Advanced Unsupervised Machine Learning Frameworks for Smart Water Usage Pattern Detection and Anomaly Resolution

Shahid Ansari
Lovely Professional University
shahidlpu786@gmail.com

***Abstract*—The rapid digital transformation of the global water sector has generated massive volumes of high-resolution consumption data, necessitating the development of advanced analytical frameworks to ensure sustainable resource management. This report examines the application of unsupervised machine learning (ML) algorithms for identifying water-use patterns and anomalies in Smart Water Networks (SWN). In contrast to energy systems, water distribution exhibits stochastic, sparse, and zero-inflated data streams that pose challenges for conventional forecasting models.**

This study presents a complete data pipeline, beginning with preprocessing methods such as seasonal–trend decomposition and followed by feature engineering based on statistical moments and Minimum Night Flow (MNF). We evaluate multiple clustering algorithms and identify K-means as a robust baseline for consumer segmentation. Furthermore, deep learning architectures—particularly LSTM Autoencoders—demonstrate enhanced capability in detecting complex non-linear anomalies, including leaks and meter tampering.

The report also highlights the role of Edge AI and TinyML as emerging solutions for on-device inference, effectively addressing bandwidth and energy limitations in battery-powered smart meters. Supported by real-world case studies from major water utilities such as Anglian Water and Thames Water, this study illustrates how unsupervised learning can transform raw telemetry into actionable intelligence, resulting in substantial economic savings and significant contributions to water conservation.

Keywords: Smart Water Networks (SWN); Unsupervised Machine Learning; Anomaly Detection; Consumer Segmentation; K-Means Clustering; LSTM-Autoencoders; Edge AI; TinyML; Non-Revenue Water (NRW); Hydro-Informatics; Leak Detection.

I. Introduction: The Hydro-Informatics Paradigm Shift

The global water sector is experiencing a digital and scarcity-induced paradigm shift that fundamentally redefines both the challenges faced by utilities and the solutions required to address them. As rapid urban expansion intensifies demand and climate change drastically alters traditional rainfall patterns, the management of this vital resource has shifted from a branch of civil engineering and hydraulics to a data-driven discipline known as hydro-informatics[1]. The legacy water distribution model—characterized by ageing infrastructure, infrequent manual meter readings, and predominantly reactive maintenance—is increasingly inadequate for meeting the operational demands of modern utilities. In response, the Smart Water Network (SWN) has emerged, driven largely by Advanced Metering Infrastructure (AMI), which provides continuous high-resolution consumption data[3]. This manuscript highlights the decisive role of unsupervised machine learning (ML) in extracting actionable insights from such large and complex datasets. Water consumption patterns are inherently stochastic and discontinuous, shaped by human behaviour and diverse appliance signatures. Unlike the energy sector, the water industry lacks extensive labelled datasets required for supervised learning. Events such as leaks and theft are rarely recorded with sufficient accuracy in historical databases, rendering supervised classification methods impractical for large-scale deployment[6]. Consequently, unsupervised learning—which seeks to uncover hidden structures within unlabelled data—has become a primary source of intelligence for modern water utilities. This study follows the complete methodological pathway for deploying unsupervised learning within the water sector: preprocessing noisy sensor data, generating statistical and Minimum Night Flow (MNF)-based features, segmenting consumers through advanced clustering algorithms, and detecting anomalies using deep learning autoencoders. In addition, we explore the emerging fields of Edge AI and TinyML, where computational intelligence is shifted from cloud systems directly to smart meters, overcoming bandwidth constraints and enabling real-time decision-making[8]. By outlining conceptual frameworks and presenting quantitative case studies from major utilities such as Anglian Water and Thames Water, this paper demonstrates that these innovations are not merely theoretical enhancements but essential requirements for economic viability and environmental protection[10].

II. The Data Infrastructure of Smart Water Networks

The data acquisition layer forms the foundation of any machine learning (ML) application, and in Smart Water Networks (SWN) this layer consists of a geographically distributed set of sensor technologies, communication protocols, and data management systems. Together, these components determine the resolution, quality, and reliability of the data available for analysis.

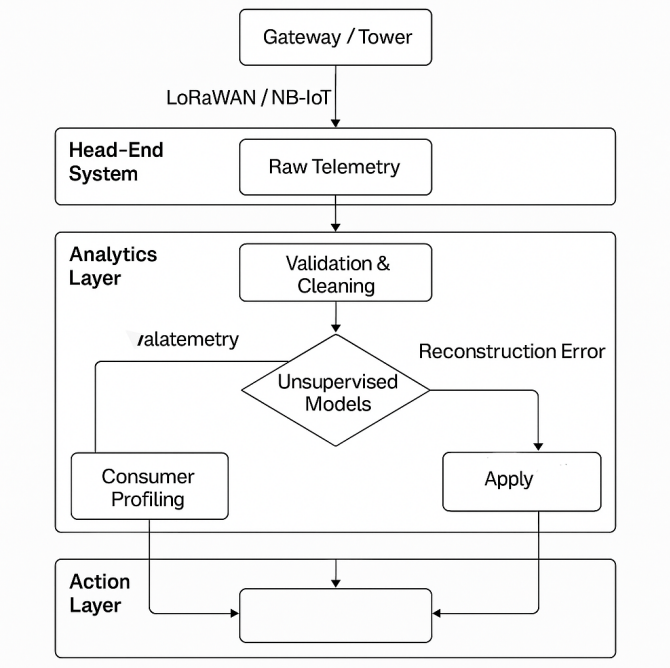


Fig. 1. Code Flow

A. Advanced Metering Infrastructure (AMI) and Telemetry

The transition from mechanical meters to smart meters represents a shift from recording cumulative volumes to continuously analysing flow rates. Smart meters capture water consumption at high temporal resolutions—typically every hour, 15 minutes, or even every second—and transmit this data through telemetry systems such as LTE-M, NB-IoT, LoRaWAN, or proprietary RF networks[3]. High-resolution data provides significant analytical value. A monthly meter reading produces a single data point, whereas an hourly reading yields 720 data points per month, revealing detailed consumption structures. Such granularity enables distinction between continuous low-flow leaks and legitimate high-volume usage (e.g., swimming pool filling). However, this increase in resolution introduces a "Big Data" challenge: a utility serving one million consumers generates more than 8.7 billion data points annually from hourly AMI readings[5]. This necessitates robust data pipelines, scalable storage, and efficient processing architectures.

B. Characteristics of Hydro-Informatics Data

Water consumption data exhibits several statistical characteristics that pose challenges for machine learning. Understanding these is essential for effective feature engineering and model selection.

1) Intermittency and Zero-Inflation: Unlike electricity, which maintains a non-zero base load due to always-on appliances, residential water consumption frequently drops to zero for long periods (e.g., nighttime or unoccupied hours). This zero-inflated nature violates assumptions of standard forecasting models such as ARIMA[5].

2) Stochasticity and Burstiness: Water usage is driven by discrete human actions such as flushing toilets, opening taps, or taking showers. These produce sharp spikes followed by immediate returns to zero, generating non-Gaussian, highly bursty distributions often modelled using Poisson pulse processes[13].

3) Seasonality: Consumption patterns exhibit strong cyclic behaviour:

- **Diurnal:** Morning and evening peaks associated with household routines.
- **Weekly:** Differences between weekdays and weekends.
- **Seasonal:** Weather-driven changes, including increased outdoor irrigation during summer months[3].

C. Data Preprocessing Architectures

Raw telemetry from smart meters is noisy, contains missing values, and may include transmission errors. A rigorous preprocessing pipeline is therefore essential.

1) Handling Missing Data and Imputation: Missing data typically results from LPWAN connectivity disruptions. Simple interpolation is unsuitable due to the stochastic and bursty nature of water usage. For instance, interpolating between a morning shower and an evening wash cycle would incorrectly imply steady usage throughout the day. More effective imputation methods rely on the consumer's historical behaviour patterns rather than system-wide averages, preserving behavioural signatures that are critical for downstream tasks[13].

2) Seasonal-Trend Decomposition: Multiple Seasonal-Trend decomposition using LOESS (MSTL) separates time series data into:

- **Trend:** Long-term consumption changes.
- **Seasonality:** Repeating daily or weekly patterns.
- **Residual:** Irregular fluctuations or anomalies[3].

Analysing the residual enables anomaly detection algorithms to focus on deviations not explained by predictable behaviour. For example, a high flow at 7:00 AM may correspond to normal activity, while the same flow at 3:00 AM appears anomalous.

3) Normalization Strategies: Consumption magnitudes vary widely across consumers, from small apartments to industrial facilities. Without normalization, clustering algorithms such as K-Means may group users based on total volume rather than behavioural shape. Techniques such as Min–Max scaling or Z-score standardization allow comparisons based on temporal patterns rather than absolute usage[3], enabling categorisation into behavioural archetypes such as "morning larks" and "night owls."

III. Feature Engineering: The Statistical Signature of Water Usage

Although deep learning models are capable of learning representations directly from raw data, explicit feature engineering remains highly valuable, particularly for lightweight unsupervised models. Time-series water consumption data can be summarised using statistical descriptors that reduce dimensionality while preserving essential behavioural characteristics.

A. Statistical Moments

Statistical moments provide a compact "fingerprint" of consumption behaviour. The four standardized moments describe key aspects of the distribution.

1) Mean: The mean represents the average level of consumption. While useful for establishing a baseline, it is highly sensitive to outliers and may not reflect typical behaviour.

2) Variance / Standard Deviation: Variance quantifies the degree of fluctuation in water usage. A household with a continuous leak may exhibit low variance relative to its mean, whereas irregular usage patterns produce high variance due to frequent bursts.

3) Skewness: Water consumption data is typically positively skewed, consisting of many zero or low-flow values and occasional high-consumption events. Skewness is calculated as:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Changes in skewness may indicate behavioural or mechanical changes. For example, a persistent leak increases the baseline flow, reducing right-skewness by making the distribution more symmetrical.

4) Kurtosis: Kurtosis measures the concentration of data around the mean and in the tails.

- **Leptokurtic (High Kurtosis):** Indicates a sharp peak and long tails, often reflecting a stable baseline with occasional extreme bursts (e.g., weekly irrigation).
- **Platykurtic (Low Kurtosis):** Indicates a flatter distribution, possibly due to irregular routines or persistent leakage.

These moments serve as key features for clustering and anomaly detection tasks.

B. Frequency Domain Features

Fast Fourier Transform (FFT) converts time-series consumption data into the frequency domain, revealing periodic patterns.

1) Dominant Frequencies: Strong frequency peaks, such as a 24-hour cycle, confirm daily behavioural routines.

2) Harmonics: Unexpected harmonics may indicate timed irrigation systems, mechanical oscillations, or irregular repetitive behaviours.

C. Minimum Night Flow (MNF)

Minimum Night Flow is one of the most important domain-specific indicators for leak detection. In residential areas, legitimate water usage is normally near zero between 2:00 AM and 4:00 AM.

1) Logic: During this window, the baseline flow should ideally approach zero.

2) Interpretation: A non-zero MNF is a strong indicator of a downstream leak or appliance malfunction, such as a continuously flushing toilet. Feature engineering pipelines commonly extract the mean, median, minimum, and slope of the MNF window for anomaly detection models.

IV. Unsupervised Clustering for Consumer Segmentation

Consumer segmentation aims to divide a heterogeneous customer base into smaller, behaviourally homogeneous groups. Traditional segmentation approaches rely on static demographic variables such as household size or income. In contrast, behavioural segmentation using smart meter data provides a dynamic and detailed representation of water use patterns. Most existing studies in this area rely entirely on unsupervised clustering algorithms.

A. K-Means Clustering: The Benchmark Standard

K-Means remains the most widely used clustering algorithm in the water sector due to its simplicity, scalability, and interpretability. The method partitions the dataset into K non-overlapping clusters, with each data point assigned to the cluster whose centroid is closest.

1) Mechanism: The algorithm iteratively:

1) assigns each data point to the nearest centroid,
2) recomputes centroids based on updated cluster memberships,
3) minimises the within-cluster sum of squares (WCSS) until convergence.

2) Performance: Comparative studies demonstrate the strong performance of K-Means on daily water consumption profiles. One evaluation reported a Silhouette Coefficient Index (SCI) of 0.6315 and a Calinski–Harabasz Index (CHI) of 305.92, outperforming Hierarchical and Spectral clustering [20].

3) Application: K-Means is effective at uncovering behaviour-based consumption archetypes such as:

- **Low-volume/flat users** (vacant or infrequently occupied homes),
- **Morning peakers** (working families),
- **Dual peakers** (typical 9–5 households).

Users who deviate significantly from their assigned cluster centroid may indicate anomalous behaviour, forming the basis for anomaly detection applications [20].

B. Agglomerative Hierarchical Clustering (AHC)

Agglomerative Hierarchical Clustering (AHC) builds a hierarchy of clusters rather than requiring the number of clusters to be specified in advance. The resulting dendrogram provides a visual summary of the merging process.

1) Mechanism: AHC begins by treating each data point as its own cluster. It then repeatedly merges the closest pair of clusters until all points are connected within a single hierarchy.

2) Pros and Cons:

- **Advantages:** Provides a clear hierarchical structure and allows exploration of segmentation at various levels.
- **Limitations:** Computationally expensive ($O(n^3)$), making it difficult to scale to millions of smart meters; tends to yield lower separation metrics (SCI = 0.5922) compared to K-Means [20].

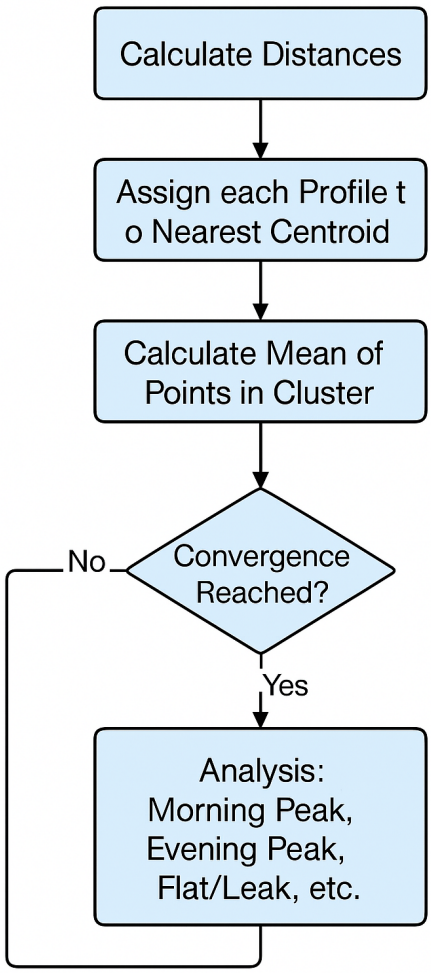


Fig. 2. Code Flow

C. Spectral Clustering

Spectral clustering uses graph-based similarity to identify cluster structures, making it suitable for detecting non-convex or complex-shaped groups.

1) Mechanism: The algorithm:

1) constructs a similarity graph,
2) computes the eigenvalues of the graph Laplacian for dimensionality reduction,
3) applies clustering in the reduced eigenspace.

2) Pros and Cons:

- **Advantages:** Effective at identifying non-convex clusters that K-Means may misclassify.
- **Limitations:** Computationally intensive due to eigenvalue decomposition; less suitable for large-scale deployment; performance metrics (SCI = 0.6272) slightly lower than K-Means [20].

D. Comparative Analysis of Clustering Efficacy

Table **??** summarises the performance of major clustering algorithms in the context of daily water consumption pattern identification, as reported in recent benchmarking studies [20].

E. Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) provide a neural-network-based approach to clustering, particularly advantageous for extremely high-dimensional data. SOMs project high-dimensional input vectors onto a low-dimensional (typically 2D) grid while preserving the topological relationships within the data.

1) Application in Water Analytics: In the context of smart water networks, SOMs are widely used to visualise behavioural patterns across consumer populations. Each node on the SOM grid represents a prototype consumption pattern, and consumers with similar behaviours are mapped to adjacent nodes. This produces a smooth, interpretable landscape of consumption phenotypes.

SOMs are especially useful for:

- identifying gradual transitions between behavioural groups;
- detecting users located near cluster boundaries, who often represent evolving habits or emerging leaks;
- providing visual interpretability for utility analysts.

Because of these properties, SOMs complement algorithms like K-Means by offering strong visualisation and topological insight, even though they may require more computation for training [20].

V. Dimensionality Reduction: Navigating the High-Dimensional Hydro-Space

A single year of hourly smart-meter data produces a vector with 8,760 dimensions. Direct analysis or visualisation of such high-dimensional data is impractical due to the *curse of dimensionality*, where distance metrics degrade and computational requirements escalate. Dimensionality reduction techniques are therefore essential preparatory steps for clustering, anomaly detection, and pattern discovery.

A. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear transformation method that identifies orthogonal directions—principal components—along which data variance is maximised.

1) Working Principle: PCA projects high-dimensional data onto a low-dimensional subspace spanned by mutually perpendicular axes. The first few components typically capture dominant patterns such as total usage volume and major seasonal cycles.

2) Use Cases:

a) Noise Reduction: By retaining only the leading components and discarding the remainder, PCA filters noise and stabilises downstream machine learning models [5].

b) Anomaly Detection: PCA can operate as a standalone anomaly detector. Training on normal consumption creates a principal subspace. When anomalous data points are projected into this subspace, they yield high reconstruction error because their variance lies in the discarded components [**?**].

B. t-Distributed Stochastic Neighbor Embedding (t-SNE)

While PCA preserves global structure, it may fail to capture complex nonlinear relationships. t-SNE is a manifold learning technique tailored for visualisation of high-dimensional data.

1) Operation: t-SNE converts Euclidean distances between point pairs into conditional probabilities using Gaussian kernels in the high-dimensional space and Student's t-distribution in the low-dimensional space. It then minimises the Kullback–Leibler (KL) divergence between these distributions through gradient descent.

2) Use Cases: t-SNE excels at revealing visually distinct behavioural clusters. For example, it can separate users into clear groups such as those who irrigate at night versus those who irrigate in the morning, whereas PCA might overlap these categories. Because the method preserves local neighbourhoods, consumers with similar habits appear near each other in the embedding [22].

C. UMAP (Uniform Manifold Approximation and Projection)

UMAP provides an alternative to t-SNE, offering competitive visualisation quality while preserving both local and global structures.

1) Working Principle: UMAP constructs a high-dimensional graph reflecting local topological relationships between points, then optimises a corresponding low-dimensional graph to maintain similar structure.

2) Why UMAP: Compared with t-SNE, UMAP:

- preserves global cluster relationships more faithfully,
- is computationally much faster,
- scales effectively to millions of smart-meter time-series.

These properties make UMAP particularly suitable for large-scale water utilities deploying behavioural analytics at operational scale [21].

VI. Anomaly Detection Architectures: From Statistics to Deep Learning

Anomaly detection—identifying deviations from expected behaviour—is the most financially impactful application of unsupervised learning in the water sector. Anomalies typically indicate one of three issues:

- **Physical Losses:** leaks, bursts.
- **Apparent Losses:** meter tampering, theft, calibration drift.
- **Data Issues:** transmission failures or sensor faults.

A. Taxonomy of Hydro-Anomalies

1) Point Anomalies: A single observation that is significantly different from the rest, such as a sudden spike in flow caused by a burst pipe.

2) Contextual Anomalies: A data point that is normal in magnitude but abnormal for its context. For example, a flow of 500 L/hr may be typical at 8:00 AM but highly anomalous at 3:00 AM.

3) Collective/Pattern Anomalies: A sequence of individually normal points that together indicate abnormal behaviour, such as a gradual rise in minimum flow over several weeks due to a growing leak.

B. Differentiating Leaks from Theft

A key strength of unsupervised learning in water analytics is its ability to distinguish between leaks and theft—both major contributors to Non-Revenue Water (NRW).

1) Leak Signatures: Leaks tend to be continuous. They elevate the Minimum Night Flow (MNF) and reduce the proportion of zero-flow intervals. From a statistical standpoint, leaks raise the lower bound of the consumption distribution.

2) Theft Signatures: Theft or tampering typically manifests as:

- abrupt drops in recorded consumption (e.g., bypassing the meter),
- magnetic interference causing the meter to stop logging,
- suspicious gaps or flatline segments in the data.

While leaks mainly modify consumption magnitude, theft disrupts the structure or continuity of the time series.

C. Statistical and Proximity-Based Methods

1) Z-Score and IQR Methods: These methods identify points outside statistical thresholds (e.g., more than 3σ from the mean or outside the interquartile range). Although computationally cheap, they assume near-Gaussian behaviour—an assumption frequently violated in zero-inflated, skewed water usage distributions.

2) Isolation Forest (iForest): Isolation Forest excels in high-dimensional settings. Rather than modelling normal behaviour, it isolates anomalies directly. Because anomalous points are "few and different," they require fewer splits in randomly generated isolation trees. iForest has been widely adopted in smart water analytics due to its scalability and robustness.

3) One-Class SVM (OC-SVM): OC-SVM maps data into a high-dimensional feature space using a kernel function (typically RBF) and learns a decision boundary that encloses the normal data while treating the origin as the "anomaly." Although effective for complex boundaries, its cubic computational complexity ($O(n^3)$) limits its applicability to large datasets.

D. Deep Learning Architectures

Deep learning models outperform classical methods when detecting subtle contextual or temporal anomalies because they learn complex nonlinear representations.

1) Autoencoders (AE): An Autoencoder is a neural network trained to reproduce its input. It consists of:

- an **Encoder**, which compresses data into a latent representation,
- a **Decoder**, which reconstructs the original input.

When trained on normal water consumption patterns, the Autoencoder learns an efficient latent representation of typical behaviour. Anomalous sequences—such as leak patterns—cannot be reconstructed accurately, resulting in a high

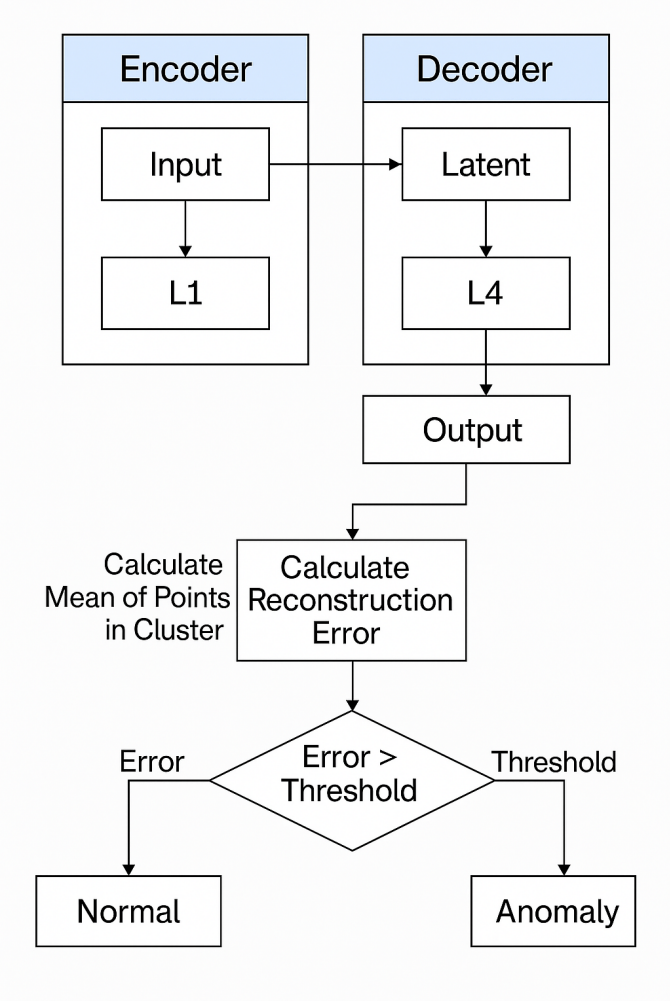


Fig. 3. Code Flow

Reconstruction Error (commonly Mean Squared Error). Any sequence whose reconstruction error exceeds a learned threshold is flagged as anomalous.

2) LSTM-Autoencoders: Traditional AEs treat inputs as fixed vectors, ignoring temporal structure. LSTM-Autoencoders incorporate Long Short-Term Memory (LSTM) layers, allowing them to model sequential dependencies.

a) Mechanism:

- The **Encoder LSTM** processes the time series step-by-step, compressing temporal dynamics into a latent vector.
- The **Decoder LSTM** reconstructs the sequence from this latent representation.

LSTM-AEs detect anomalies in temporal orderings, not just magnitudes. For example, if high flow at 7:00 AM is normally followed by a drop at 9:00 AM, a persistently high flow at 9:00 AM produces a high reconstruction error—even if the magnitude is not unusual globally.

3) Attention-Based Models (AT-DCAEP): The AT-DCAEP (Attention-based Double Convolutional Autoencoder-based Prediction) model represents a significant advance in unsupervised anomaly detection for multivariate time series.

a) Architecture: AT-DCAEP integrates:

- an **Attentive Prediction Network** to capture temporal dependencies,
- a **Convolutional Autoencoder** to extract spatial features across multiple sensors.

b) Mechanism: The attention mechanism directs the model to focus on the most informative sections of the time series—similar to how a domain expert inspects nighttime flow patterns for leaks.

c) Results: By jointly optimising reconstruction and prediction losses, AT-DCAEP captures complex correlations (e.g., between pressure and flow channels) and provides robust anomaly detection in noisy industrial water systems.

VII. Edge Computing and TinyML: Intelligence at the Source

Traditional Smart Water Systems rely on a centralized "cloud intelligence" paradigm in which smart meters transmit raw or minimally processed data to cloud servers for analytics. However, this approach suffers from several limitations, including high bandwidth costs, latency, privacy concerns, and accelerated battery depletion due to frequent high-power radio transmissions. Edge Computing and TinyML fundamentally reshape this paradigm by embedding intelligence directly into the meter [9].

A. Constraints of the Edge

Smart meters typically operate on resource-constrained microcontrollers (MCUs), such as ARM Cortex-M4 or M33 units, featuring modest clock rates (tens of MHz), limited RAM (KB to MB), and batteries designed for 10–15 years of operation.

a) Energy Budget: Wireless transmission is one of the most energy-expensive operations for battery-powered smart meters. Sending a single bit can consume energy equivalent to executing thousands of local CPU instructions. Thus, performing computations locally and transmitting only high-level insights (e.g., "Leak Detected") drastically improves battery life [9].

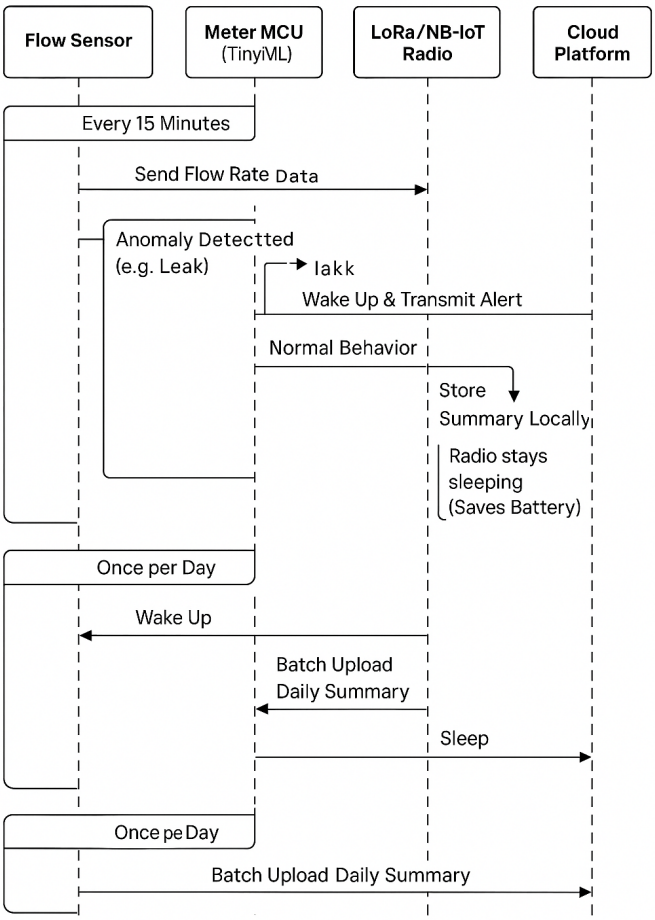
B. TinyML Frameworks and Optimisation

TinyML refers to machine learning techniques tailored for low-power embedded devices.

1) Frameworks: TensorFlow Lite for Microcontrollers (TFLM) is a widely used framework for MCU deployments. It provides an ultra-lightweight runtime suitable for devices with stringent memory constraints.

2) Model Compression: Conventional deep learning models are far too large for MCUs. Therefore, optimisation techniques such as quantisation (e.g., converting 32-bit floating-point weights to 8-bit integers) and pruning (removing redundant parameters) are applied. These techniques can reduce model size by factors of 4–10 with minimal loss in accuracy [8].

3) Hardware Acceleration: Modern embedded chipsets, such as the BK7258 (dual-core ARM Cortex-M33), now incorporate dedicated instructions for neural network inference, offering local acceleration for TinyML workloads [39].



Energy-saveng sieqaening-saving f a TinyML-enabled smart meter meter, interaf a trashldom

Fig. 4. Code Flow

C. Use Cases for On-Device AI

1) Real-Time Leak Alerts: On-device LSTM or CNN models continuously monitor the flow profile. When persistent non-zero flow indicative of a leak is detected, the meter activates the high-power radio to send a notification. This reduces "time-to-detection" from days (typical cloud batch-processing delay) to minutes [38].

2) Analog-to-Digital Conversion: Legacy analog meters can be retrofitted with camera modules running TinyML-based OCR models (e.g., MobileNet or lightweight YOLO variants). These models extract readings locally, enabling digital reporting without infrastructure replacement [41].

3) Data Compression via Autoencoders: Tiny autoencoders embedded in the meter can encode raw time-series data into a small latent vector. The compressed representation is transmitted to the cloud, where reconstruction occurs. This can reduce data transmission volumes to less than 10% of the original, significantly extending battery life [12].

VIII. Validation and Evaluation Strategies

Validation is a major challenge in unsupervised learning for water systems because ground-truth labels are typically unavailable. For instance, clusters representing "wasteful" behaviour cannot be verified, and anomalies such as leaks cannot be confirmed without field inspections. Several strategies address this limitation.

A. Internal Validation Indices

In the context of clustering, internal validation indices measure the geometric quality of clusters without requiring labels.

1) Silhouette Coefficient Index (SCI): SCI evaluates cohesion and separation. Values close to +1 indicate well-separated and internally consistent clusters.

2) Calinski–Harabasz Index (CHI): CHI is the ratio of between-cluster dispersion to within-cluster dispersion. Higher scores denote clearer segmentations. Studies consistently show that K-Means achieves the highest SCI and CHI for smart water datasets [20].

B. Evaluation for Anomaly Detection

Assessing anomaly detection accuracy in an unsupervised setting is inherently difficult. Common evaluation strategies include:

1) Synthetic Anomaly Injection: Artificial anomalies (e.g., constant added flow to simulate a leak) are injected into normal data. Performance is then measured using recall and precision based on the model's ability to detect the injected events.

2) Golden Batches: A small historical dataset containing verified anomalies—such as burst events corroborated by customer support logs—is used as a reference test set [5].

3) Accurately Diverse Ensembles: One advanced approach constructs a diverse ensemble of unsupervised models (e.g., Isolation Forest, Autoencoder, OC-SVM). When multiple models with different mathematical assumptions agree on an anomaly, confidence increases. Recent research proposes consensus metrics that allow hyperparameter tuning without labelled data [7].

IX. Quantitative Case Studies: Real-World ROI

Unsupervised learning technologies are no longer theoretical constructs; they are now central to achieving operational efficiency, reducing Non-Revenue Water (NRW), and enhancing service delivery across major water utilities worldwide. The following case studies illustrate quantifiable returns on investment (ROI) from large-scale deployments.

A. Anglian Water (United Kingdom)

Anglian Water, operating in one of the driest regions of the UK, has deployed more than 1.3 million smart meters. Their integrated analytics platform employs unsupervised learning methods to continuously monitor network health.

a) Impact: Over 600,000 customer-side leaks have been identified and resolved since 2021.

b) Water Savings: The programme achieved an average saving of 14.75 litres per household per day.

c) Financial ROI: The initiative has delivered annual savings of approximately £15 million for consumers (around £250 per affected customer).

d) Technology Stack: In addition to smart meter analytics, the utility employs thermal imaging drones and satellite-based Synthetic Aperture Radar (SAR) for leak detection, demonstrating the effectiveness of multimodal data fusion [10].

B. Thames Water (United Kingdom)

Thames Water, serving the London metropolitan region, utilises smart meter data to address leakage, affordability, and operational resilience.

a) Leak Detection: More than 84,000 customer-side leaks have been identified, saving approximately 120 megalitres of water per day—the equivalent of filling 1.5 million bathtubs daily.

b) Affordability and Demand Reduction: Analytics revealed that customers flagged with affordability indicators and consuming more than 500 L/day could reduce their bills by 8–17% (£40–£166 per year) following targeted water-efficiency visits.

c) Operational Efficiency: Deployment of smart data products increased the speed of clearing sewage blockages by a factor of 10 compared to pre-analytics operations [11].

C. Texas Water Utilities (USA)

Deployment of Cellular AMI across geographically challenging terrain enabled Texas Water Utilities to realise significant benefits.

a) The "Invisible" Leak: Unsupervised analytics detected a major leak in a cliffside community abstracting water directly from a lake—undetectable through visual inspection. Repairing it reduced NRW by nearly 60%.

b) Disaster Resilience: During the 2021 Texas Freeze, real-time analytics identified frozen pipe locations requiring shutoff. This prevented millions of gallons of water loss and mitigated extensive infrastructure damage [49].

X. Future Directions and Conclusion

The integration of Unsupervised Machine Learning within the water sector marks a significant step toward realising the broader vision of intelligent, adaptive Smart Cities. Water utilities are transitioning from mere data collection to genuine data-driven intelligence.

From an algorithmic standpoint, the sector is adopting robust, scalable methods such as K-Means for consumer segmentation and Isolation Forests or Autoencoders for anomaly detection. At the architectural level, the shift toward Edge Computing and TinyML has become unavoidable, driven by the constraints of battery life, bandwidth cost, and the need for real-time decision-making.

From a business perspective, the ROI is both substantial and measurable. The ability to detect leaks weeks before they surface or to identify customer-specific consumption behaviours for targeted engagement leads directly to multi-million-dollar savings and large-scale water conservation measured in gigalitres.

As global water scarcity intensifies, these "invisible" algorithms—embedded in the cloud and running on microcontrollers underground—are becoming indispensable stewards of the world's most valuable resource. The transition from traditional hydraulic engineering to hydro-informatics is complete: the future of water lies in data.

References

[1] Predictive Modeling for Water Anomaly Detection Using Machine Learning. IEEE Xplore. Accessed December 11, 2025. https://ieeexplore.ieee.org/document/11019440/

[2] Anomaly Detection in Smart Water Management System. SCRS Book Series. Accessed December 11, 2025. https://www.publications.scrs.in/chapter/pdf/view/360

[3] Modelling and Clustering Patterns from Smart Meter Data in Water Distribution Systems. SciTePress. Accessed December 11, 2025. https://www.scitepress.org/Papers/2025/132005/132005.pdf

[4] Using Smart Meters and Data Mining to Inform Demand Management. CRC for Water Sensitive Cities. Accessed December 11, 2025. https://watersensitivecities.org.au/wp-content/uploads/2016/07/TMR_C5-1_Smart_metering_data_mining.pdf

[5] Anomalous Water Use Detection Using Machine Learning. CEUR-WS. Accessed December 11, 2025. https://ceur-ws.org/Vol-3575/Paper3.pdf

[6] Detection of Anomalous Patterns in Water Consumption. Accessed December 11, 2025. https://upcommons.upc.edu/bitstreams/3bb39f2b-f9bb-4b86-9ab3-e8bd28422faf/download

[7] Towards Unsupervised Validation of Anomaly-Detection Models. arXiv. Accessed December 11, 2025. https://arxiv.org/html/2410.14579v1

[8] Edge AI & TinyML. Verpex. Accessed December 11, 2025. https://verpex.com/blog/reseller-hosting/edge-ai-tinyml

[9] TinyML: Enabling Inference Deep Learning Models on Ultra-Low-Power IoT Edge Devices. PMC - NIH. Accessed December 11, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9227753/

[10] Record Breaking Number of Smart Meters Installed in Anglian Water's Region. Accessed December 11, 2025. https://www.anglianwater.co.uk/news/record-breaking-number-of-smart-meters-installed-in-anglian-water-region/

[11] Smarter Ways Out of Water Poverty. Thames Water. Accessed December 11, 2025. https://www.thameswater.co.uk/media-library/home/about-us/responsibility/affordability/water-saving-affordability-study.pdf

[12] An Energy Efficient Smart Metering System Using Edge Computing in LoRa Network. IEEE Xplore. Accessed December 11, 2025. https://ieeexplore.ieee.org/ielaam/7274860/9976243/9316214-aam.pdf

[13] Exploring the Statistical and Distributional Properties of Residential Water Demand. MDPI. Accessed December 11, 2025. https://www.mdpi.com/2073-4441/10/10/1481

[14] Extracting Urban Water Usage Habits from Smart Meter Data: A Functional Clustering Approach. Accessed December 11, 2025. https://www.esann.org/sites/default/files/proceedings/legacy/es2017-31.pdf

[15] Multimodal Learning for Automatic Anomalies Detection in Water Consumption Profiles. IEEE Xplore. Accessed December 11, 2025. https://ieeexplore.ieee.org/document/11156494/

[16] Anomaly Detection of Water Level Using Deep Autoencoder. PMC - NIH. Accessed December 11, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC8512605/

[17] Understanding Skewness and Kurtosis in Data Analysis. Cuvette Tech. Accessed December 11, 2025. https://cuvette.tech/blog/understanding-skewness-and-kurtosis-in-data-analysis

[18] Statistical Notes for Clinical Researchers: Skewness & Kurtosis. PMC - NIH. Accessed December 11, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC3591587/

[19] Segmentation Analysis of Residential Water-Electricity Demand. ResearchGate. Accessed December 11, 2025. https://www.researchgate.net/publication/320576188

[20] Identifying Daily Water Consumption Patterns Based on K-Means. Accessed December 11, 2025. https://iwaponline.com/aqua/article/73/5/870/101909/

[21] Exploring the Influence of Dimensionality Reduction on Anomaly Detection Performance in Multivariate Time Series. IEEE Xplore. Accessed December 11, 2025. https://ieeexplore.ieee.org/iel8/6287639/10380310/10559232.pdf

[22] Introduction to t-SNE: Nonlinear Dimensionality Reduction & Visualization. DataCamp. Accessed December 11, 2025. https://www.datacamp.com/tutorial/introduction-t-sne

[23] How t-SNE Outperforms PCA in Dimensionality Reduction. Towards Data Science. Accessed December 11, 2025. https://towardsdatascience.com/how-t-sne-outperforms-pca-in-dimensionality-reduction-7a3975e8cbdb/

[24] Difference Between PCA and t-SNE. GeeksforGeeks. Accessed December 11, 2025. https://www.geeksforgeeks.org/machine-learning/difference-between-pca-vs-t-sne/

[25] Dimensionality Reduction: PCA, t-SNE, and UMAP. Medium. Accessed December 11, 2025. https://medium.com/@prathik.codes/dimensionality-reduction-dive-into-pca-t-sne-and-umap-9af37f6d9cb3

[26] An Automated Machine Learning Approach for Detecting Anomalous Peak Patterns. arXiv. Accessed December 11, 2025. https://arxiv.org/html/2309.07992v2

[27] How to Detect a Water Leak. LADWP. Accessed December 11, 2025. https://www.ladwp.com/account/customer-service/meter-information/how-detect-water-leak

[28] Tech Dive: Using Smart Meters to Detect Water Theft. Aquatech Amsterdam. Accessed December 11, 2025. https://www.aquatechtrade.com/news/utilities/tech-dive-using-smart-meters-to-detect-water-theft

[29] AMI + Analytics Curb Water Loss & Boost Bottom Line. Hubbell Blog. Accessed December 11, 2025. https://blog.hubbell.com/en/aclara/ami-analytics-curb-water-loss-and-boost-the-bottom-line

[30] Real-Time Anomaly Detection: Algorithms, Use Cases & SQL Code. Tinybird. Accessed December 11, 2025. https://www.tinybird.co/blog/real-time-anomaly-detection

[31] Unsupervised Anomaly Detection Algorithms on Real-World Data: How Many Do We Need? JMLR. Accessed December 11, 2025. https://jmlr.org/papers/volume25/23-0570/23-0570.pdf

[32] Towards Reliability in Smart Water Sensing Technology. PubMed Central. Accessed December 11, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11244236/

[33] Enhanced Water Leak Detection with CNN + OC-SVM. arXiv. Accessed December 11, 2025. https://arxiv.org/html/2511.11650v1

[34] Unsupervised Anomaly Detection in Water System Networks Using RNNs. Fenix. Accessed December 11, 2025. https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043837655/83425-Margarida-Costa_dissertacao.pdf

[35] Unsupervised Real-Time Anomaly Detection in Hydropower. MDPI. Accessed December 11, 2025. https://www.mdpi.com/2227-7080/13/11/534

[36] Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series. MDPI. Accessed December 11, 2025. https://www.mdpi.com/1424-8220/23/5/2844

[37] Unsupervised Deep Anomaly Detection for Industrial Multivariate Systems. MDPI. Accessed December 11, 2025. https://www.mdpi.com/2076-3417/14/2/774

[38] Smart Buildings: Water Leakage Detection Using TinyML. MDPI. Accessed December 11, 2025. https://www.mdpi.com/1424-8220/23/22/9210

[39] TensorFlow Lite Micro Developer Guide. BK7258 Documentation. Accessed December 11, 2025. https://docs.bekencorp.com/arminodoc/bk_aidk/bk7258/en/v2.0.1/api-reference/tensorflow.html

[40] TensorFlow Lite Micro with ML Acceleration. Accessed December 11, 2025. https://blog.tensorflow.org/2023/02/tensorflow-lite-micro-with-ml-acceleration.html

[41] Water Meter Reading Recognition Using Attention Mechanisms. PMC. Accessed December 11, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12459845/

[42] Edge-Computing Flow Meter Reading Algorithm for Agricultural IoT. Accessed December 11, 2025. https://arroma.uiowa.edu/docs/publication/paper_pdf/2023/Le_et_al_Flow_meter.pdf

[43] An Energy-Efficient Smart Metering System Using Edge Computing. Scholars' Mine. Accessed December 11, 2025. https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=2216&context=comsci_facwork

[44] Towards Unsupervised Validation of Anomaly Detection Models. ResearchGate. Accessed December 11, 2025. https://www.researchgate.net/publication/385091698_Towards_Unsupervised_Validation_of_Anomaly-Detection_Models/download

[45] Anglian Water's Smart Meters Save Millions. Smart Water Magazine. Accessed December 11, 2025. https://smartwatermagazine.com/news/anglian-water/anglian-waters-smart-meters-save-millions-litres-and-millions-bills-customers

[46] SUEZ and ASTERRA Partner with Anglian Water. Accessed December 11, 2025. https://www.suez.com/en/uk/news/press-releases/suez-and-asterra-partner-with-anglian-water-to-detect-invisible-leaks

[47] Smart Metering: A Step Towards Water's Digital Twin? Construction Management Magazine. Accessed December 11, 2025. https://constructionmanagement.co.uk/smart-metering-a-step-towards-waters-digital-twin/

[48] Million Plus Smart Meters Coming to Thames Water Region. Thames Water. Accessed December 11, 2025. https://www.thameswater.co.uk/news/2024/nov/million-plus-smart-meters-coming-to-thames-water-region

[49] Enhance Water Management with Smart Water Meters. Performance Services. Accessed December 11, 2025. https://www.performanceservices.com/resources/smart-water-meters-smarter-water-management-for-cities-and-towns/

[50] Texas Water Utilities Leads the Way with AMI & Leak Detection. Badger Meter. Accessed December 11, 2025. https://www.badgermeter.com/case-studies/texas-water-utilities/