AIT

ARIFA INSTITUTE OF TECHNOLOGY
EASANOOR,KEEZHAYUR(PO), NAGAPATINAM-611103

SMART WATER MANAGEMENT

STUDENT NAME : J.MOHAMED JASSUR

EMAIL ID : mdjassur17122003@gmail.com

DEPARTMENT : Electronics & communication Engineering

Abstract: Water resource management represents a fundamental aspect of a modern society. Urban areas present multiple challenges requiring complex solutions, which include multidomain approaches related to the integration of advanced technologies. Water consumption monitoring applications play a significant role in increasing awareness, while machine learning has been proven for the design of intelligent solutions in this field. This paper presents an approach for monitoring and predicting water consumption from the most important water outlets in a household based on a proposed IoT solution. Data processing pipelines were defined, including K-means clustering and evaluation metrics, extracting consumption events, and training classification methods for predicting consumption sources. Continuous water consumption monitoring offers multiple benefits toward improving decision support by combining modern processing techniques, algorithms, and methods.

Keywords: water consumption monitoring; machine learning; deep learning; Internet of Things

1. Introduction

Water represents an essential resource for survival. Today, the quality and quantity of water have decreased considerably. These are the effects of global industry development and the overexploitation of land and sea resources. Moreover, climate change has a strong impact on water resources, and drought is becoming more prevalent. All these factorscan cause major damage to water resources, so intelligent water management systemsare essential for maintaining efficient management in terms of the quality and quantity of drinking water.

One of the most important aspects of water sustainability is the continuous monitoring of water consumption in order to make the right decisions regarding the good management of this vital resource. Water scarcity, together with diseases caused by contaminated water, are major dangers that threaten humanity. Therefore, additional attention should be paid to this area, and the necessary resources should be allocated to monitoring water consumption. The collected data provide decision support for streamlining water resources.

Another important factor to note is population growth. The demand for water increases, and the need to develop an intelligent system becomes indispensable [1]. An alternative that has become increasingly popular, with multiple benefits in many areas, is based on the concept of the Internet of Things (IoT) [2]. IoT has made significant contributions in various fields, including being integrated into intelligent water management systems [3]. In households, the IoT concept can be implemented by installing sensors and collecting data in real time to provide continuous data monitoring. Solutions that include IoT in water management have multiple advantages including low costs and real-time remote data access. Furthermore, the integration of smart sensors in an existing system does not require many changes; IoT offers flexibility and only requires a few configurations to extend functionality [4].

The concept of IoT has evolved considerably, and it is almost indispensable today. For example, most intelligent water management systems use IoT sensors in the process of streamlining water consumption. In [5], the authors describe an IoT network that monitors water consumption in residential areas, including solutions for analyzing the data collected. Thus, this platform has multiple benefits for consumers and water suppliers, being a consistent basis for supporting the decision-making process regarding the control of water use in urban households. In the same way, Nie et al. [6] proposed an IoT solution for water monitoring in terms of quality and quantity, with the aim of detecting leaks or changing parameters that could signal a possible contamination of water. The proposed solution promotes awareness of the use of water resources, encouraging consumers and suppliers to make the right decisions to maintain a sustainable water supply system. An IoT-based pilot study was described in [7] for monitoring water consumption in a residential apartment in Naples. The data collected are public and can be used by the community to train different algorithms in order to raise awareness of responsible consumption, promoting water conservation.

Continuous monitoring and data analysis are essential for decision support systems and are a good starting point for water suppliers in terms of developing new strategies to optimize water consumption [8]. Moreover, the forecast plays an important role in this process, helping to streamline the water consumption [9]. Another proposal is an IoT system developed in India that aims to monitor water consumption in the tanks of an urban housing complex. In order to collect the data, the authors used ultrasonic sensors, while for the processing and analysis of data in terms of forecasting daily consumption and detecting leaks, the authors used machine learning algorithms [10]. A new perspective is presented in [11], where the authors outlined an IoT solution based on smart sensors to collect data and used algorithms to analyze them. The solution provides a forecast of water consumption and leaks that may occur in a household. The innovation of this article is the approach of the authors, who also considered the meteorological data to make a seasonal analysis.

Anomaly detection is a real benefit in time-series analysis for making predictions with high accuracy. Such an approach is presented in [12], where a rules-based decision system is proposed that improves the detection of anomalies in multivariate time series using the point of change detection.

In order to identify leaks or changes that may occur in water distribution networks as quickly as possible, the forecast needs to be accurate [13]. To this end, several solutions have been identified for predicting real-time water consumption [14]. These include: statistical methods, intelligent and filtering techniques, fuzzy logic, and combinations of several methods. Compared with classic techniques, machine learning (ML) techniques have shown encouraging results [15]. An in-depth analysis found that deep neural networks have the most effective prediction results. It should be noted that not only the data collected help to ensure forecasting performance but also data related to location, environment, weather conditions, and consumer patterns; data on the available infrastructure of the water distribution system are also vital. All these criteria are important factors in the decision-making process regarding streamlining water management [16]. A recent approach to data collection that also offers the possibility of monitoring and analyzing daily consumption is the advanced metering infrastructure (AMI), which is based on the IoT concept [17]. Furthermore, using ML algorithms such as long short-term memory (LSTM) [18] or back-propagation neural network (BPNN) [19], daily water consumption can be accurately predicted.

According to Koo et al. [20], a universal method by which water demand can be predicted has not yet been discovered, as this depends on many factors that can vary. For example, the forecast can be made in the short term (hours, days, or weeks) [21], medium term (between one and two years) [22], or long term (more than two years) [23]. Making a short-term forecast can be essential in the decision-making process of water suppliers, organizations, or companies at the managerial level. By reducing water consumption, the energy demand decreases and the water pressure in pipes increases, which is a great advantage for consumers.

For the analysis of the collected data to provide results with high accuracy, it is recommended that they not vary too much and instead have a certain continuity and uniformity. From this point of view, Kofinas, Spyropoulou, and Laspidou [24] implemented an algorithm that can generate synthetic data related to domestic water consumption that they tested in two European countries with good results. In the data analysis stage, the data are pre-processed to remove zero values or small leaks that could negatively influence the data related to the use of water in households.

Profiling consumers in a water distribution system is a key factor in ensuring a sustainable water system. Such an approach is presented in [25], who applied several methods that result in different patterns of consumers. Moreover, an extension of this study is described in [26] that consists of testing the classification and clustering methods in the consumer assessment process. A rather important feature in the development of a decision support system is the geolocation of households. Such a study is defined in [27] and presents an overview of multiple areas in a water distribution network. Going further with the research, an evaluation of the consumption activities in a household was undertaken in [28].

Additional attention should be paid to external factors that may influence water demand, e.g., weather, season, days of the week, and public holidays [29,30]. Recent research has shown that machine learning has much faster and more accurate results compared with traditional statistical models. Among the most common input data for machine learning models that aim to predict water time series were climate values [30]. A real challenge to obtaining highly accurate water demand prediction is choosing the relevant algorithms for data processing [31]. Prediction models can be grouped into two broad categories: machine learning and statistical models, which are often used to establish different relationships between data [32]. Statistical models are used in predicting water demand but have limitations in terms of data volume [33] and require a predetermined structure [32] compared with those of machine learning, where the large volume of data helps to achieve higher accuracy. Moreover, machine learning is being increasingly used and offering encouraging results in several fields [34]. Machine learning models can be further divided into two categories: simple (using a single algorithm, e.g., decision tree, SVM), or complex (using ensemble algorithms, e.g., random forests). Ensemble algorithms are more efficient for big data, involving multiple similar models [35], where each model makes separate predictions, and the final hypothesis is established on the principle of majority voting.

Another perspective based on machine learning for water control and monitoring is presented in [36] with the aim of creating a multimodel control supervisor for pumping stations that includes increased adaptability to changing operating conditions.

Khoi Nguyen et al. proposed in [37] a proactive approach to water demand awareness and management. This study combines advanced technology with machine learning to provide new strategies for more cost-effective water management. Mainly, various functions were used, among which were autonomous disaggregation, demand forecasting and recommendations offered to consumers based on the data collected. Disaggregation techniques are also used in [38], who proposed a system for monitoring water consumption in a residential apartment, in order to extract user behavior. A preliminary analysis of the data set was also performed.

Regarding the integration of machine learning in short-term water demand forecasting, the most used algorithms were LSTM (long short-term memory) [39], SVM (support vector machine) [40], and RF (random forests) [21]. Given that the development of intelligent solutions in terms of water demand forecasting is deeply researched, both traditional forecasting methods and improved methods, i.e., integrating machine learning algorithms, have been tested [41]. To obtain an efficient generalization with high accuracy, Candelieri et al. described how they used an SVM in the forecast evaluation process [42]. In order to improve the results obtained by a simple SVM, Brentan et al. highlighted a solution that combinesan SVM model with adaptive Fourier series [43].

This paper presents a study about water consumption using data collected from sensors installed in various households. The aim of the paper is to identify consumption patterns and to determine the source of water consumption from the collected data, using supervised and unsupervised learning methods. The novelty of this article consists in testing and analyzing various methods and algorithms for classification and clustering in order to obtain a more accurate prediction of consumption patterns in multiple households.

The rest of the paper is structured as follows: Section 2 outlines the architecture of the proposed system and describes in detail each component, with emphasis on the theoretical background; Section 3 presents the experimental results obtained using the data collected from multiple households, with references to source code and data repository and Section 4 summarizes the discussion of the results. Finally, the general conclusions are drawn in Section 5.

2. Methodology

This chapter describes the proposed methodology for the data collection and processing pipelines, which provide the context for the experimental results obtained in Section 3. The system architecture describes the water consumption monitoring solution. The data processing pipeline presents the methodology for water consumption analysis, for which the clustering and classification methods are described.

2.1. System Architecture

An intelligent water network management system is most effective when several factors are considered simultaneously, i.e., water suppliers, decision makers, and the direct involvement of consumers.

For efficient water consumption data collection, the sensor network is integrated into a cloud-based architecture as shown in Figure 1. Data acquisition is performed using a NodeMCU development board based on the ESP8266 microcontroller, with Wi-Fi communication. The platform has multiple GPIO pins (general-purpose input/output) connected to several YF-S201 flow meters and can be programmed using the Arduino environment [44] to monitor and collect the water flow through several pipes. The acquisition module is pre-programmed before installation, providing a configurable interface for connecting to the local network and defining a variable number of flow meters attached. To display the required configuration data, the sensor interface queries the server that is hosted directly on the ESP8266 microcontroller, which is configured in both station and Wi-Fi client modes.

Collecting data from the sensors is based on the MQTT (MQ telemetry transport) messaging broker, which includes a standard communication protocol, adapted to the IoT domain [45]. MQTT is a publish–subscribe communication protocol that is suitable for exchanging messages between IoT devices [46] as it supports bidirectional TCP/IP connection. Thus, IoT devices can connect to the message broker that has the role of routing the messages received from customers and can post messages on a topic or subscribe to a topic to receive messages.

Real-time data access is provided by the MQTT client (e.g., desktop application) with consideration of the required access permissions at the broker level. Furthermore, in order to access data in real time, a REDIS module has been integrated, which stores a snapshot of the latest data collected [47].

The data transmission over MQTT is done in two ways: consumption data (higher frequency for real-time flow monitoring) and volume data (lower frequency for database storage). For the sensor data to be displayed in the user interface, the application server subscribes to the MQTT broker and pushes the updates to the mobile application. The real-time feed is provided by the sensors using a sampling time of 10 s. For not overloading the database, the real-time data is stored periodically, with a lower sampling frequency, i.e., once per minute, which can be configured for each connected IoT device.

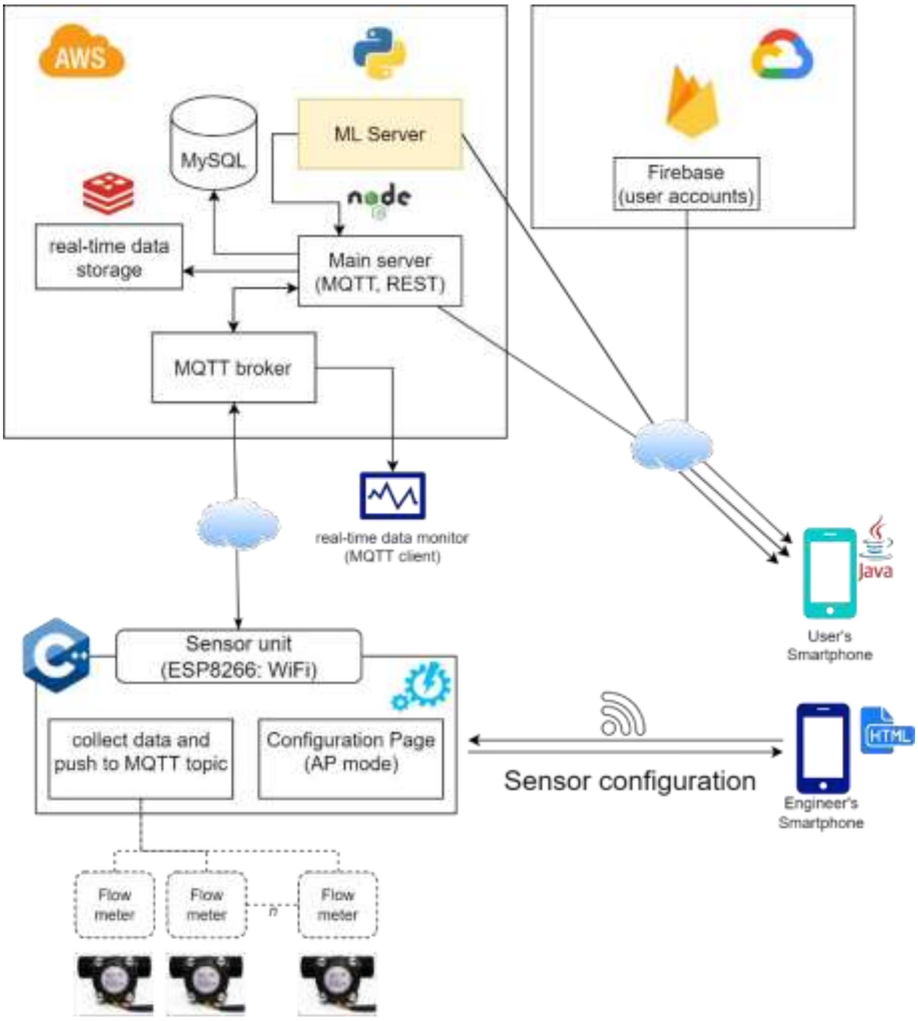


Figure 1. The water consumption monitoring system architecture.

The consumer can visualize the status of a sensor, as well as water consumption data through a mobile application. The Firebase component provides mechanisms for the authentication and management of user accounts (e.g., consumers, suppliers) and the data are displayed in the user interface through HTTP requests to the main server.

2.2. *Data Processing Pipeline*

The processing pipeline is shown in Figure 2. For a better understanding, each step will be described below. The raw data set (time series) was generated by collecting data from sensors installed in multiple households. Four types of water outlets were considered, i.e., sink (cold and hot water measured separately), toilet, and shower. The raw data set were processed to be evaluated and tested by first eliminating nonrelevant data that could negatively impact the results.

The time series of daily consumption were further extracted from the raw data set to make it easier to assess and better understand water consumption patterns for each outlet with a high level of accuracy. Furthermore, based on the identified patterns, several tests were performed, K-means clustering, daily batch processing, and event processing, in order to evaluate consumption profiles from different perspectives that are not necessarily correlated with a particular consumer. While the exact source of the consumption is known in the context of the experiments, the proposed methodology defines multiple strategies for uncovering various patterns in terms of overall consumer behaviors.

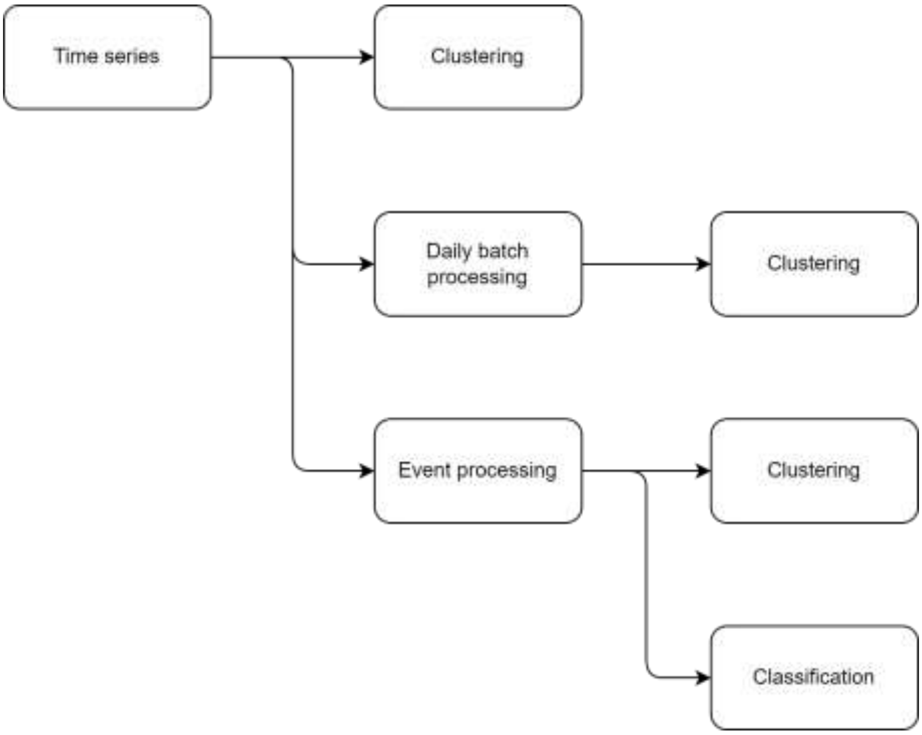


Figure 2. The processing pipeline.

In the first step, K-means clustering [48] was used over time series to extract distinctive water consumption patterns in terms of variability and total volume. To provide a better qualitative overview of consumption trends, a moving average filter was applied on the raw data set. For a quantitative evaluation of the clustering results, the metrics described in Section 2.3 were used for both original and filtered data.

Finally, a more in-depth analysis was performed, and the individual consumption events were extracted from the time series, being characterized by duration (minutes) and total volume (liters). In order to analyze the accuracy of the results in terms of identifying the types of consumer outlets, K-means clustering was again applied to the events extracted from the data set.

In the classification stage, only two measuring points (combined hot/cold sink water and toilet water) were considered. The purpose of this stage was to train the classification models to predict the two types of consumption events based on the entire data set. Four classification methods were used, two from the machine learning category and two from the deep learning category, these being described in Section 2.4.

The processing pipeline was implemented in Python using the scikit-learn package [49], version 0.24.0, for clustering and machine learning-based classification methods, while the Keras package [50], version 2.4.0, was for the deep learning models.

2.3. *Clustering Methods*

Profiling consumer outlets by water consumption involves using the K-means clustering algorithm with a predefined number of clusters. This approach was used to identify consumption patterns from the data set that were represented as time series as well as consumption events defined by duration and volume. Regardless of the data set, the algorithm iteratively assigns the data points to the nearest centroid, recalculating the centroid at each step using a stopping criterion based on Euclidean distance.

In this study, we used clustering methods to analyze the quality of the results based on the data set. Furthermore, to evaluate the accuracy of clustering with regard to the known consumption outlets, we used the confusion matrix together with the following parameters: silhouette score (S), Rand index (RI), adjusted Rand index (ARI), purity (P), and entropy (E).

The silhouette score is used to determine the separation distance between clusters by considering the intra-cluster distance and the nearest clusters for all samples. The parameter may have values from 1 to 1. If the value is closer to 1, then the separation between clusters is higher, and if the value is around 0, the sample is closer to the decision boundary. The formula for the silhouette score is:

$$S = \frac{(nc - ic)}{\max(ic, nc)} \tag{1}$$

where:

- ic is the average intra-cluster distance;
- nc is the average nearest-cluster distance.

The Rand index provides a comparison between cluster assignments (predicted and known groups), showing similarities between the two groups. In our study, this measure was used to determine the deviations of the identified consumption profiles from the known consumer outlets. RI can take values between 0 (clustering disagrees) and 1 (perfect match) considering the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The formula for the Rand index is:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

If there are different number of clusters or different assignments in two clusters, the adjusted Rand index can be used to provide a better overview:

$$ARI = \frac{RI - \text{expected\_}RI}{\max(RI) - \text{expected\_}RI} \tag{3}$$

To calculate the purity, the majority class should be identified in each cluster, and for each cluster, the tagged objects number should be calculated. The calculated value must be between 0 (for random labeling) and 1 (for identical clusters). In our study, this measure was used to evaluate the correctness of the assigned consumption profiles. The formula for purity is:

$$P = \sum_{i} p_i max_j \frac{p_{ij}}{p_i} \tag{4}$$

The definitions for p_{ij}, p_i, p_j (using the confusion matrix) are the following:

$$p_{ij} = \frac{n_{ij}}{n}; \quad p_i = \frac{n_i}{n}; \quad p_j = \frac{n_j}{n} \tag{5}$$

where:

- n_{ij} is the number of matches for cluster i and class j;
- n_i, n_j is the sum of matches for given row/column.

The entropy presents the similarity (for low values) or the dissipation (for high values) of the data and is defined by the following formula:

$$E = \sum_{i} p_i \sum_{j} \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i} \tag{6}$$

2.4. Classification Methods

Consumer outlets are classified according to the water consumption using supervised learning algorithms. Training models are created with the main purpose of predicting the right class for a consumer outlet.

To test and evaluate the classification methods on our data set, we considered four different supervised learning algorithms: decision tree (DT) using the CART algorithm; ensemble methods using the random forests (RF) algorithm where we grow multiple CART trees; multilayer perceptron (MLP) using the Dense algorithm; and recurrent neural network (RNN) using LSTM cells.

The above four types of algorithms are divided into two groups. DT and RF are part of machine learning algorithms, while MLP and RNN represent deep learning algorithms. In the algorithm evaluation process, 80% of the data set was used for training, and the remaining 20% was be predicted by the supervised learning algorithms.

In the context of linear data sets, DTs present fast and efficient solutions. Given this aspect, we decided to implement them in our study. For greater accuracy, the RF algorithm provides better results, as will be seen from the tests applied to the data set. For comparing the two machine learning algorithms, RF combines several DTs to be more complex and more accurate, with lower overfitting.

In terms of deep learning algorithms, they have had a remarkable evolution in recent years, offering facilities such as multilayer architectures, normalization, and pruning [50]. It should be noted that deep learning is recommended for use when the data set is larger, the results being more accurate. In the proposed study, we chose to use the Dense and RNN/LSTM algorithms for deep learning, making a comparison with the results obtained by the two traditional machine learning algorithms: RF and DT.

Decision trees are supervised algorithms used to build prediction models using a tree structure, either for classification or for regression. They are constructed incrementally by splitting the data set into smaller subsets. The splits are determining by the best splitting attribute's values. Thus, the final output of the algorithm is a tree structure where nodes, labeled with attributes, act as decisions and leaves, labeled with classes for classification or target values for regression, as predictions. In our implementation, we use the CART algorithm.

CART (classification and regression tree) is a greedy decision tree algorithm that constructs the tree top down. At each step, a "best" split attribute is chosen among all the sample's (D) attributes (A_t, $t \in \{1, 2, \ldots, m\}$) by computing its homogeneity using the GINI impurity:

$$G_{impurity}(D \mid A_t) = \sum_{i=1}^{r} \frac{n_i}{n} G_{index}(D_i) \quad (7)$$

where:

- n the size of D;
- n_i the size of subsample D_i;
- r the different values of attribute A_t;
- $G_{impurity}(D_i)$ the GINI impurity for the subsample D_i computed using the GINI Index $G_{index}(D_i) = \sum_{j=1}^{k} p(c_j)(1 - p(c_j))$ with $p(c_i)$ the probability of class c_j in D_i.

Ensemble methods are used to minimize overfitting and improve the stability and performance of machine learning algorithms. Bagging is an ensemble method that avoids overfitting by employing sampling with replacement, which reduces the variance within the data set. Bagging employs the following steps: (1) build multiple data sets, named bootstrap samples, from the original data set using sampling with replacement; (2) create a model using each of the bootstrap samples; and (3) combine the results of each model to obtain a final prediction. Bagging is used together with decision trees, which are greedy and tend to overfit, to obtain bagged decision trees. Unfortunately, bagged decision trees produce tree structures that are very similar and that as such offer highly correlated results among the different tree classifiers built.

Random forests produce less-similar trees by limiting the number of attributes used by the decision tree algorithm, here CART, from which to search for the best split. Similar to bagged decision tree, random forests also construct multiple decision trees. The difference between them consists in the heuristic used for obtaining the best splitting attribute. Thus, the improvement of random forests over bagged decision tree lays in limiting the sampling features to a random selection from all the attributes. Using this random selection, the produced trees are less similar, and the prediction correlation among them is lowered. The final prediction, as in the case of bagging, is the aggregation of the results of all the obtained different classifiers.

Multilayer perceptron (MLP) is a feed-forward artificial neural network (ANN) consisting of perceptron units added in three fully connected weighted directed layers, i.e., input, hidden, and output.

By adding multiple hidden layers, an MLP architecture evolves into a deep feed-forward neural network.

The perceptron is a basic processing unit that associates its input x_i to one binary value, i.e., $\hat{y}_i \in \{0, 1\}$ It can be generalized to a multiclass perceptron that predicts $\hat{y} = argmax_y\, f(x, y)\; w$ using $f(x, y)$ a feature representation function that associates each *input/output* pair to a finite-dimensional feature vector and multiplies it by a weight vector w.

In our implementation, we used a Dense layer as input and relu as the activation function. The hidden layer consists of several Dense layers to which are added Dropout layers to prevent overfitting. Another Dense layer with the softmax activation function is used as the output layer. The final layer is used for classification.

Recurrent neural networks (RNNs) are successfully used in sequence processing due to their capability to cache the previous outputs and add their information to the current inputs.

For long-term dependencies, the classical RNN encounters a phenomenon that affects the gradient, i.e., the vanishing gradient that can completely stop the network from further training. To deal with this problem, a new type of RNN was introduced, i.e., LSTM (long short-term memory). This model uses the gates functions in addition to the loops in the RNN and also has extra information called memory that helps resolve the aforementioned problem.

In an LSTM model, there are three gates: input, output and forget. The model receives as input at each time step: the current input (x_t), the previous hidden state (h_{t-1}), and the previous memory state (c_{t-1}). The outputs at every time step consist of the current hidden state (h_t) and the current memory state (c_t).

The expressions for each gate (i.e., input i_t, output o_t, and forget f_t) as well as the ones for the current hidden state (h_t) and current memory state (c_t) are as follows:

$$i_t = \sigma_s(W_i x_t + U_i h_{t-1} + b_i)\; o_t$$
$$= \sigma_s(W_o x_t + U_o h_{t-1} + b_t)\; f_t =$$
$$\sigma_s\; W_f x_t + U_f h_{t-1} + b_f$$
$$\tilde{c}_t = \sigma_h (W_c x_t + U_c h_{t-1} + b_c)$$
$$c_t = f_t \circledS\; c_{t-1} + i_t \cdot c_t$$
$$h_t = o_t \circledS\; tanh\,(c_t)$$

where:

- x_t is the input at time step t;
- h_t is the next hidden state and h_{t-1} is the previous hidden state;
- c_t is the cell input activation vector;
- c_t is the current memory state and c_{t-1} is the previous memory state;

- W_i, W_o, and W_f are the weights for each gate's current input state, while U_i, U_o, and U_f are the ones for the previous hidden state;
- b_i, b_o, b_f, and b_c are the bias vectors;
- σ_s is the sigmoid activation function;
- σ_h is the hyperbolic tangent activation function;
- Ⓢ is the Hadamard product.

In our implementation, we use three LSTM layers with 32 units each. These layers are followed by a Dense layer used for classification.

3. Experimental Results

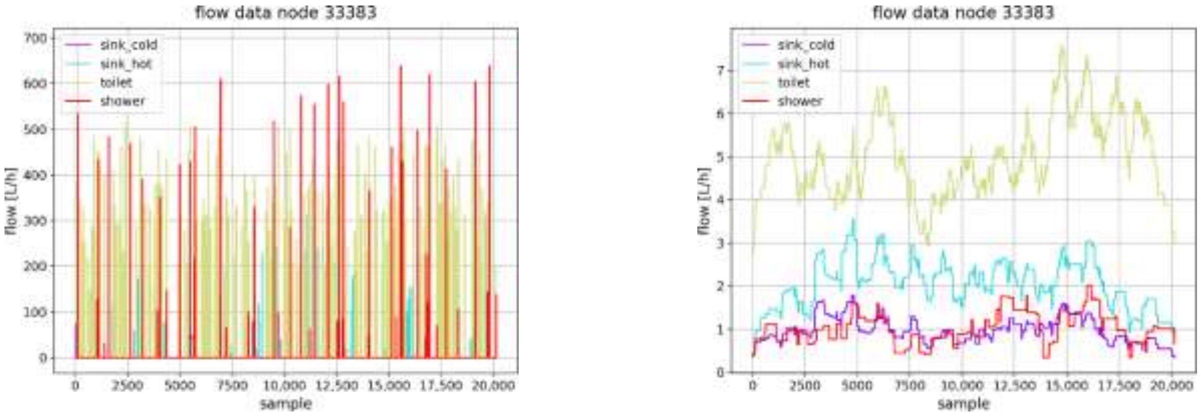
This section describes the results obtained using the proposed methodology. The consumption profiles are evaluated using clustering methods, showing weekly and daily consumption patterns. Consumption events are processed and evaluated using clustering and classification methods to provide an additional level of detail.

The data sets, collected from smart meters installed in households, over a time frame of several months, and the processing scripts defined for the clustering and classification methods used in this article, are included in the following GitHub repository, created by the authors: https://github.com/alexp25/watergame-other/tree/main/sensors (accessed on 30 May 2022).

The scripts include batch processing for extracting water consumption events from the raw data sets and visualization scripts for the proposed evaluation scenarios. The classification scripts include the training and evaluation of machine learning and deep learning models.

3.1. Consumption Data Evaluation

The data collected from a multisensor node are shown in Figure 3a, showing water consumption from 4 types of water outlets: sink cold (cold tap water), sink hot (hot tap water), toilet, and shower, from 33 sources, over a time frame of 1 week, with a sampling time of 60 s. The instantaneous flow is measured by the sensors and represented in L/h, while the individual samples are represented on the x-axis.



(a)

(b)

Figure 3. Sample data set for multisensor node: (**a**) raw data; (**b**) filtered data.

The individual consumption events can be observed in this case as the duration is very short compared with the entire time frame. To evaluate the relative consumption patterns, a moving average filter was applied to the data set, as shown in Figure 3b. It is now possible to compare the average water consumption volumes between the water outlet types as follows: toilets accounted for the most water consumption (1716 L), followed by hot tap water (699 L), shower water (357 L), and cold tap water (335 L).

3.2. *Consumption Profiles Evaluation*

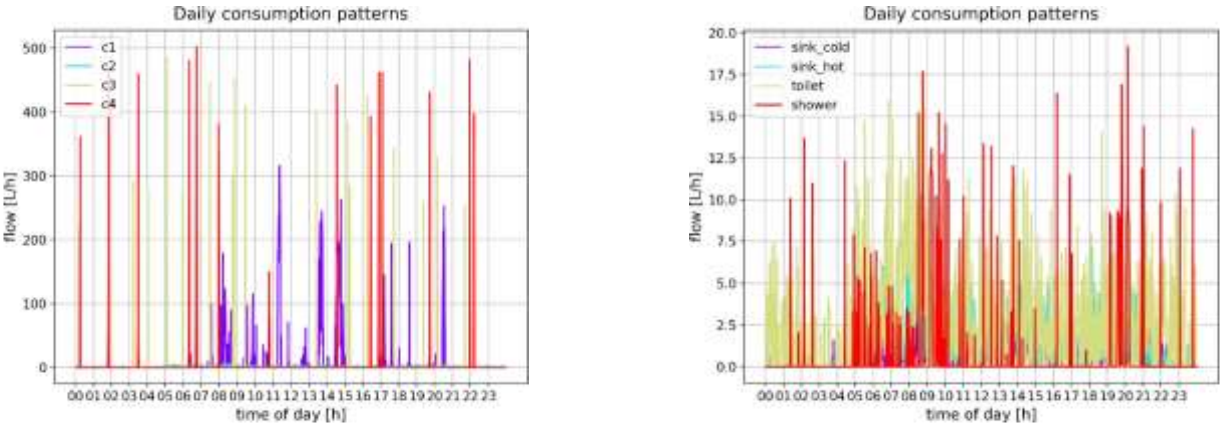
To provide a side-by-side comparison between the clustering results and the consumer outlet classes defined in the data set (i.e., sink cold, sink hot, toilet, and shower), four clusters were used for K-means clustering that revealed distinctive patterns in terms of variability and total volume.

The first evaluation was performed on the raw data set and the filtered one. We used the metrics described in Section 2 to provide a numerical evaluation of the results, and we obtained the scores presented in Table 1, which shows moderate similarity between the clustering assignments and the consumer outlet classes. The silhouette score is higher for the filtered data set, showing a higher quality of clustering, while the other metrics show mixed results in terms of the correlation of the identified clusters with the known consumer outlet groups.

Table 1. The evaluation results for the time-series clustering.

Evaluation Method	Score (Raw Data Set)	Score (Filtered Data Set)
Silhouette score	0.188	0.297
Rand index	0.356	0.454
Adjusted Rand index	0.010	−0.013
Purity	0.424	0.393
Entropy	0.334	0.335

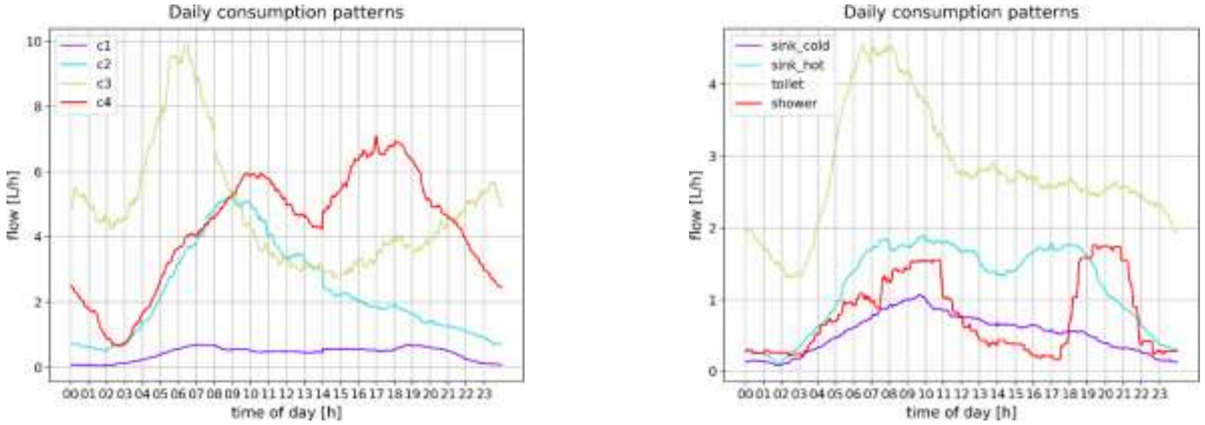
For a better understanding of daily consumption patterns, the daily consumption time series were extracted from the raw data set. Then, K-means clustering was used to extract the daily consumption patterns as shown in Figure 4a. For evaluating the clustering performance, the consumption patterns extracted from actual consumer outlet classes are shown in Figure 4b.



(**a**) (**b**)

Figure 4. The daily consumption patterns on the raw data set: (**a**) the consumer outlet clusters; (**b**) the consumer outlet groups.

To provide another perspective in terms of consumption trends, the clustering was performed on the daily consumption trends extracted from the filtered data set (using moving average filtering), as shown in Figure 5a. The consumption trends for the actual consumer outlet classes are shown in Figure 5b.



(a)

(**b**)

Figure 5. The daily consumption trends on the filtered data set: (**a**) the consumer outlet clusters; (b) the consumer outlet groups.

In this second evaluation performed on the raw and filtered data set, we used the metrics described in Section 2 to provide a numerical evaluation of the results, and we obtained the scores presented in Table 2, which show moderate similarity between the clustering assignments and the consumer outlet classes. The silhouette score is higher for the raw data set, showing a higher quality of clustering, while the other metrics show better results using the filtered data set in terms of the correlations of the identified clusters with the known consumer outlet groups.

Table 2. The evaluation results for the daily consumption patterns.

Evaluation Method	Score (Raw Data Set)	Score (Filtered Data Set)
Silhouette score	0.502	0.455
Rand index	0.293	0.544
Adjusted Rand index	0.001	0.091
Purity	0.339	0.471
Entropy	0.209	0.182

While the two evaluations provide different overviews of the consumption patterns, the K-means clustering evaluation metrics are presented in Figure 6 for determining the best scenario for clustering in terms of overall performance. The visual representation shows the relative clustering quality considering the proposed clustering scenarios for the consumption profile evaluations.

The first two scenarios, i.e., patterns raw and patterns filtered, represent the weekly consumption patterns, while the last two scenarios, i.e., trends raw and trends filtered, represent the daily consumption patterns extracted using the batch processing component. For each type of scenario, the raw data and filtered data were processed to evaluate the clustering quality.

The relative score for each evaluation method was computed based on the deviation from the average score determined from the four analyzed scenarios. It was revealed that clustering works best for evaluating the daily consumption trends, using the filtereddata set.

3.3. *Consumption Event Clustering*

To provide a more in-depth analysis, the individual consumption events were extracted from the raw data set and characterized by their duration (minutes) and overall volume (L). The event processing stage resulted in 9831 individual events generated by the corresponding water outlet.

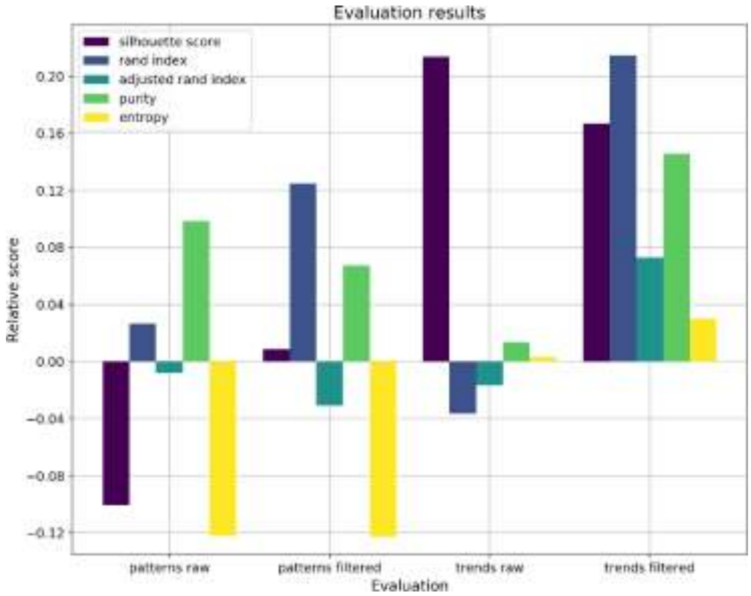


Figure 6. The consumption profile evaluation results.

In Figure 7, the consumption event clusters for sink water outlets can be characterized by their duration and total volume according to the level of variability as follows: (1) the purple cluster shows low consumption events having moderate variability in terms of duration and lower variability in volume; (2) the red cluster shows moderate consumption events having moderate variability in both parameters; and (3) the green cluster shows high consumption events having high variability in both parameters, with many outliers.

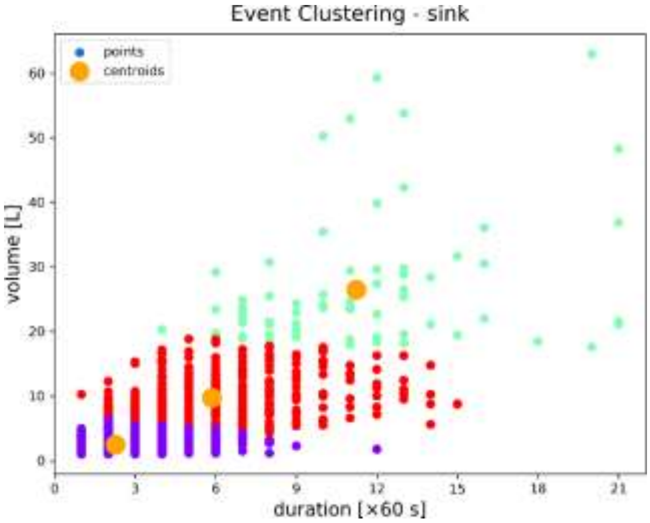


Figure 7. Consumption event clustering—sink.

The sink water data set includes the combined measurements obtained from cold water and hot water outlets to provide more consistent results in terms of the overall water consumption. Otherwise, the water temperature can be adjusted based on consumer preferences, resulting in variable consumption patterns for the hot versus cold water taps, influenced by the required water temperature, which is less relevant for this study. Nonetheless, the installation of different sensors for hot and cold water was necessary from a practical perspective, as these were installed directly on the flexible water supply tubes that lead to the sink in order to be less intrusive for the end user.

In Figure 8, the consumption event clusters for toilet water outlet can be characterized by their duration and total volume, according to the level of variability as follows: (1) the green cluster represents low consumption events having a moderate level of variability in terms of both duration and volume; (2) the purple cluster shows moderate consumption events having a low level of variability in both parameters; and (3) the red cluster shows high consumption events having high levels of variability in both parameters, with many outliers.

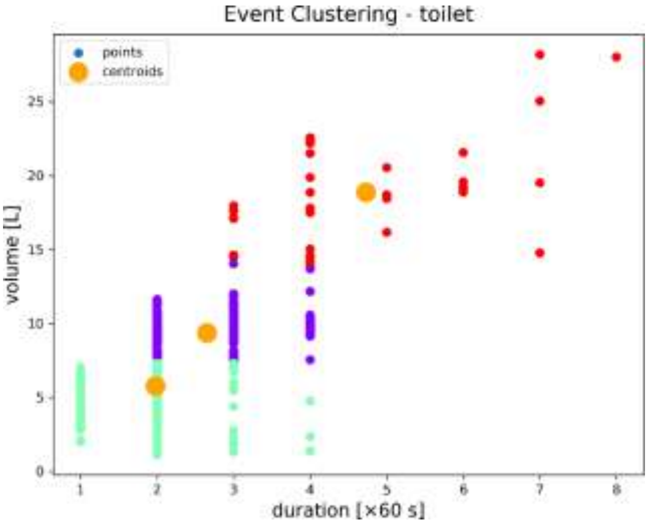


Figure 8. Consumption event clustering—toilet.

In Figure 9, the consumption event clusters for shower water outlet can be characterized by their duration and total volume, according to the level of variability as follows: (1) the green cluster represents low consumption events having low variability in terms of both duration and volume; (2) the purple cluster shows moderate consumption events having high variability in both parameters; and (3) the red cluster shows high consumption events that can be considered outliers.

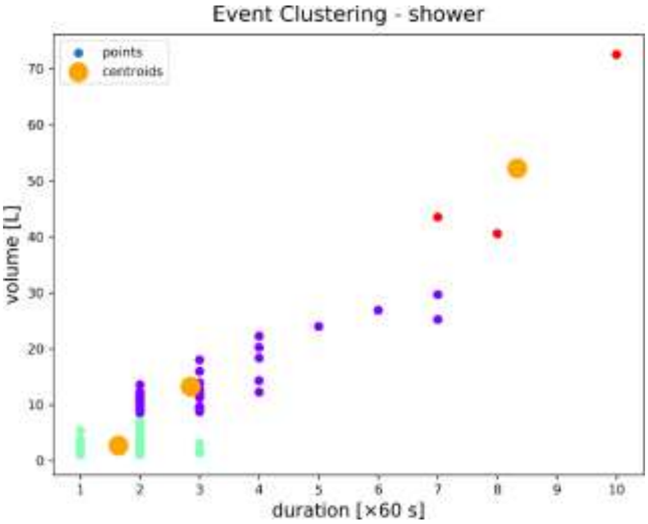


Figure 9. Consumption event clustering—shower.

To evaluate the variability/dispersion for the analyzed consumer outlet types, a visual overview is shown in Figure 10, with the average results ordered by the highest within-cluster dispersion as follows: shower (79 L), sink (46 L), toilet (5 L). This evaluation shows that shower events are characterized by generally higher variability, while toilet events are characterized by generally lower variability. This observation suggests that consumption events could be identified from an unknown data set based on the level of variability (i.e., high variability for shower data, moderate variability for sink data, and low variability for toilet data).

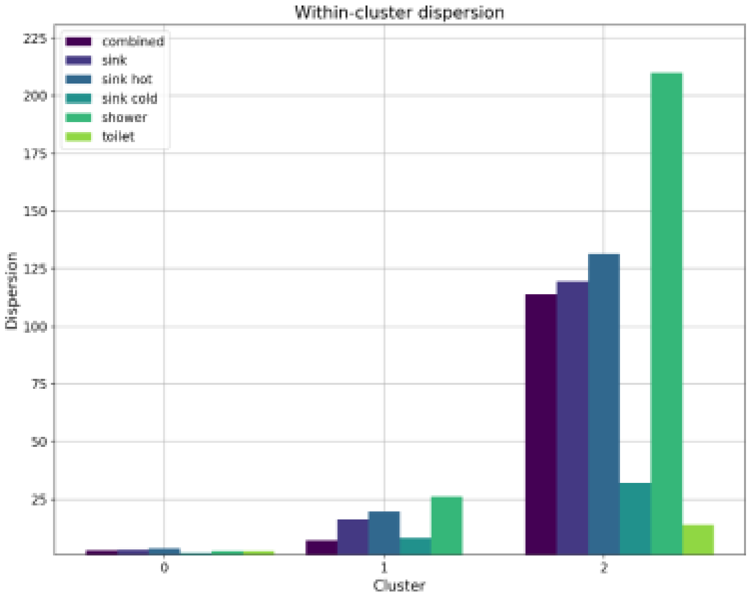


Figure 10. The evaluation of within-cluster dispersion.

To evaluate the possibility of accurately identifying the consumer outlet types from a mixed data set, we performed the clustering of consumption events using three clusters representing the three types of outlets, i.e., shower, combined hot/cold water sink, and toilet and analyzed the variability in each cluster. In Figure 11, the consumption event clusters for combined water outlets can be characterized by their duration and total volume, according to the level of variability as follows: (1) the green cluster represents low consumption events having a low level of variability in terms of volume and moderate level of variability in terms of duration; (2) the purple cluster shows moderate consumption events having a moderate level of variability in both parameters; and (3) the red cluster shows high consumption events having high levels of variability in both parameters, with many outliers.

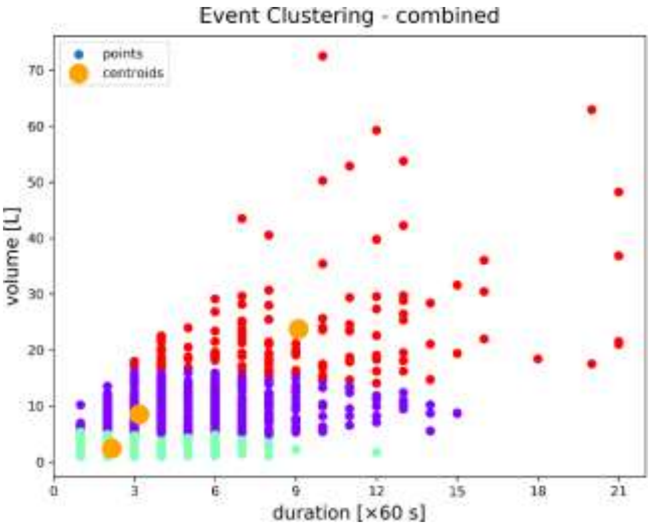


Figure 11. Consumption event clustering—combined.

The results presented in Figure 11 emphasize three levels of variability that can be assigned to the previously identified levels (i.e., green cluster showing low variability, purple cluster showing moderate variability, and red cluster showing high variability). To demonstrate this assumption, the clustering performance was analyzed using the same evaluation metrics, with the results presented in Table 3, showing a high level of similarity between the clustering assignments and the actual consumer outlet types.

Table 3. The evaluation results for the event clustering.

Evaluation Method	Score
Slihouette Score	0.590
Rand index	0.678
Adjusted Rand index	0.356
Purity	0.792
Entropy	0.068

Therefore, the accuracy for identifying consumer outlet types is higher by clustering consumption events rather than the raw time series.

3.4. Consumption Event Classification

Due to the potential interdependency between hot and cold tap water caused by variable water temperatures, a single class was considered for the sink water outlet. Moreover, because the available data were unequally distributed into the multiple consumer outlet types (i.e., sink cold, sink hot, toilet, shower), the classification problem was limited to using two target classes (i.e., combined hot/cold water sink and toilet) to provide consistent results.

Therefore, the main objective of the supervised learning in this study was to train the classification models to predict between sink and toilet consumption events from the water consumption data set, which accounted for 9414 consumption events. An additional balancing of the data set resulted in 4785 consumption events, which were used for training the classification models.

The classification accuracy is shown in Figure 12 for each model. The machine learning models, i.e., DT and RF, can be characterized by lower prediction accuracy (i.e., 91.84% and 92.46%), which combined with a higher accuracy on the training data set (i.e., 96.31% and 96.40%) are indicators of higher overfitting. The deep learning models, i.e., Dense and RNN, on the other hand, show higher accuracy for prediction (i.e., 93.52% and 94.62%), which can be attributed to lower levels of overfitting and superior prediction capabilities.

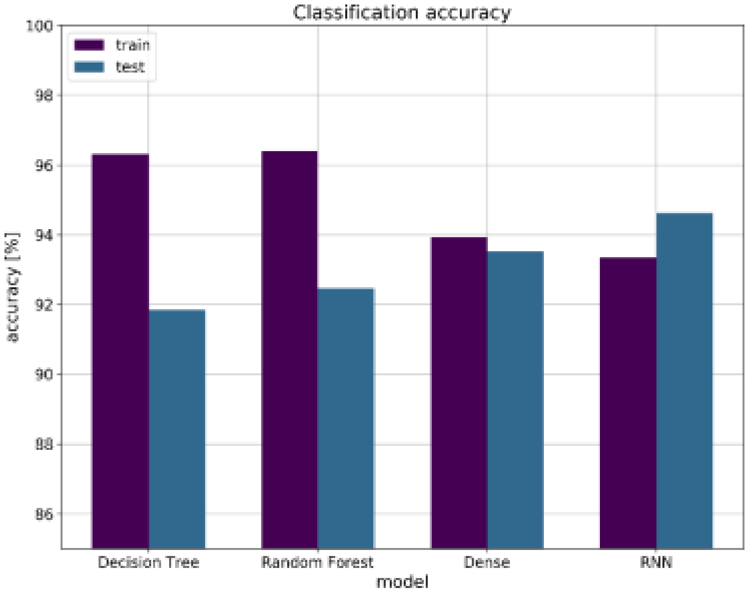


Figure 12. Consumption event type—classification accuracy.

4. Discussion

As presented in the results section, the results are promising in terms of predicting and identifying consumer outlet types of the four measurement points (hot water sink, cold water sink, toilet, and shower). Data were collected from 33 sources representing various water outlets for one week with a sampling time of 60 s, which allowed for a high level of detail in terms of consumption patterns and events.

In the first stage, the K-means clustering algorithm and evaluation metrics were applied to observe the consumption patterns in terms of variability. Due to a moderate similarity between the extracted clusters and the consumer classes, we extracted daily consumption patterns for a better understanding of the consumers. During this stage, clustering on daily consumption patterns and evaluation metrics was applied and showed better results in terms of similarity, which was higher for the filtered data set.

For the better understanding of the daily consumption patterns, the time series were extracted from the raw data set, while moving average filtering was applied for a better overview. Next, K-means clustering was applied to extract the daily consumption patterns. For a numerical evaluation of the results obtained by filtering and daily consumption, the five metrics were applied again, showing better scores on the filtered data set compared with the raw data set.

Given that the two approaches offer different results in terms of extracting consumption patterns, a graph was made containing the evaluation metrics, where the relative score for each method was calculated based on the deviation from the average score. It turned out that the clustering method applied on the filtered data set offers the best results.

Going in depth with the analysis stages, the individual consumption events were also extracted from the time series. An initial limitation was caused by the separation of hot and cold tap water, and the accuracy was not so encouraging. This is because consumers' preferences can be very varied in terms of water temperature. As a solution for the sink data to be more consistent, and for the study to be relevant, the hot water tap and cold water tap were combined, making it easier to evaluate the entire sink data set. The results show moderate variability for sink events and high variability for shower events, while the toilet registers small variability; toilet water consumption is mostly constant because the water tank has the same volume each time and requires the same period of time to be refilled.

Next, the clustering of the events extracted from the time series was performed using three clusters corresponding to the three consumer outlet types, i.e., sink, shower, and toilet, to analyze the variability. Three levels of variability could be visualized that matched the characteristics of the three consumption types; we concluded that this scenario is the most suitable for evaluating consumption events on a data set. Furthermore, the evaluation metrics showed a high similarity between the clustering assignments and the known water outlets.

The final evaluation involved testing the machine learning and deep learning algorithms to predict consumption events. For this stage, only two of four events were chosen, i.e., hot water and cold water events taken together and toilet events, these being the most consistent in terms of variability. The machine learning models offer comparably higher accuracy for training and lower accuracy for testing when compared with the deep learning models, which have higher accuracy for prediction.

5. Conclusions

Monitoring water consumption is essential today for predicting leaks and eliminating water waste that can lead to water scarcity. The current paper proposes the evaluation of several methods for predicting water at consumer outlet types: hot tap water sink, cold tap water sink, toilet, and shower.

The proposed architecture is based on scalable components and requires minimal configuration, which makes it possible to extend the study to multiple households. The proposed configuration using separate sensors for each consumption outlet allowed for evaluating different methods for characterizing consumer behavior in great detail.

Moreover, the same methods can be applied to extract the consumption patterns from combined measurements, i.e., using sensors installed at the mains, and predict the consumption activities based on the overall water consumption data. Possible extensions include location-based clustering and demographic analysis in large-scale deployments.

In contrast with other studies in this field, the four types of household activities were analyzed in this study, applying clustering, classification methods, and evaluation metrics. In the case of clustering methods, high accuracy was obtained by extracting the consumption events from the time series, as confirmed by the evaluation metrics. In the case of classification methods, using both machine learning algorithms and deep learning algorithms, good results were achieved in terms of prediction accuracy.

As in any study, there were some limitations. First, the solution could be extended to more households and the installation of additional sensors to improve the accuracy of the results. Second, the installation process proved to be quite tedious due to the wiring of the flow sensors, which have to be connected to the IoT boards. Another limitation is the high volume of data that is collected regardless of the actual water consumption, which adversely impacts the data access.

As future work, the results can be used to define a decision support system that achieves multiple objectives for improved efficiency of water resource management, i.e., shaping consumption patterns to maintain a constant flow, which can be especially beneficial for central heating systems; showing consumption statistics for household consumers; comparing consumption events between households; and promoting consumer involvement in regulating water flow and overall resource consumption.

Author Contributions: Conceptualization, D.A. and A.P.; methodology, D.A., A.P., C.-O.T. and E.-S.A.; software, A.P.; validation, A.P.; formal analysis, D.A., C.G.C., C.-O.T. and E.-S.A.; investigation, D.A.; resources, A.P.; data curation, A.P.; writing—original draft preparation, D.A., A.P., B.P., C.-O.T. and E.-S.A.; writing—review and editing, C.G.C.; visualization, A.P.; supervision, C.G.C.; project administration, C.G.C.; funding acquisition, C.G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by UEFISCDI, grant number PN-III-P2-2.1-PED-2019-4993, Smart Urban Water-Based on Community Participation Through Gamification—Watergame Project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the GitHub repository referenced in Section 3.

Conflicts of Interest: The authors declare no conflict of interest.]