysis, six supervised learning models were evaluated to classify daily water consumption into four categories: Low, Normal, High, and Anomalous. The selected algorithms — Decision Tree, Random Forest, k-Nearest Neighbors, Support Vector Machine, Logistic Regression, and XGBoost — represent a range of model families, from interpretable rule-based methods to ensemble and margin-based approaches. This selection enables comparison across varying levels of model complexity and nonlinearity. All supervised models were evaluated using stratified 10-fold cross-validation to improve statistical robustness given the limited dataset size and class imbalance. The percentile-based labels were defined prior to model training and remained fixed throughout all validation folds.

The supervised models were evaluated using overall accuracy and macro-averaged F1-score, which together capture both global prediction performance and class-balanced evaluation. Table VI summarizes the average results obtained using stratified 10-fold cross-validation.

TABLE VI
PERFORMANCE METRICS FOR SUPERVISED MODELS
USING STRATIFIED 10-FOLD CROSS-VALIDATION

Model	Accuracy	Macro F1-Score
Decision Tree	0.988 ± 0.025	0.969 ± 0.079
Random Forest	0.988 ± 0.025	0.969 ± 0.079
XGBoost	0.988 ± 0.025	0.969 ± 0.079
k-Nearest Neighbors	0.931 ± 0.065	0.872 ± 0.136
Support Vector Machine (SVM)	0.906 ± 0.075	0.871 ± 0.115
Logistic Regression	0.698 ± 0.096	0.499 ± 0.114

Among the evaluated models, Decision Tree, Random Forest, and XGBoost achieved the highest average performance under stratified 10-fold cross-validation, exhibiting consistent accuracy and macro F1-scores across folds. In contrast, k-NN and SVM showed moderate reductions in macro F1-score, suggesting confusion in boundary cases where operational consumption profiles overlap. Logistic Regression produced the lowest performance, reflecting its limited ability to model the nonlinear relationships present in the feature space.

Figs. 9 and 10 present the confusion matrices obtained through stratified 10-fold cross-validation for the Random Forest and Logistic Regression models, respectively. Both models exhibit a strong diagonal structure, indicating good classification performance. However, the Random Forest classifier shows fewer misclassifications, whereas Logistic Regression exhibits greater confusion among neighboring consumption categories. Fig. 11 shows the PCA projection of the labeled dataset. The visualization reveals distinguishable structures for High Consumption and Anomalous Activity, while partial overlap remains between neighboring operational categories. This distribution supports the feasibility of supervised classification using shift-based volumetric features.

Inspection of the full time series revealed that most anomalous cases occurred toward the end of the monitoring period. This temporal concentration suggests that such events may be associated with broader contextual factors, including seasonal demand variations or maintenance cycles. Incorporating tem-

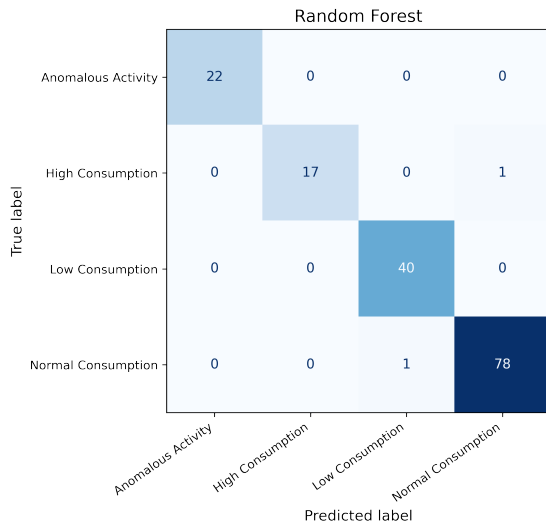


Fig. 9. Confusion matrix of the Random Forest classifier obtained using stratified 10-fold cross-validation.



Fig. 10. Confusion matrix of the Logistic Regression classifier obtained using stratified 10-fold cross-validation.

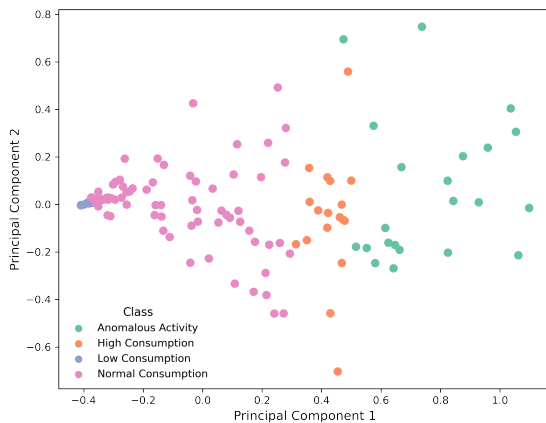


Fig. 11. PCA projection of the labeled water consumption dataset.

poral features in future models could therefore improve the early detection of atypical consumption behaviors.

All models were implemented using the Scikit-learn (v1.0.2) and XGBoost (v1.5.1) libraries in Python with default hyperparameters and `random_state=42` where applicable to improve experimental consistency. Supervised evaluation was performed using stratified 10-fold cross-validation, while no hyperparameter tuning was applied at this stage, as the objective was to establish baseline performance under standard configurations. Despite the simplified setup, several models achieved consistent classification performance, supporting the feasibility of the proposed feature representation for operational water consumption monitoring.

VII. DISCUSSION

The results demonstrate that machine learning can effectively classify urban water consumption patterns using volumetric data alone. By combining percentile-derived thresholds with operationally grounded features, the proposed framework distinguishes **Low**, **Normal**, **High Consumption**, and **Anomalous Activity** categories even under resource-constrained sensing conditions.

Implications of non-stationarity: EDA revealed regime shifts driven by calendar and supply constraints. While fixed percentile thresholds provide a reproducible baseline, their alignment may drift as operating conditions evolve. In deployment, two lightweight safeguards are advisable: (i) periodically re-estimating percentiles using a recent window (e.g., 60–90 days) so thresholds track the current regime; and (ii) monitoring distribution shifts in daily volumes or shift shares to trigger scheduled recalibration. These adjustments reduce false alarms during operational transitions without increasing sensing complexity and should be logged for traceability.

A. Operational Relevance and Model Performance

The supervised evaluation produced consistent results under stratified 10-fold cross-validation. Random Forest, Decision Tree, and XGBoost achieved the highest average accuracy and macro F1-scores, while SVM and k-Nearest Neighbors showed moderate reductions associated with overlapping operational profiles. Logistic Regression yielded the lowest performance, reflecting the nonlinear structure of the feature space and the limitations of linear decision boundaries under class imbalance.

Although the dataset remains limited in size, the cross-validation results indicate stable classification behavior across multiple data partitions, supporting the feasibility of the proposed framework under realistic monitoring conditions. A key contribution is the exclusive reliance on volumetric flow measurements, avoiding additional sensing modalities while still enabling robust classification using operationally interpretable features.

B. Insights from Unsupervised Learning

Although unsupervised clustering did not align with the predefined operational labels, it revealed latent structure in

the dataset. K-Means and Agglomerative Clustering produced partitions broadly associated with usage intensity, while DB-SCAN identified atypical events through its noise assignments, frequently corresponding to refill operations or maintenance activities. Nevertheless, clustering alone lacked the precision and interpretability required for operational classification, reinforcing the role of supervised learning for high-confidence automated decision-making.

C. Limitations and Path Forward

This study adopts several simplifying assumptions to preserve analytical tractability. Labels were defined using fixed global thresholds, and anomalies were restricted to high-volume events ($\geq 8.0 \text{ m}^3$), excluding intra-day or pattern-based deviations. Low-consumption and non-operational days were grouped to avoid sparse classes, and the dataset size limits seasonal generalization. These limitations motivate future extensions incorporating adaptive temporal and contextual modeling strategies.

D. Broader Implications

This work establishes a baseline for context-aware water consumption monitoring under constrained sensing conditions. Demonstrating that meaningful classification can emerge from volumetric data alone supports scalable deployment in cost-sensitive urban environments. It also provides a foundation for future predictive, adaptive, and multiscale monitoring strategies across buildings and urban water systems.

VIII. CONCLUSION AND FUTURE WORK

This paper presented the design, deployment, and evaluation of a facility-scale LPWAN-based IoT monitoring system for institutional water consumption analysis. Over a five-month period, the proposed architecture captured hourly-resolution volumetric data using a minimal sensing configuration and energy-efficient daily-burst telemetry. The collected dataset revealed clear operational rhythms and enabled both supervised classification and unsupervised exploration of consumption behavior.

A key contribution of this work is the demonstration that operationally meaningful consumption classification can be achieved using volumetric flow data alone, without requiring complex multi-sensor infrastructures. By combining percentile-based thresholding, contextual shift segmentation, LPWAN telemetry, and comparative machine learning evaluation, the proposed framework provides a practical baseline methodology for scalable and interpretable water consumption monitoring. The resulting architecture remains simple to instrument while sufficiently expressive to support pattern-informed management decisions under real operational constraints.

Supervised models, particularly Random Forest, Decision Tree, and XGBoost, achieved strong and consistent classification performance under stratified 10-fold cross-validation despite the limited dataset size and absence of temporal encoding. Complementary clustering analysis revealed latent

usage structure and highlighted atypical operational events, supporting the value of combining supervised classification with unsupervised diagnostics for context-aware monitoring.

Future work will focus on extending the framework toward broader urban-scale deployments by incorporating adaptive thresholding, temporal profiling, external contextual variables, and multi-season datasets to improve robustness under evolving operational conditions. Additional efforts will evaluate telemetry summarization strategies that further balance data freshness, communication overhead, and long-term energy efficiency.

ACKNOWLEDGMENTS

This work was partially supported by the Ramón y Cajal grant RYC2020-029925-I, funded by MCIN/AEI/10.13039/501100011033 and by the European Social Fund (ESF), "Investing in your future."

The authors would like to thank the Faculty of Nursing and Nutrition at the Universidad Autónoma de San Luis Potosí for its support during the execution of this study. The authors also acknowledge WND México for providing the Sigfox connectivity used by the deployed IoT infrastructure.

STATEMENT

AI-based tools were used for language refinement. All scientific content and decisions are the sole responsibility of the authors.

REFERENCES

- [1] United Nations (UN), the Department of Economic and Social Affairs, 2018. World Urbanization Prospects: The 2018 Revision, Key Facts. United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA.
- [2] C.J. Vörösmarty, P.B. McIntyre, M.O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S.E. Bunn, C.A. Sullivan, C.R. Liermann, P.M. Davies, "Global threats to human water security and river biodiversity," 2010, *Nature*, 467(7315), p.555. 10.1038/nature09549.
- [3] M. Rodell, et al., "Emerging trends in global freshwater availability," 2018 *Nature*, 557(7707), p.651.
- [4] A.Y. Rosinger, "Biobehavioral variation in human water needs: How adaptations, early life environments, and the life course affect body water homeostasis," *American Journal of Human Biology*, vol. 31, no. 5, Oct. 2019, e23338. doi: <https://doi.org/10.1002/ajhb.23338>
- [5] H. Pontzer, et al., "Daily energy expenditure through the human life course," *Science*, vol. 373, pp. 808-812, Aug. 2021, doi: 10.1126/science.abe5017.
- [6] J.M. Luna-Rivera, J. Rufo, J. Rabadan et al. "Urban genome: a new paradigm for sustainable cities." *npj Urban Sustain* 5, 77, 2025. <https://doi.org/10.1038/s42949-025-00265-1>
- [7] I. Andrić, A. Vrsalović, T. Perković, M. Čuvic, p. Šolić, "IoT approach towards smart water usage," *Journal of Cleaner Production*, Vol. 367, pp. 133065, 2022, doi: 10.1016/j.jclepro.2022.133065.
- [8] E. Syrmos, V. Sidiropoulos, D. Bechtis, F. Stergiopoulos, E. Aivazidou, D. Vrakas, P. Vezinias, I. Vlahavas, "An Intelligent Modular Water Monitoring IoT System for Real-Time Quantitative and Qualitative Measurements", *Sustainability*, Vol. 15, No. 3, pp 2127, 2023, doi:10.3390/su15032127
- [9] D. Arsene, A. Predescu, B. Pahonțu, C. Chiru, E. Apostol, C. Truică, "Advanced Strategies for Monitoring Water Consumption Patterns in Households Based on IoT and Machine Learning", *Water*, Vol. 14, No. 14, pp. 2187, 2022, doi:10.3390/w14142187
- [10] M. Pilotto Figueiredo and L. de Souza Oliveira, "AquaOculus: A Cost-effective Advanced Metering Infrastructure for Urban Water Distribution Systems", *IEEE LAT AM T*, vol. 24, no. 2, pp. 174–186, Jan. 2026.

- [11] M. K. Benhaoua *et al.*, "SW3T-IoT: Smart Water Monitoring Based on a Three-Tier IoT Architecture," in *IEEE Access*, vol. 13, pp. 206331-206354, 2025, doi: 10.1109/ACCESS.2025.3640507.
- [12] A. Morain, P. Vilar, L. Chen, "Artificial Intelligence for Water Consumption Assessment: State of the Art Review," *Water Resources Management*, vol. 38, pp. 3113-3134, 2024, doi: <https://doi.org/10.1007/s11269-024-03823-x>
- [13] K. Görenekli, A. Gülbağ, "Comparative Analysis of Machine Learning Techniques for Water Consumption Prediction: A Case Study from Kocaeli Province," *Sensors* 2024, 24, 5846. <https://doi.org/10.3390/s24175846>
- [14] J. V. Filho, A. Scortegagna, A. P. de Sousa Dias Vieira y P. A. Jaskowiak, "Machine learning for water demand forecasting: case study in a Brazilian coastal city," *Water Practice and Technology*, vol. 19, no. 5, pp. 1586-1602, 2024. doi: 10.2166/wpt.2024.096
- [15] M. Kavya, M. Aneesh, P. R. Shekar, Sarwesh P, "Short term water demand forecast modelling using artificial intelligence for smart water management," *Sustainable Cities and Society*, Vol. 95, 2023, 104610, doi: <https://doi.org/10.1016/j.scs.2023.104610>
- [16] L. Cao, X. Yuan, F. Tian, H. Xu, and Y. Zhang, "Forecasting of water consumption by integrating spatial and temporal characteristics of short-term water use in cities," *Physics and Chemistry of the Earth, Parts A/B/C*, Vol. 130, 2023, pp. 103390, doi: 10.1016/j.pce.2023.103390
- [17] J. Kim, H. Lee, M. Lee, H. Han, D. Kim, H.S. Kim, "Development of a Deep Learning-Based Prediction Model for Water Consumption at the Household Level, *Water*, Vol. 14, no. 9, 2022, pp. 1512, doi: 10.3390/w14091512.
- [18] J. Adrian SOMONTINA and E. Queen MACABEBE, "A Non-Intrusive Water Consumption Monitoring System," 2020 IEEE Global Humanitarian Technology Conference (GHTC), Seattle, WA, USA, 2020, pp. 1-6, doi: 10.1109/GHTC46280.2020.9342898.
- [19] M. Wang, Z. Li, Y. Liu, Lu Lin, C. Wang, "A hybrid model of machine learning for classifying household water-consumption behaviors," *Cleaner and Responsible Consumption*, Vol. 16, 2025, pp. 100252, doi: <https://doi.org/10.1016/j.clrc.2025.100252>
- [20] N. Gourmelon, *et al.*, "Implications of Experiment Set-Ups for Residential Water End-Use Classification," *Water* 2021, 13, 236. <https://doi.org/10.3390/w13020236>
- [21] A. Dahesh, *et al.*, "Classification of water subscribers by machine learning algorithms," *Water and Environment Journal*, vol. 38, pp. 45-58, 2024. doi: 10.1111/wej.12892.
- [22] H. Guo, X. Liu y Q. Zhang, "Identifying daily water consumption patterns based on K-means Clustering, Agglomerative Hierarchical Clustering, and Spectral Clustering algorithms," *Journal of Water Supply: Research and Technology-Aqua*, vol. 73, no. 5, pp. 870-887, 2024. doi: 10.2166/aqua.2024.294.
- [23] S. Lee, B. Kim, "Machine Learning Model for Leak Detection Using Water Pipeline Vibration Sensor," *Sensors* 2023, 23, 8935. <https://doi.org/10.3390/s23218935>
- [24] S. Pei, L. Hoang, G. Fu y D. Butler, "Real-Time Pump Scheduling in Water Distribution Networks Using Deep Reinforcement Learning," *Journal of Water Resources Planning and Management*, vol. 151, n.o 6, 2025. doi: 10.1061/JWRMD5.WRENG-6476
- [25] C. Maußner, M. Oberascher, A. Autengruber, A. Kahl, R. Sitzenfrei, "Explainable artificial intelligence for reliable water demand forecasting to increase trust in predictions," *Water Research*, vol. 268, Part B, 2025, 122779, doi: <https://doi.org/10.1016/j.watres.2024.122779>.
- [26] C. Luciani, F. Casellato, S. Alvisi, M. Franchini, "Green Smart Technology for Water (GST4Water): Water Loss Identification at User Level by Using Smart Metering Systems" *Water* vol. 11, no. 3, 2019. <https://doi.org/10.3390/w11030405>
- [27] R. Zese, E. Bellodi, C. Luciani and S. Alvisi, "Neural Network Techniques for Detecting Intra-Domestic Water Leaks of Different Magnitude," in *IEEE Access*, vol. 9, pp. 126135-126147, 2021, doi: 10.1109/ACCESS.2021.3111113.



J. M. Luna-Rivera received the B.S. and M.Eng. degrees in Electronics Engineering from the Autonomous University of San Luis Potosí, Mexico, in 1997 and 1998, respectively, and the Ph.D. degree in Electrical Engineering from The University of Edinburgh, U.K., in 2003. He is currently a Full Professor at the Faculty of Sciences, Autonomous University of San Luis Potosí. His research interests include signal processing for wireless communications, visible light communications, IoT systems, and vehicular communications.



Carlos A. Hernández-Morales received the B.S. degree in Telematics Engineering from the Polytechnic University of San Luis Potosí, Mexico, in 2016, the M.S. degree in electronics engineering and the Ph.D. degree in engineering sciences from the Autonomous University of San Luis Potosí (UASLP), Mexico, in 2019 and 2023, respectively. He is an adjunct professor at the UASLP. His research interests include applied IoT, wireless communications and data-driven modeling, with an emphasis on the experimental validation of intelligent systems.



Luis E. Baez-Salazar received the B.S. degree in electronics engineering from the Autonomous University of San Luis Potosí (UASLP) in 2019. He is currently pursuing the M.S. degree in electronics engineering at the Faculty of Sciences of UASLP. His research interests includes IoT systems and statistical analysis.



Julio Rufo received the Ph.D. degree in Telecommunications from the Universidad de Las Palmas de Gran Canaria (ULPGC), Spain. In 2016, he received the Spanish National Award for the Best PhD Thesis from the Official College of Telecommunication Engineers (COITT). His research interests include optical wireless communications, visible light communications (VLC), and optical communication systems for IoT and underwater environments. He is currently a Ramón y Cajal researcher at the Universidad de La Laguna, Spain.



José A. Rabadán M.S. (1995) and Ph.D. (2000) degrees from the Universidad de Las Palmas de Gran Canaria (ULPGC), Spain, where he is currently a Full Professor. His research focuses on wireless optical communications, particularly Visible Light Communications (VLC) and Optical Camera Communications (OCC) for IoT sensor networks in free-space, indoor, and underwater environments. He has participated in more than 40 national and international research projects funded by Spanish and European institutions, as well as 25 industry-funded projects. He is the author of 3 book chapters, over 60 journal articles, and more than 100 conference papers.