Retraction

Retracted: Real-Time Water and Electricity Consumption Monitoring Using Machine Learning Techniques

Shariq Bashir

IEEE Access

10.1109/ACCESS.2023.3241489

<p>Notice of Retraction S. Bashir, “Real-Time Water and Electricity Consumption Monitoring Using Machine Learning Techniques,” IEEE Access, vol. 11, pp. 11511–11528, 2023, doi: 10.1109/ACCESS.2023.3241489.</p> <p>After careful and considered review of the content of this article by a duly constituted expert committee, this article has been found to have violated IEEE publication principles. Specifically, this article copied portions of content from the following article without appropriate reference:</p> <p>Aida Boudhaouia, “Analyse, classification et prédiction de consommation d’eau et d’électricité par des techniques de machine learning”, Thesis, Université de Haute Alsace, January 2022</p> <p>Therefore, IEEE has retracted the content of this article from Xplore. When informed of the retraction, the author did not respond.</p>

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 December 2022, accepted 23 January 2023, date of publication 1 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241489

RESEARCH ARTICLE

Real-Time Water and Electricity Consumption Monitoring Using Machine Learning Techniques

SHARIQ BASHIR
Institute of Applied Data Analytics (IADA), Universiti Brunei Darussalam (UBD), Gadong BE1410, Brunei
e-mail: shariq.bashir@ubd.edu.bn

This work was supported by Universiti Brunei Darussalam (UBD), Faculty Research, under Grant UBD/RSCH/1.11/FICBF(b)/2022/006.

ABSTRACT This work studies the automatic classification of water consumption patterns and electrical devices, both supervised and unsupervised. This involves training machine learning algorithms to identify normal and abnormal water consumption patterns and differentiate between different types of electrical devices. We performed an unsupervised classification of consumer water patterns in direct and indirect ways. The first is to use the raw consumption patterns obtained directly from the server. The second one corresponds to the use of the sampled consumption patterns. This classification is performed using hierarchical bottom-up classification and a self-organizing map. A probabilistic analysis of daily water consumption is performed to extract the percentage of daily consumption with the most information. It enables us to identify water consumption patterns more quickly by reducing the number of data points for each daily pattern, allowing us to recognize and classify anomalous behaviors as soon as feasible. Then, the signatures of electrical devices are classified using three ML algorithms: multilayer perceptron (MLP), k nearest neighbor (KNN), and decision tree (DT). Furthermore, assembly approaches are also studied. These are based on the OAA (One Against All) principle, which presents one class against all other classes, and the ECOC (Error-Correcting Output Codes) philosophy, which allows the classification error to be corrected. According to the bias/variance trade-off, both techniques enhanced classification by ensuring accuracy and generalization. A more in-depth analysis of the properties of the electrical devices is handled with an esemplastic approach based on gradient-boosting decision trees. The features of electrical devices can be extracted using this analysis based on the nature of their harmonic signatures.

INDEX TERMS Water consumption prediction, water forecasting, sustainability cities, electricity consumption prediction, energy forecasting, machine learning.

I. INTRODUCTION

New environmental challenges are emerging in the area of energy and water resource management [1], [2]. To overcome these issues, novel advances must be explored, particularly in the everyday usage and processing of this limited quantity of resources. Thanks to recent advancements in IoT (Internet of Things) and machine learning new systems can be developed to efficiently manage and allocate resources in commercial and residential buildings [3], [4], [5]. The developments in digital technology provide greater opportunities for data collection, enabling closer monitoring of the usage of water and power as well as quicker detection of network problems (leaks, wastage, fraud, etc.). The purpose of this article is to continuously monitor the electricity and water distribution networks in real time.

IoT is the connection of everyday devices to the internet, such as appliances, automobiles, and home systems [6], [7]. This enables these gadgets to connect and be remotely controlled over the internet. The purpose of IoT is to make daily life easier and more efficient by automating various operations and collecting data for better

The associate editor coordinating the review of this manuscript and approving it for publication was Roberto Sacile.

decision-making [8]. This technology automates data collection through the internet. It has made it possible to use Automatic Meter Reading (AMR) techniques to automatically gather water or electricity consumption data and create advanced metering infrastructure (AMI) [9]. The AMI is a crucial technology driving the development of smart buildings [10]. Building managers may monitor consumption in real-time and make modifications to preserve resources and decrease costs by connecting equipment such as water and energy meters to the internet. This data may also be utilized to enhance building systems and boost overall energy efficiency. The next step is to evaluate this data to acquire insights into how the building is used and how resources are consumed. This may be accomplished through the use of sophisticated analytics and machine learning algorithms, which can identify patterns and trends in data and make recommendations for improvement.

The data collected from AMI is mostly represented as time series data, specifically as consumption load curves (CLCs) [11], [12]. A CLC provides a change in consumption of a specific resource (e.g. electricity, water) over time. The CLCs serve as the foundation for further analysis and modeling. They can be used for categorization, grouping, and forecasting. For example, CLCs can be used to identify different types of usage patterns, such as peak usage or low usage, which can be used to optimize building systems and reduce water or energy consumption. We want to use machine learning algorithms to identify patterns and trends in the CLCs data and make predictions about future consumption. One such technology is wireless communication, which allows for the transfer of data from smart meters to a central hub or cloud-based platform. This data can be accessed through a web portal or mobile app, allowing customers to view their usage in real-time and set up alerts if their usage exceeds a certain threshold. By providing customers with real-time access to their water usage data and alerts for leaks, people may become more conscious of their usage and take steps to prevent wasting resources. This can not only help them to reduce their water bills but also conserve a precious resource. The study and comprehension of water and power use through smart metering and IoT technology will also lead to several research avenues. For example, the same technology and methods can be applied to other types of resources such as gas consumption. By installing smart meters and sensors to monitor gas usage, customers can track their usage and detect leaks in the same way as with water. Additionally, the data collected can be analyzed and used to predict future usage patterns, optimize building systems, and improve energy efficiency.

A. MAIN CONTRIBUTIONS

The goal of this article is to examine and assess the use of both supervised and unsupervised machine learning techniques on data represented as CLCs and time series of consumption in order to uncover patterns, make predictions, and support decision-making for water and energy efficiency and smart building operations. We cannot interpret the behavior or dynamics of CLCs of water consumption as a time function based on the acquired CLCs as the consumption patterns may not be predictable or consistent over time. Making groups of CLCs using clustering is an alternative approach to exploring the data, as it can help to reduce the quantity of data and focus on the CLCs that contain the most information. Earlier research on this topic for example in the case of water consumption has focused on predicting water use by integrating other characteristics in addition to consumption. These characteristics include the time of day, climate, residence size, etc. Our article aims to predict the daily water usage dynamics in home and residential installations using historical data. This approach is based on the idea that by analyzing past water consumption patterns, it is possible to make accurate predictions of future water usage. To make this forecast, only one exogenous variable, the operating day is considered implicitly.

The article proposed the first contribution which is to use a clustering approach to group the consumption load curves (CLCs) of water consumption. This clustering aims to categorize each CLC into a specific class based on its daily profile, such as normal/abnormal, activity days/weekends, or vacations. In this context, a probabilistic technique is presented that investigates users' daily water usage patterns and seeks the part of this CLC with the greatest consumption. This method allowed us to classify the CLCs of water usage with less amount of data and low error data. Moreover, by using the CLC portions that contain the most consumption, it is possible to classify daily consumption in real-time throughout the day and alert as soon as possible in the event of abnormal consumption. With a case study of the water usage of a residential unit, we undertook a more thorough analysis of the abnormal consumption. This case study aimed to better understand the causes of abnormal water consumption patterns, and how they differ from normal consumption patterns.

Then, we provide our second contribution by classifying electrical devices based on harmonic features inferred from aggregate load current. A variety of machine learning (ML) approaches including the k-nearest neighbor method, a decision tree (DT) that resembles CART, and multi-layer perceptron (MLP) are developed and assessed to classify electrical devices. These approaches entail classifying eight electrical devices based on their harmonic signatures, represented by 16 features. The results indicate that the DT produced the optimal results. However, if the data is substantial and the tree is too large and deep, there is a risk of overlearning. To handle this problem we presented two ensemble approaches that use OAA (One Against All) and ECOC (Error-Correcting Output Codes) with MLP. By comparing OAA and ECOC to the traditional ML approach (DT, KNN, and MLP), we were able to analyze their effectiveness. This comparison also allowed us to understand the strengths and limitations of the ensemble methods. In the case of our application, the OAA ECOC performed better than the MLP, DT, and KNN models. We also compared the performance and accuracy of OAA with ECOC. We found that OAA outperformed all other classifiers in terms of performance.

To streamline the computations, another contribution that we proposed is based on minimizing the number of features of the electric devices. This is achieved through analyzing and extracting relevant features. To assign relevance scores to features, a gradient-boosting method was presented. It is based on several decision trees. Gradient boosting is a powerful ML technique that improves the accuracy of a model by combining multiple weak learners into a strong ensemble. Four distinct datasets, noisy, random, normal, and tampered, are used to test this technique. After reducing features, we found that the accuracy of classification and the time required to learn and perform the classification were improved.

II. RELATED WORK

Short-term forecasting of usage of water [13], [14], electricity [15], [16], or gas [17] have been reported in the literature using a variety of methodologies and horizons. However, only a small percentage of them have provided high-resolution service to individual customers in residential buildings [18]. The technique proposed in [19] utilizes a non-homogeneous Markov chain model, allowing for an understanding of the water consumption dynamics. This model can predict daily consumption patterns based on additional inputs, such as external factors like weather conditions [20], the day of the week, etc. A different study [21] concentrates on estimating water demand on a weekly and hourly basis using an autoregressive model that is based on a time-based periodic component of time series data to refine daily demand estimates and hours. Many different period models are employed in this forecast. The majority of these studies aim to anticipate consumption by including supplementary parameters through various predictive models that are chosen based on the type of input data and the outcomes sought. Indeed, we observe that the input databases of models play a significant role in the offered forecast horizon. Typically, these databases feature annual, seasonal, monthly, weekly, daily, or hourly resolutions. Even when using intelligent techniques, most of the task depends on extra knowledge. Support vector machines are used in [22] to analyze monthly water usage, user counts, and overall water bill costs. Reference [23] discussed regulating domestic water use depending on costs, restrictions, weather, and demographics. Currently, no study uses learning architectures like Hopfield networks, LSTM, direct or recurrent BPNN, or direct or recurrent BPNN to forecast water demand based solely on historical data from a single measurement point. On the other hand, we suggest more accurate forecasting using high-resolution data from smart meters without any additional contextual information. This article focuses on predicting water consumption from a private building without knowing the number of occupants or the devices that use electricity.

TABLE 1. List of notations and symbols used in the article.

Symbol	Description of Symbol
R	Raw data
W	Consumption load curve after correction
C	Unevenly spaced time series
e	Coefficient
LC	Cumulative load curve
CLC	Consumption load curve
CRC	Cumulative reference curve
DC	Electrical device class
OAA	One against all
$ECOC$	Error-correcting output codes
$NILM$	Non-Intrusive Load Monitoring

III. WATER AND ELECTRICITY CONSUMPTION DATA COLLECTION

The section outlines a comprehensive smart meter data collection infrastructure. Figure 1 shows a graphical representation of data collection infrastructure. In our case study, multiple smart meters were installed in a residential building. With the aid of this technology, data collecting is automated, and a database is used to store all data. The AMI (Advanced Metering Infrastructure) that is part of the IoT platform enables non-intrusive continuous monitoring of usage from a single central collection point. This platform enables utilities to monitor consumption patterns, detect abnormal usage, and predict future consumption. In real-time operational conditions, data on water and electricity usage is transmitted to a web server. A time series and CLC are used to transform the data that came from a measurement point. Data shortages might emerge from anomalies that arise during data gathering. To effectively utilize the data, it is also necessary to analyze and condense a large amount of data into meaningful patterns of consumption, both on a daily and hourly basis. This process helps identify key usage trends and patterns that can be used for forecasting and resource management.

The majority of the information in the database comes from smart water and energy meters [24], [25], [26]. Smart meters are able to collect and transmit consumption data to a central location, allowing for real-time monitoring of usage. They also have the ability to detect and diagnose problems in the network, such as leaks or tampering, which can lead to cost savings and improved efficiency in the distribution system. As a result, the use of smart meters and the collection of data from a central location allows for more accurate and efficient monitoring of water and energy consumption in residential and commercial settings. Furthermore, smart meters are designed to provide accurate and real-time information on water and energy consumption, which can be used for a variety of applications such as monitoring, billing, and load management. The built-in computational capabilities of smart meters allow for the deployment of advanced techniques for energy management, measurement processing, data compression, message transmission, and local interaction with the user and the environment [27].

A. CONSUMER CONSUMPTION EVENTS

Consumer consumption events refer to the occurrence of a unit of consumption, such as one liter of water or one watt-hour (Wh) of electricity. A strategy is proposed in [28] which transmits all raw data to the server allowing for detailed analysis of water and electricity consumption. The raw data are specified as time-stamped events occurring at times $t_1, t_2, \ldots, t_n$. These events are triggered whenever there is a change of one liter, or one Wh, in the measured quantity. In the case of a water meter, the consumption event corresponds to the flow of water measured by the meter, which is based on the meter's size and sensor specifications. This variation is fixed and does not change. In our situation, a DN15 meter has a one-liter flow accuracy. With a larger meter (DN100), the water flow is ten liters [28]. This allows for more accurate and frequent data collection, as well as the ability to detect and respond to changes in consumption patterns in real time. For the case of electricity consumption, the meters have an accuracy of 0.1 and 1 Wh, and the collection scripts are designed to gather and transmit this data to a central server for storage and analysis.

B. CONVERTING RAW DATA TO MEANINGFUL DATA

The gathered information is really valuable and has to be examined. To achieve this, it is necessary to understand the raw data and link it to various theoretical frameworks and models. Unevenly spaced time series, sometimes known as consumption load curves (CLCs), is among them.

1) CONVERTING LOAD CURVE TO TIME SERIES CURVE

Time-stamped events refer to a series of numerical observations that are chronologically ordered in time according to the occurrence of each event. These events reflect the temporal variation of consumption data and can be used to analyze patterns and trends in the data. This is known as a time series curve with unequal spacing, denoted by C in Equation 1 [29]. An event in the context of water consumption corresponds to each consumed liter. Thus C is a series of scalar values of the incremented variable $R_{i+1} = R_i + 1$. The raw data for one smart meter that was taken from the platform previously mentioned is represented by the value C. It is the end result of an observed process during a time period T. The platform and AMI suggested by [9] provide the ability to record the exact moments that each liter is consumed.

$$C = [R_1\left(t_1\right), R_2\left(t_2\right), \ldots, R_T\left(t_T\right)] \tag{1}$$

2) CUMULATIVE WATER CONSUMPTION

A smart meter index is represented by each R_i and is the total amount of water used at each instant t_i. The time interval between t_i and t_{i-1} is not constant. A cumulative load curve (LC) is the evolution of R_i over a time period T. Examining and comparing consumption across years, months, weeks, and days is made easier with the help of the load curve. Then we address LC on a monthly, weekly, and daily basis.

3) SAMPLED WATER CONSUMPTION

To make the obtained data compatible with the machine learning algorithms we used a sampling strategy. The sampling strategy refers to a method that maintains the obtained data in a regular time-spaced interval for the purpose of analysis. It is an important aspect of machine learning as it can affect the accuracy and representativeness of the results. The sampling can be done in minutes or hours, yielding 1440 or 24 data points daily. We also choose to compute the LC by processing sequences of n data, where n is the number of liters used each minute or hour. Therefore, a chronological time series with an implicit sampling of the order of appearance is as follows:

$$W = \lfloor R_1, R_2, \ldots R_n \rfloor \tag{2}$$

C. DATA INTEGRITY AND PREPROCESSING

The data's integrity must be validated under real-world working situations. The database may contain missing raw measurements as a result of errors or malfunctions. As a result, we suggest a raw data preprocessing stage to check the data and eventually complete missing data through interpolation. Figure 1 depicts the whole proposed preprocessing technique. The database is used to retrieve the raw time series for every day. The data are gathered at a precision of minutes since a one-hour accuracy is preferred for a forecast of water usage (i.e., 1440 minutes per day). Daily preprocessing is carried out in this manner. Interpolation is used to identify and correct any periods during which no consumed liters events occurred.

Machine learning methods must be used for forecasting and data analysis with no missing or inconsistent results. It is therefore critical to differentiate and separate abnormal consumption from typical and regular consumption. Abnormal water consumption is always the outcome of unusual and infrequent user behavior. [30]. The detection of anomalous water usage is accomplished in the following manner. A cumulative reference curve (CRC) is calculated for each day of the week. The CRC is computed for each day with a maximum and minimum LC. Typically, there is a significant correlation between the load profile for one day j and the day before $(j-1)$ and the prior week's day $(j-7)$ [31]. The CRC is computed as follows:

$$W_j\left(t_i\right) = \text{avg}\left(W_{j-7}\left(t_i\right), W_{j-1}\left(t_i\right)\right) \tag{3}$$

Equation 4's average $avg()$ function is replaced with the $min()$ and $max()$ functions to calculate the minimum and maximum LCs for each day in the same manner. The following criteria is used to detect normal consumption:

$$\text{abs}\left[R_j(t) - \text{avg}\left(\sum_{i=1}^{n}\left(R_j(t)\right)\right)\right] \geq \alpha \times \text{std}\left(\sum_{i=1}^{n}\left(R_j(t)\right)\right) \tag{4}$$

where α is an empirically selected numerical variable, in experiments we used $\alpha = 5$. For each value of the LC, $std()$ represents the standard deviation. Additional tests can be run to determine if the instantaneous consumption is outside the

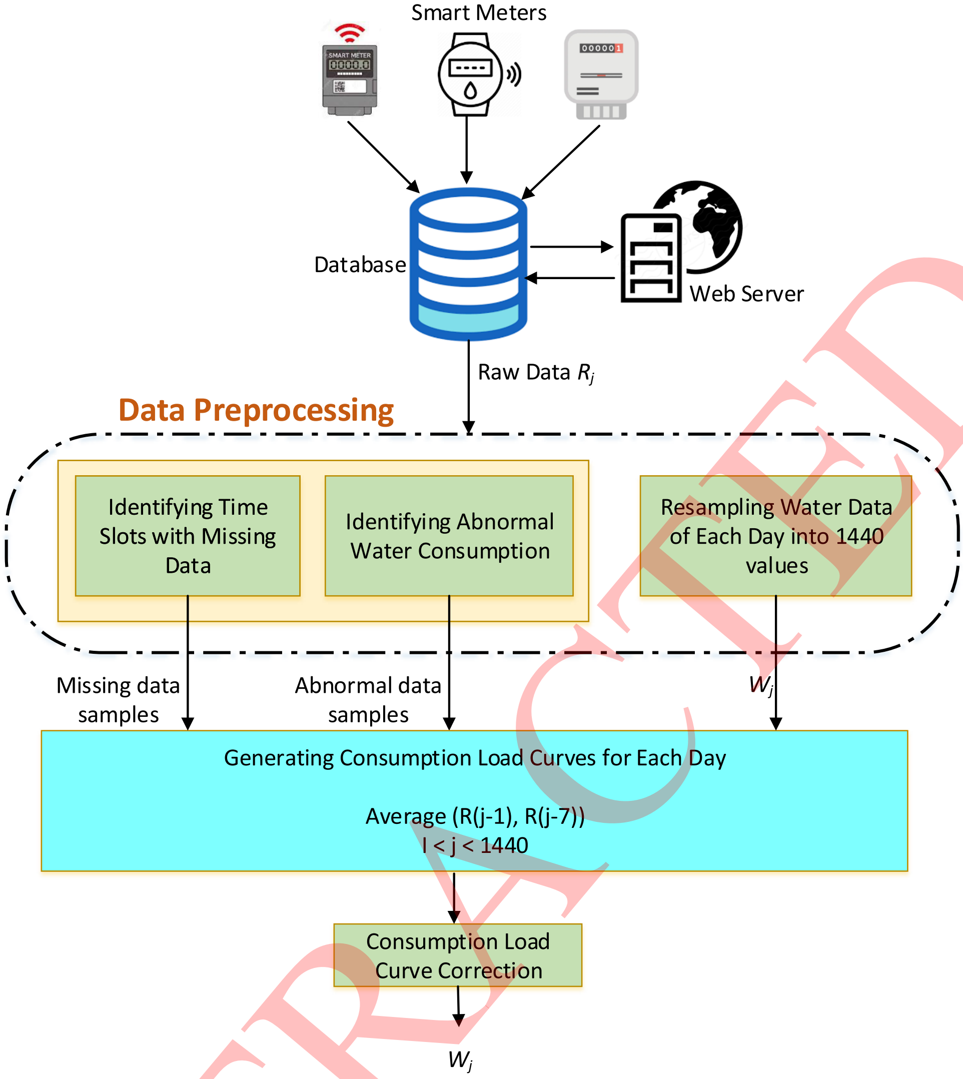

Smart Meters
SMART METER
000001
Database
Web Server
Raw Data R_j
Data Preprocessing
Identifying Time Slots with Missing Data
Identifying Abnormal Water Consumption
Resampling Water Data of Each Day into 1440 values
Missing data samples
Abnormal data samples
W_j
Generating Consumption Load Curves for Each Day
Average (R(j-1), R(j-7))
l < j < 1440
Consumption Load Curve Correction
W_j

FIGURE 1. Architecture to convert raw data to consumption load curve.

bounds set by the lowest and highest LCs for the same day of the week and to spot any additional consumption that significantly deviates from the average consumption [32]. It is important to note that only data from water usage and a few statistical markers are used to identify abnormal and unexpected consumptions [32]. Consumptions odd or unusual, are adjusted by interpolation during their duration and are not considered in the learning procedures. Finally, we have a time series $\hat{W}_j$ sampled in minutes that matches the LC (W_j) without data loss and without abnormal and unusual consumptions.

D. FINAL CONSUMPTION DATASETS

For the categorization and prediction of electricity and water use, we created two datasets. To forecast the amount of consumed water liters in the upcoming hour, a four-month database (from February 2022 to May 2022) is used. This consumption was measured from a domestic apartment in Brunei Darussalam, which 47 families occupied in 47 units, each consuming 1034 liters per day on average. The electricity database is used to categorize eight different electrical devices. Each sample of the database corresponds to harmonic signatures of electrical devices. The harmonic signature is used to identify the device's characteristics and distinguish it from other types of devices. A harmonic signature of an electrical device refers to the unique pattern of harmonics present in its current waveform. The harmonic signature can provide information about the electrical properties of the device, its operating conditions, and any potential problems. The measurement of the harmonic signature is usually performed by analyzing the current waveform of the device and determining the presence of harmonics and their amplitudes. Only Odd harmonics are relevant in terms of amplitude because they can contribute to the total harmonic distortion (THD) of a current waveform. THD is a measure of the deviation of the waveform from a pure sine wave and can indicate the presence of harmonics in the current. The amplitude of odd harmonics: 1 (fundamental frequency), 3, 5, 7, 9, 11, 13, and 15 is important to consider when evaluating the quality of the power being supplied to an electrical system [33]. As a result, the feature vector has a total of sixteen features [34] for the devices under consideration. In Figure 9 harmonic signatures of eight electrical devices are graphically presented.

As electrical devices are recognized by their electrical signatures. Therefore, it is possible to monitor a device's electrical consumption in a variety of situations, including noisy, normal, and random ones. Through this, we constructed a database that we call the global signature database. It contains 40k samples. With sixteen characteristics per sample where each sample acts as the device's distinctive signature are used as the only inputs for training the classifiers. Random 67% and 33% samples from the global signature database were obtained for the model's training and testing. The electrical devices are classified into the following eight classes: C1 stands for inverter air conditioner, C2 for bulb, C3 for TV, C4 for battery charger, C5 for lamp, C6 for computer, C7 for electric geyser, and C8 for refrigerator. To further investigate the categorization task, this database is derived into four sub-datasets, each with 10,000 samples. These sub-datasets are explained in Table 2.

IV. CLASSIFYING WATER CONSUMPTION CLCs

This section investiages an unsupervised categorization of CLCs for water utilization. The CLCs of the database discussed in Section III-D are first analyzed using the Pearson Correlation Coefficient. Only 100 days with water consumption greater than zero are considered. Each CLC is sampled, resulting in 1440 data points per day for each CLC. The correlation $corr(W_{avg}, W_j)$ and the distance $dist(W_{avg}, W_j)$ are two tools for determining the similarity and resemblance of two vectors (W_{avg} and W_j) which refer to reference CLC and daily CLC respectively. The correlation is computed using the following equation.

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}, \tag{5}$$

$$\mathrm{cov}(X, Y) = \frac{1}{N} \sum_{n=1}^{p} \sum_{j=1}^{q} n_{ij} \left(x_i - \bar{X}\right) \left(y_j - \bar{Y}\right) \tag{6}$$

The covariance between Y and X is represented by $cov(X, Y)$ represents. The average values of Y and X are $\overline{Y}$ and $\overline{X}$, and the standard deviation of Y and X is represented by σ_y and σ_x. Figure 2 displays the outcomes for a few days. We can see that just one correlation (day 06/03/2022) with W_j is less than 90%. This approach attempts to find days with daily CLCs that deviate considerably from the reference CLC. Although it is simple, but if a database has a lot of CLCs, it is not an effective approach. Furthermore, because of the diversity and the selection of the reference CLC, the interpretation of the data is its major drawback.

A. CLUSTERING WATER CONSUMPTION CLCs USING HIERARCHICAL AGGLOMERATIVE CLUSTERING (HAC)

Hierarchical clustering is one of the most extensively used algorithms for unsupervised clustering. However, its quadratic computational complexity limits it to small data sets. HAC is applied in an agglomerative manner. The HAC begins with single individuals, which are then grouped into subsets, and so on. Unsupervised classification can be used to find clusters using this technique. The HAC computes a distance matrix (Euclidean Distance (ED) or Dynamic Time Warping (DTW)) for each pair of data, such as a pair of CLC. It attempts to construct classes at each stage by aggregating each pair of series of values that are closest in distance to the previous grouping. The consecutive groupings are represented by a binary tree known as a dendrogram [35].

We start this section by proposing clustering techniques for CLCs by three direct and indirect clustering. A cluster in unsupervised machine learning is a group of similar data points. In unsupervised learning, the algorithms are used to find the inherent structure in the data without any prior knowledge of the target variables or class labels. Raw (unsampled) CLCs are grouped together with a process called direct clustering. Then, sampled data are created by indirect clustering, which uses data collected at different periods. To carry out this form of clustering a number of classes must be pre-initialized. The needed number of clusters is calculated using a statistical approach of deviations. We used kmax = 10 to find the optimal number of clusters using gap statistics [36]. Six classes, the optimum number were produced by using this strategy. Thus, six classes are used as parameters for clustering.

B. HAC CLCs CLUSTERING WITH EUCLIDEAN DISTANCE (ED)

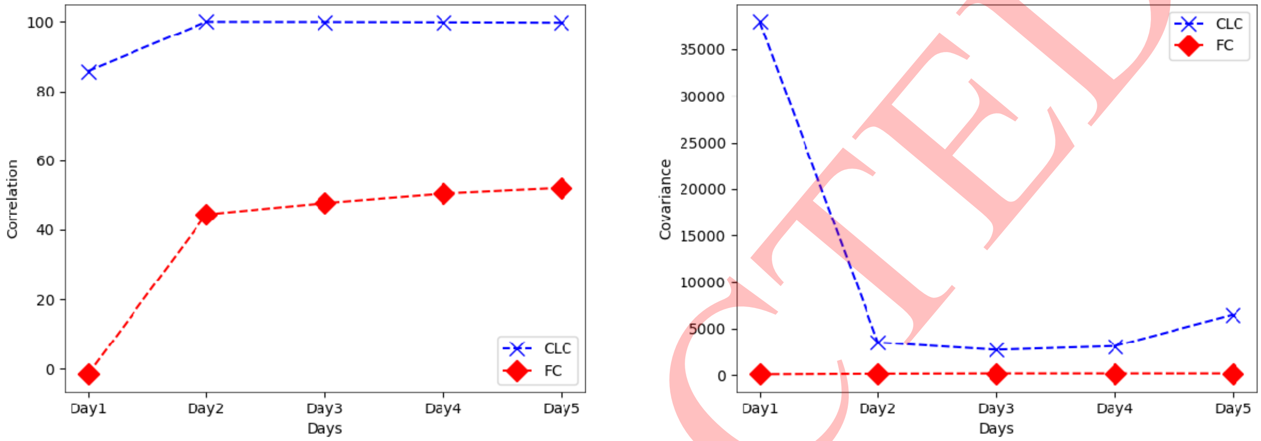
The CLCs should be the same length to achieve a HAC with ED. Therefore, the CLCs to be clustered are sampled. The gathered data demonstrates customers' water use patterns. To the best of our knowledge, there won't be any significant water usage within a minute. Based on this, a granularity of one minute is selected. This results in 1440 data points collected daily from the 1440 minutes. This makes calculating the ED easier. Figure 3 shows the CLCs with ED for the various types of water use. From the results, we can see that class 2 accounts for 71.11 percent of CLCs, which demonstrates that this class includes the majority of water consumption behavior. The CLCs with leakage fall under classes 6, 5, and 1. We want to emphasize once again how this clustering approach helped us identify the CLCs of water use that had different behaviors. The exceptionally high EDs when compared to the other classes show this to be the case. The CLCs for water use over the weekend are assigned to class 4.

C. HAC CLCs CLUSTERING WITH DYNAMIC TIME WARPING (DTW)

By reducing CLC disparities and temporal delays, DTW allows for increased attention to be paid to the diverse

TABLE 2. The four variations (sub-datasets) of global signature database. Each dataset contains 10,000 samples.

Sub-Datasets	Description
Normal	The normal dataset contains raw harmonic signatures. The signatures are directly obtained from the Fourier transform of the current.
Noisy	This dataset is generated by inserting 10% white noise. The dataset contains sixteen features of the harmonic signature.
Random	This dataset contains random harmonic signatures. The random features make it very challenging to establish a relationship between the sixteen features and the classes to which the devices belong.
Tampered	This dataset contains tampered signatures. The constructed tampered signatures are designed so that only one of the sixteen features is significant and directly reveals the category of the device.



Correlation

Covariance

FIGURE 2. The graphs show the covariance and correlation of a few flow curves (FC) and water consumption CLCs. These are calculated using the average flow curve and the CLC, respectively.

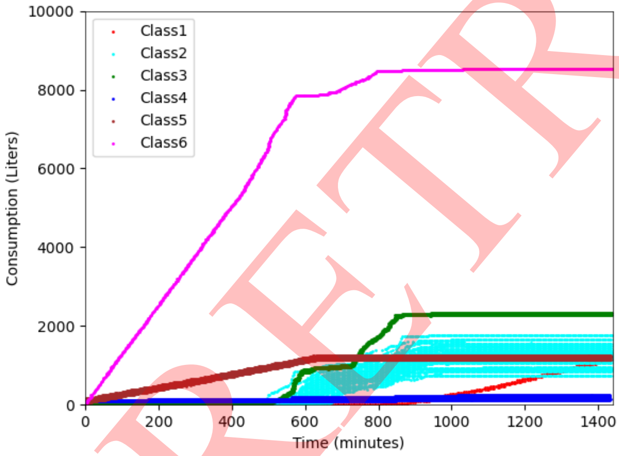


FIGURE 3. Consumption load curve (CLC) classification with HAC approach using euclidean distance (ED) with 6 classes.

consumption patterns throughout the day as well as the sequencing of each liter consumed. In fact, the DTW distance emphasizes behavior and how it is organized in time, minimizing the importance of temporal differences. With this method, clustering can be done without sampling the data because the DTW distance emphasizes the shape and internal slopes of each CLC.

D. CLUSTERING CLCs USING SELF-ORGANIZING MAP (SOM)

Applying the SOM concept is another method for categorizing gathered CLCs [37]. To the best of our knowledge, this approach is rarely used for clustering series of values such as the CLCs of water use in our case study [38]. A SOM is a type of artificial neural network used for unsupervised machine learning. With SOM the input space of the training samples is discretized and represented by a single layer of artificial neurons in this neural network, which is why it is known as a map. Using a neighborhood function, the placements of the neurons are changed throughout training in order to retain the topological characteristics of the input space. Each neuron in the classification process represents a class, and the data is then grouped based on the similarity of the neurons. In our example, CLC classes are created using 1440 data points taken from the same database and a SOM map of 6 neurons (equivalent to 6 classes). Figure 4 clustering results with SOM.

E. EVALUATING CLC CLUSTERING APPROACHES

Table 4 displays the distribution of CLCs of water consumption into six classes, using three different approaches. The CLCs are assigned a number between 1 and 100. A crucial step in evaluating the clustering approaches is analyzing the

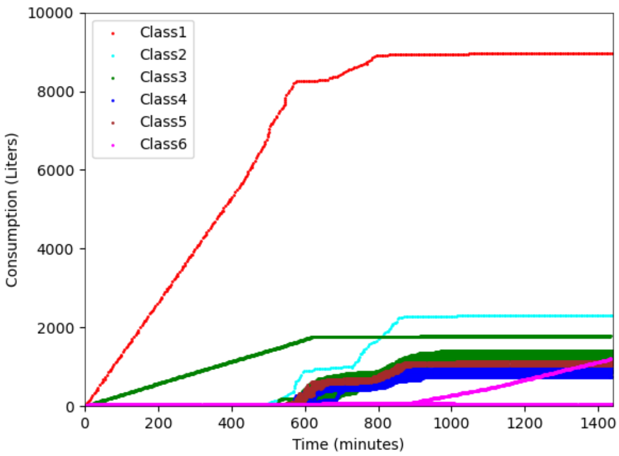


FIGURE 4. **The classification of water consumption load curves (CLCs) into 6 classes using the self-organizing map (SOM).**

similarity of the samples. For each pair of sequences, the similarity metric gives an absolute similarity value. An impartial way to contrast different clustering approaches is to use the cophenetic correlation coefficient [39]. The cophenetic correlation coefficient is a measure of the similarity between the original hierarchical clustering solution and the clustering solution obtained by a different method [40]. The cophenetic correlation coefficient compares the pairwise distances between observations in the original clustering solution to the pairwise distances between the observations in the new clustering solution. A value of 1 indicates a perfect match between the two clustering solutions, while a value of 0 indicates no correlation between the two solutions. It is often used as a way to evaluate the performance of different clustering techniques.

$$e = \frac{\sum_{i<j}(\text{ Distance }(i,j) - \bar{d})(e(i,j) - \bar{e})}{\sqrt{\left(\sum_{i<j}(\text{ Distance }(i,j) - \bar{d})^2(e(i,j) - \bar{e})^2\right)}} \tag{7}$$

The $e(i,j)$ is the cophenetic distance between x_j and x_i. The dissimilarity matrix calculates the distance between x_j and x_i as $Distance(i,j)$. The mean of $e(i,j)$ is represented by $\bar{e}$ and the mean of $Distance(i,j)$ is represented by $\bar{d}$.

The Table 3 shows the cophenetic correlation coefficient that has been calculated for daily CLC and DTW, ED, along with the time required to compute the CLC distances. The clustering with ED is used as a baseline to calculate the cophenetic correlation coefficient. Now, if we contrast the grouping produced by the SOM map and the one produced by ED. Figure 4 clearly shows that there are CLCs in classes 3 and 5 that are entirely distinct from the rest of the CLCs in the same class. However, according to the results of (Figure 3) the HAC with ED can classify them into separate classes. The best approach for clustering, according to a number of studies [41] is neural networks. In situations when the number of classes is important, this seems to be useful. Additionally, even in the presence of outliers and irrelevant data dimensions, the SOM maintains its overall stability and robustness. Our case study involves a small number of 100 CLCs and a small number of classes equal to 6.

Daily water use is an unpredictable process with erratic statistics and little underlying knowledge. There is a limitation in the above experiments that a CLC of a day can be only classified or analyzed when all of its 1440 data points are available. This is not suitable in the event of an abnormality (such as water leakage or low usage). An important piece of information that we want to identify is the period during which the water consumption is abnormal. An unevenly spaced data transmission period is defined by the meter's internal architecture. The sending times vary depending on how much water is consumed. Because of heavy water use, data are collected in shorter periods *delta*(*ti*). In contrast, the data is obtained with higher time delays *delta*(*ti*) when less water is consumed. To express this analysis mathematically, algorithm 1 uses the minimum (Min) and maximum (Max) *delta*(t_i) values of each day. This method evaluates the high consumption in relation to the moment the intelligent platform's indexes are received. This is accomplished by comparing variances in data received. A slight variation explains abnormal water use. The two slopes in a daily CLC displayed in red in Figure 5 indicate high consumption. According to the results, the peak times for consumption are from 7:45 am to 11 am and from 12 pm to 3 pm. However, consumption is minimal between midnight and 8:00 am, at noon, and in the late afternoon and evening from 3:00 pm to 8:00 pm. With the use of this straightforward technique, we were able to pinpoint different times of day when people consume water.

Algorithm 1 Zone Divided (CLC,$\delta(t_i)$)

```
if $\text{Min}(\delta(t_i)) \leqslant \delta(t_i)$ and $\delta(t_i) \leqslant \frac{[\text{Max}(\delta(t_i))-\text{Min}(\delta(t_i))]}{3}$ then
Existing water consumption (Contained consumption)
else if $\text{Max}(\delta(t_i)) \geq \delta(t_i)$ and $\delta(t_i) \geq 2 \times \frac{[\text{Max}(\delta(t_i))-\text{Min}(\delta(t_i))]}{3}$ then
No or a little water consumption
end if
```

F. SUPERVISED CLCs CLASSIFICATION IN REAL TIME WITH MULTILAYER PERCEPTRON (MLP)

The ED is used in this part to give unsupervised clustering. This clustering is only conducted on a subset of each CLC represented by the interval I with 400 data points. Figure 6(b) shows the clustering findings. Figure 6(a) displays the clustering results with full CLC (1440 data points). According to Table 3, the cophenetic correlation coefficient e determined using the $ED(CLC(I))$ with a number of data is equal to 0.99. The coefficient e, which was calculated using the $ED(CLC)$ distance and 1440 data points for each day, was 0.98. This result is better than the previous one. The time it took to create the cluster was in fact shortened by around 7 seconds.

TABLE 3. The figure shows computation time and cophenetic correlation coefficient for computing the similarity of HAC clustering with ED and DTW.

Approach	Analysis	
	Cophenetic Correlation Coefficient	Total Computation Time for Clustering (s)
ED (CLC)	0.98	32.34
DTW (CLC)	0.51	52.37
ED (CLC(I))	0.99	25.75

TABLE 4. The graph hows how the three clustering approaches distribute the water consumption load curves into different classes.

Cluster#	ED (% of Samples)	DTW (% of Samples)	SOM (% of Samples)
Cluster 1	1%	1%	1%
Cluster 2	1%	1%	1%
Cluster 3	72%	1%	18%
Cluster 4	24%	1%	31%
Cluster 5	1%	95%	22%
Cluster 6	1%	1%	27%

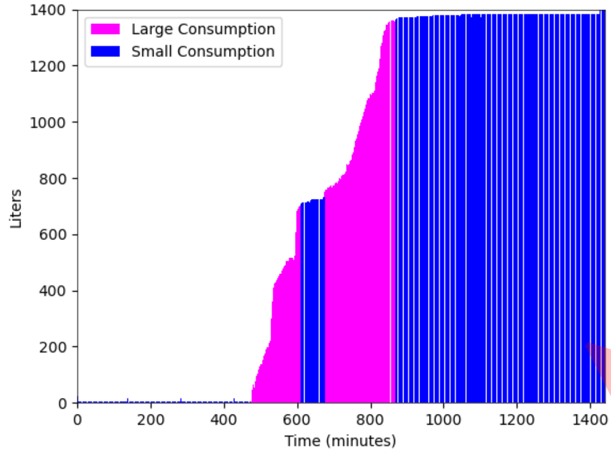


FIGURE 5. Automatic decomposing large and small water CLCs into zones.

1) SUPERVISED CLASSIFICATION WITH MLP MODEL

In this section, we will use the MLP model and want to categorize the CLCs that exist in period I. The sampled consumption, represented by 400 data points is the MLP's input. Each data point of I represents the liters consumed at a specific minute. We analyzed many MLP configurations and chose the one that offered the best categorization. The best MLP configuration includes output neurons representing the six classes denoted by k = 6, and the hidden layer contains 12 neurons. The hidden neurons have a sigmoid activation function. The partitions were made in the data using the following. 70% of the CLC samples were randomly picked to construct a training dataset. The test dataset is the remaining data. The MLP classifies the full CLCs by maintaining the same parameters. By comparing the classification accuracies, it is possible to analyze, (a) the effectiveness of clustering (presented in section IV-A), and (b) the effect of the reduction of CLC data points.

2) EXAMINING THE CLASSIFICATION OF CLCs: OFFLINE AND REAL-TIME

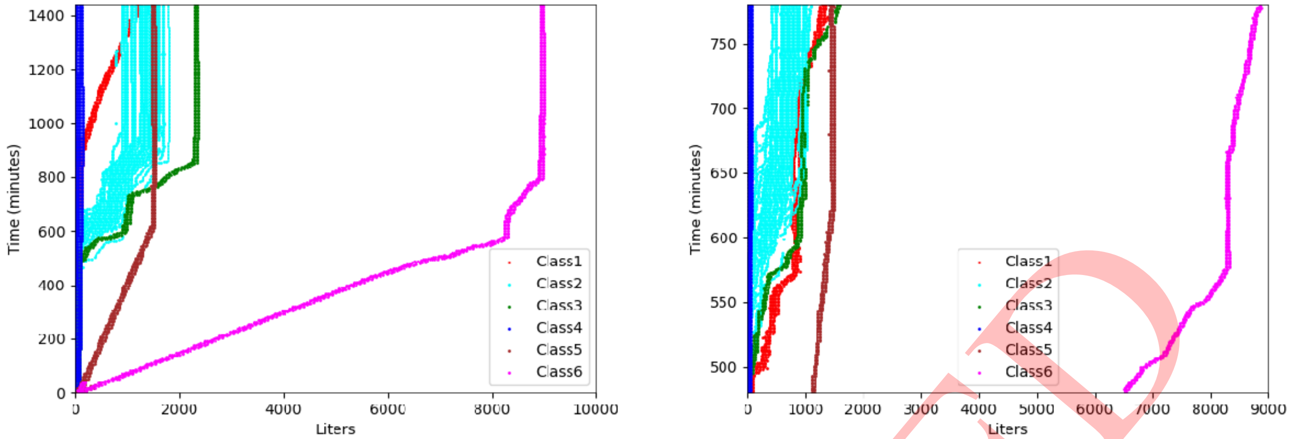
We achieved the results of Figure 7 by applying MLP clustering on full CLC and CLC(I) (containing a subset of data points). We can observe from the results that the clustering with HAC using ED is acceptable. This is supported by the MLP(CLC) classification accuracy of 99.92%. The complexity of the CLCs could be used to explain why the MLP model was unable to generalize its categorization. In addition, we can note that performing classification with period I (fewer data points) gives fast and more accurate results than the classification with full data points. The classification accuracy on training and test datasets with fewer data points are 99.97% and 100%, respectively. Furthermore, the time required to train the classification algorithm with full CLCs was 468.53 seconds, which was decreased to 41.44 seconds for CLCs with 400 data points. We have made available the first script that uses an MLP to categorize the CLCs of water use. This script records the water consumed per minute with interval I. Then, before the end of the day, it classifies each CLC(I). Similar to the first script, we developed another script that collects all 1440 data points of a day. However, it should be noted that the classification with the full CLC can be possible only at the end of the day.

V. ELECTRICAL DEVICES CLASSIFICATION USING SUPERVISED LEARNING

A cumulative load curve of electricity usage provides only the overall quantity used but it does not describe which set of electrical devices are connected to a power system. This is referred to as NILM (Non-Intrusive Load Monitoring) [33], [42]. To identify the type of electric device, there is a need to identify the signature of the electrical device from the consumption data [26], [33]. In this section, electrical devices in a residential building will be categorized.

A. IDENTIFYING CURRENT HARMONIC SIGNATURES

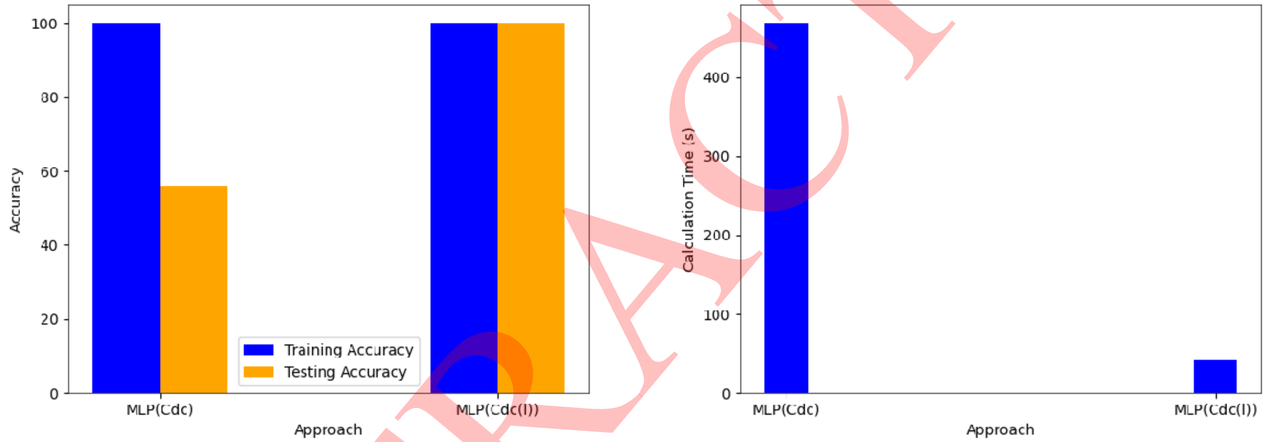
It is possible to identify specific electricity usage trends in the CLCs using ML algorithms [43]. It is, however, quite challenging to determine which device uses electricity. NILM (Non-Intrusive Load Monitoring) is a pioneering technique for analyzing the electrical power consumption of individual devices in a building without the need for physical access or measurement of each device [26]. However, obtaining precise current and voltage information is challenging in



a) Result with 1440 data points

b) Result with 400 data points

FIGURE 6. **The figure (a) shows the results of the classification of daily CLC sampled with 1440 data points. The figure (b) shows the results of the classification of the CLC over period *l* sampled with 400 data points.**



Accuracy

Calculation time (s)

FIGURE 7. **MLP classification accuracy with CLC and CLC(l).**

NILM due to the non-intrusive nature of the technique. This is because NILM relies on measurements of the overall electrical power consumption of a building or facility, rather than the power consumption of individual devices. To overcome this challenge, NILM relies on advanced algorithms known as dis-aggregation algorithms to analyze the overall power consumption data and separate it into the power consumption of individual devices. These algorithms use load signatures, which are unique patterns of energy usage, to identify and distinguish between different devices. The accuracy of NILM depends on the quality of the load signatures and the effectiveness of the dis-aggregation algorithms.

In [33], a NILM method based on artificial neural networks (ANNs) was proposed. The method uses a neural network to model the fast Fourier transform (FFT) of the observed electrical signal [37]. The FFT is a mathematical tool used to transform the time domain signal into the frequency domain, which allows for better analysis and interpretation of the signal. An ANN is trained to learn and approximate the FFT of the observed signal and update its approximation in real time. The authors showed that this NILM method based on ANNs is capable of accurately disaggregating the electrical signal into its constituent loads, making it a promising approach for non-intrusive energy monitoring and management. As defined in [33], harmonics are the elements of a repeating wave that have a frequency that is a whole number multiple of the basic frequency of the power line. The fundamental frequency is the lowest frequency present in a periodic waveform and is usually 50 or 60 Hz in power systems. High-frequency current measurement is necessary to correctly identify individual electrical devices in a NILM system. This requires a high sample rate, typically less than 1 millisecond, to capture the fine details of the current waveform. The high sample rate allows the NILM system to accurately analyze the current waveform and identify the unique load signatures of different devices. The load signatures, in turn, are used to

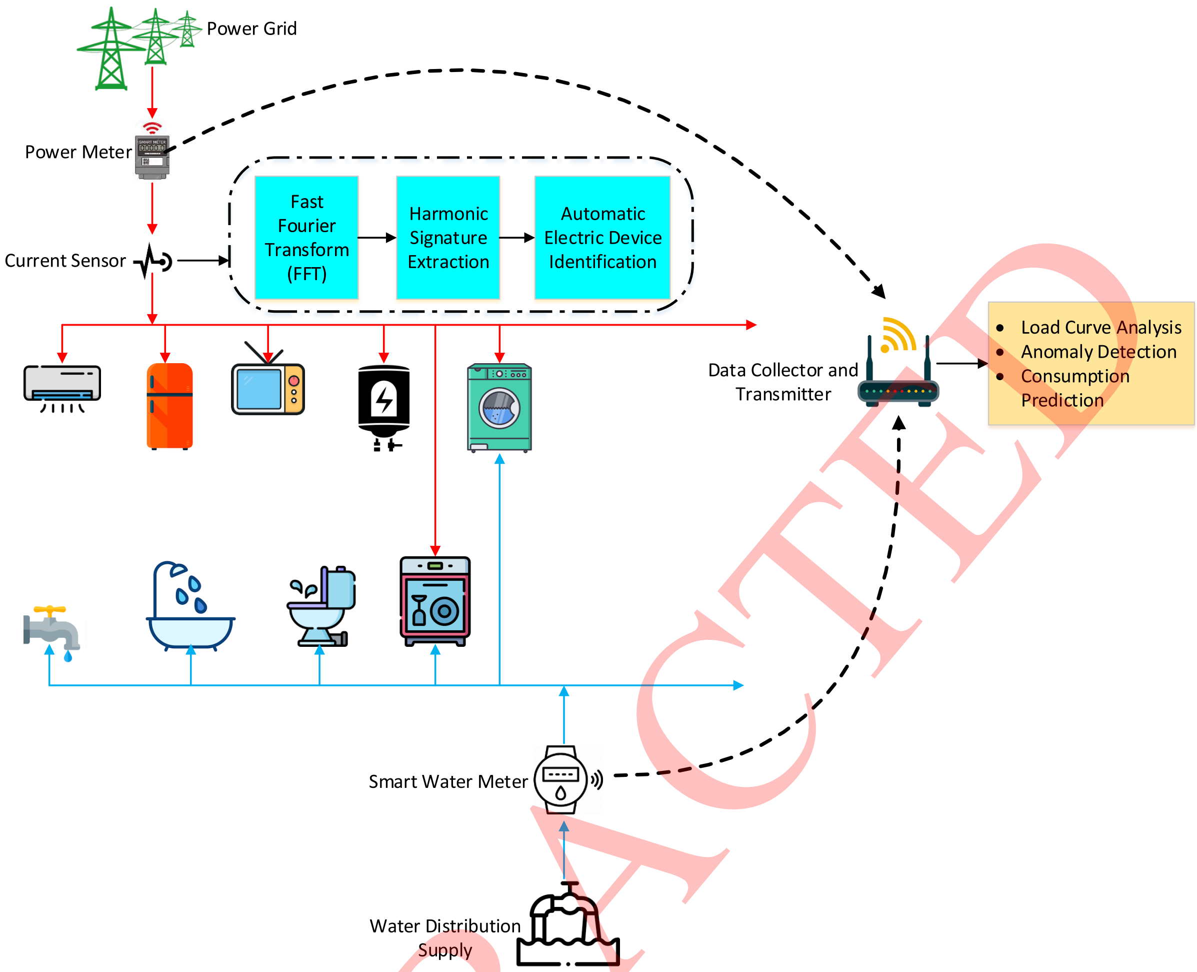

Power Grid
Power Meter
Current Sensor
Fast Fourier Transform (FFT)
Harmonic Signature Extraction
Automatic Electric Device Identification
Data Collector and Transmitter
Load Curve Analysis
Anomaly Detection
Consumption Prediction
Smart Water Meter
Water Distribution Supply

FIGURE 8. The overall design of the IoT system for measuring the water and electricity usage in a building using automated meters.

disaggregate the overall power consumption into the power consumption of individual devices. High-frequency current measurement is important for the accuracy of NILM, as it enables the system to distinguish between devices with similar power consumption patterns.

Only the first 16 most significant higher-order harmonic components are utilized as inputs to a classification algorithm. This is because higher-order harmonics are typically much weaker than the lower-order harmonics, and including more than the first 16 harmonics may not significantly improve the accuracy of the NILM system(Figure 8). Additionally, only odd harmonics with amplitudes from the first harmonic to the 15th harmonic are considered to be significant. This is because odd harmonics are typically stronger than even harmonics in non-linear loads and therefore can provide more accurate information for NILM. By considering only the odd harmonics, the NILM system can reduce the amount of data that needs to be processed and increase the accuracy of the results. This approach is commonly used in NILM methods that aim to balance the need for accuracy with computational efficiency. However, the specific harmonics considered in NILM systems can vary depending on the characteristics of the loads being monitored and the requirements of the system. Figure 9 displays the harmonic signatures of eight different electrical devices.

We proposed NILM techniques based on ML algorithms to recognize and categorize the electrical signatures of different types of devices under real-world operating conditions. The technique uses machine learning algorithms to learn the unique load signatures of different devices and identify them in real time. The electrical signatures of the devices, which provide information about their consumption, are measured under various conditions, including noisy, normal, and random ones. Through this, we obtained 40k samples. We call this a global signature database. Each sample represents a device's signature, represented by 16 features solely used as inputs by the supervised classifiers. Each sample of the global signature database represents a device's signature, and it is represented by 16 features. The model was trained on 67% random samples of the global database, and it was tested on 33% random samples. The devices are classified into eight categories: C1 = inverter air conditioner, C2 = bulb, C3 = TV, C4 = battery charger, C5 = lamp, C6 = computer, C7 = electric geyser, and C8 = refrigerator. By examining the bias-variance trade-off, we aim to build a model that achieves the maximum classification rate. Only the models that delivered the best categorization outcomes are discussed in the following sections.

B. AUTOMATIC CLASSIFICATION OF ELECTRICAL DEVICES SIGNATURES

Three supervised ML techniques were used to classify electrical devices: the KNN, the MLP, and the decision tree (DT).

- Multilayer Perceptron (MLP) Model 1: In this model, the sigmoid neurons are used in the hidden layers to introduce non-linearity into the model. This allows the model to learn complex relationships between the input features and the output. The linear neurons are used in the output layer to produce the predicted class of the electric device.
- Multilayer Perceptron (MLP) Model 2: The MLP model utilizes sigmoid neurons extensively in its hidden layer and linear neurons in its output layer. The number of neurons in the output layer varies based on the number of devices that need to be identified from the input signal. In the given example, there are 8 output neurons in the MLP. Each output neuron will provide either a 0 or 1, indicating the ON or OFF state of each device, respectively. An output $(1, 0, 0, 0, 0, 0, 0, 0)$, for instance, corresponds to $C1$ (inverter AC). Therefore, this MLP is a binary output model referred to as "MLP model 2".
- Decision tree (DT): CART [44] stands for classification and regression trees, which is a decision tree algorithm used for both classification and regression tasks. In CART, the decision tree is constructed by recursively splitting the data into two subsets based on the best feature and threshold, with the goal of creating homogeneous subsets that lead to accurate predictions. The end result is a binary tree structure, where each internal node represents a test on a feature, and each leaf node represents a prediction (eight electrical device classes that need to classify in our case).
- The k-nearest neighbor (KNN): KNN is a non-parametric and instance-based learning algorithm used for classification and regression tasks. The idea behind KNN is to classify new samples based on the majority vote of their k nearest neighbors from the training data. We used KNN with two models. KNN-Model 1 with k=1 neighbors, and KNN-Model 2 with k=10 neighbors.

C. AUTOMATICALLY IDENTIFYING ELECTRICAL DEVICES USING MACHINE LEARNING MODELS

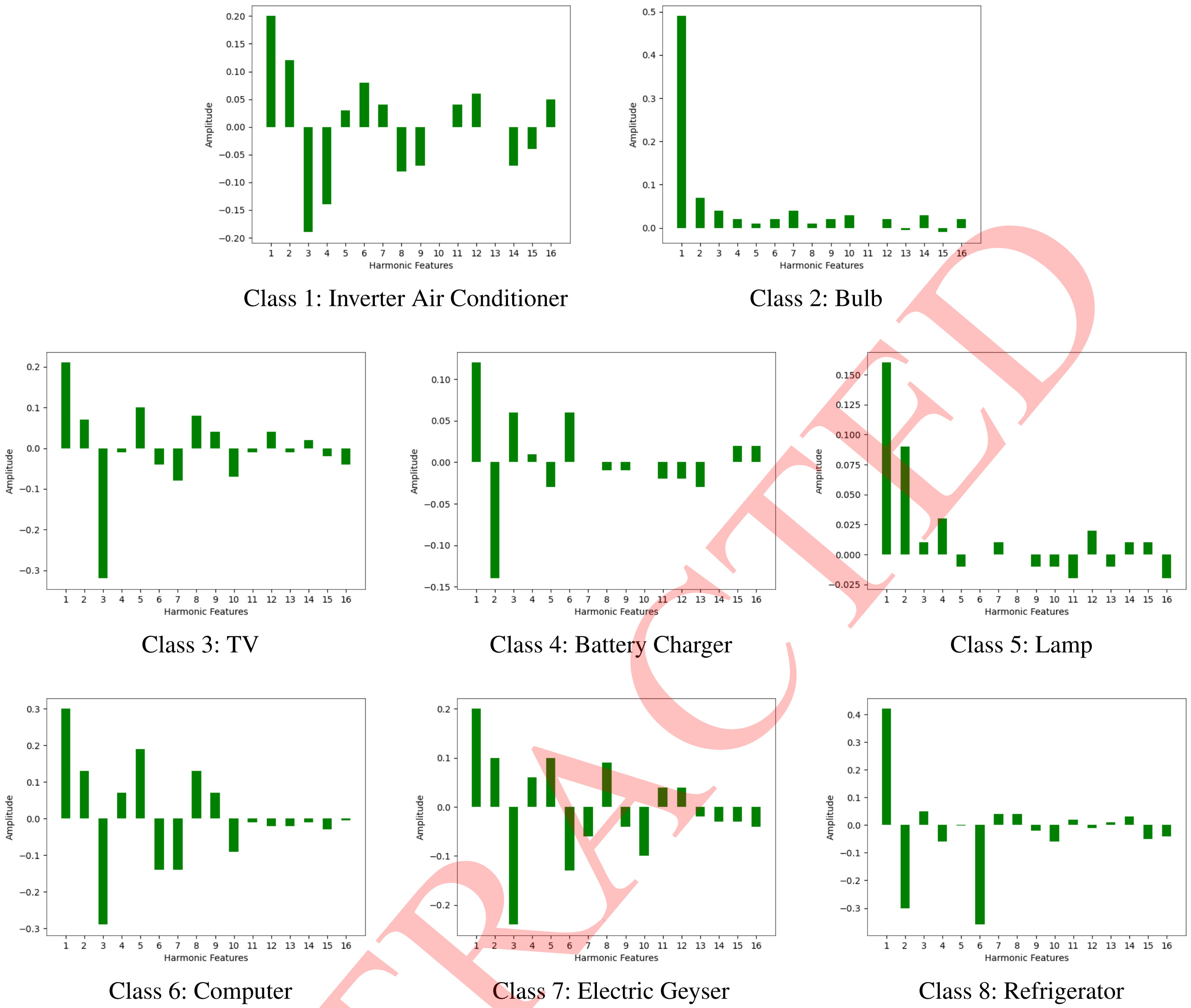
There are generally two main steps in supervised learning: learning and testing. In learning, the algorithm is trained on the available labeled data to learn the relationship between the input features and output classes. In testing the learned model is applied to new, unseen samples to make predictions and evaluate the performance of the model in terms of accuracy. We evaluated and tested the hidden layer of the MLP model from having two to twenty-five neurons to find the best configuration of the hidden layer. According to Figure 11 results, the MLP (model 1) returns the best classification accuracy with 17 neurons. The MLP (model 2) returns the best accuracy with 20 neurons in the hidden layer. Even though we can identify other better classification outcomes, we decided to keep these results for classification rates since they show a better bias/variance trade-off. The results of the electrical device classification are shown in Figure 10. If we compare the DT with KNN and MLP, then the DT is the fastest and returns higher accuracy. Only 1.8% of signatures cannot be classified to a certain class of electrical device when using DT. The decision tree structure provides a clear representation of the relationships between features and the target variable, making it easy for users to understand how the model is making its predictions. However, if the tree becomes too deep, it can lead to overfitting, meaning that the model becomes too complex and fits the training data too well, capturing the noise or random fluctuations in the data instead of the underlying patterns [45]. Another limitation of decision trees is that they are less flexible compared to some other machine learning algorithms. Once the tree structure is set during the training phase, the tree must be traversed in the same way for each new sample, based on the rules and conditions defined at each node. This means that the decision tree is not easily adaptable to changes in the distribution of the data, and it can struggle with handling new and unseen situations. Also, it should be noted that a simple classifier does not always guarantee a lower error rate.

D. CLASSIFYING ELECTRICAL DEVICES WITH ASSEMBLY APPROACHES USING OAA AND ECOC

The objective of this section is to explore assembly approaches for improving classification accuracy. Two approaches are used: ECOC (Error-Correcting Output Codes) [46] and OAA [47], [48].

1) CLASSIFICATION USING ECOC (ERROR-CORRECTING OUTPUT CODES)

This approach involves using many MLP-based classifiers. It enables us to choose a combination strategy for merging the results of different classifiers. This method is based on the ECOC (Error-Correcting Output Codes) concept [46]. This enables us to define a variety of classifiers, in this instance MLPs, each of which focuses on the classification of a certain class. In our example, there are eight categories for devices, thus we can then turn the multiple classification issue into a binary classification problem. As part of this strategy, for each collection of classes, a model that sets it apart from the others must be built [49]. First, a device class code a distinctive group of characters made up of the numbers 0 and 1, is used to identify each of the eight classes. To do this, it is necessary



Class 1: Inverter Air Conditioner

Class 2: Bulb

Class 3: TV

Class 4: Battery Charger

Class 5: Lamp

Class 6: Computer

Class 7: Electric Geyser

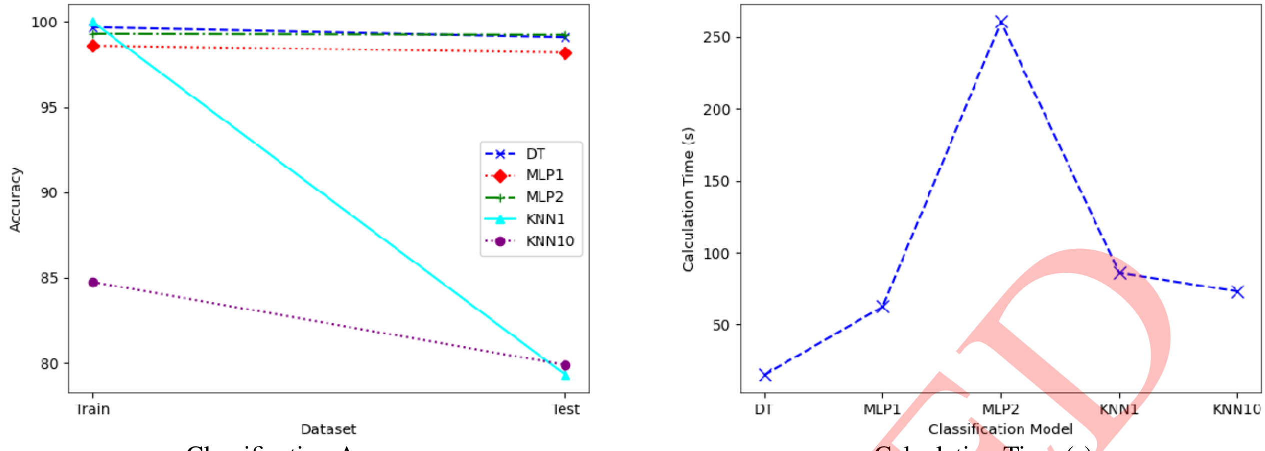
Class 8: Refrigerator

FIGURE 9. **Harmonic current features of 8 electric devices.**

to figure out how many classifiers are utilized for testing and learning [50]. Let k be the number of classes and l be the size of the device class code DC that identifies each device class. The following characteristics should be ensured by the number of classifiers and device class code size.

- The size of the device class code should be of sufficient length to allow separate class codes to be assigned to each class of device.
- The number of classifiers must be fewer than the number of classes, otherwise the classification would become an OAA (One Against All) [47], [50].

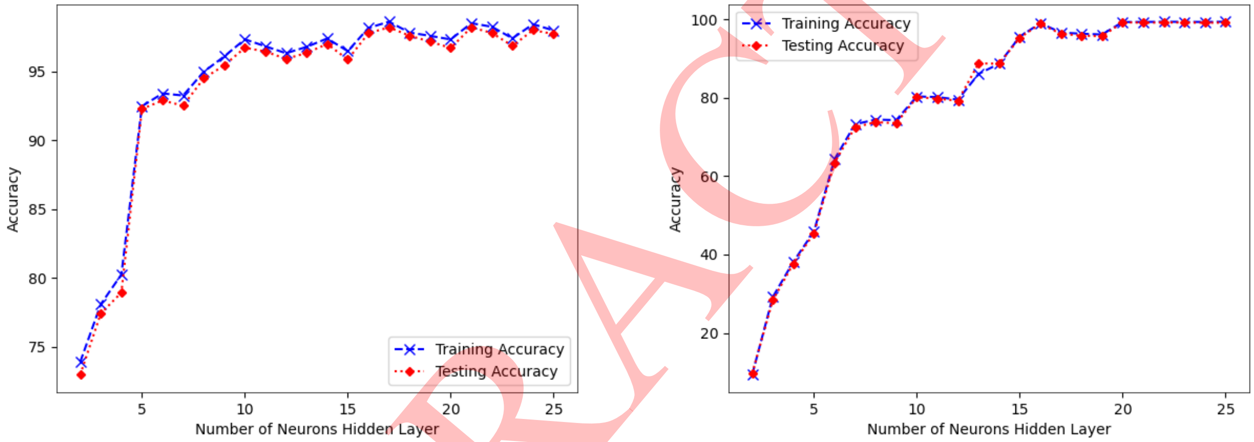
To encode the eight device classes we have $3 \leq l(DC) < k$ because $2^3 = 8$ with $DC \in 0, 1$. The class code size has four classifiers, represented in Table 5 as h_1, h_2, h_3, and h_4. Each classifier classifies n classes with $1 \leq n \leq k$. The first classifier, h_1, classifies the classes (1, 2, 3, 5, and 6). The second classifier, h_2, classifies the classes (1, 2, 4, and 6). The third classifier, h_3, classifies the classes (1, 3, 5, and 7). The final classifier h_4 learns classes 1, 2, 3, 4, and 7. Once n classes were established, we utilized four MLPs. For the classifiers h_1, h_2, h_3, and h_4, we ultimately decided to keep 20, 16, 16, and 6 neurons after testing these models by altering a different number of hidden layers. Each model generates a single output that is classified into a single class or several classes. An aggregation of the results from four classifiers is used as the new signature's test, and it is compared to the DC using the Hamming distance. This distance enables the comparison of two binary data strings. The difference between $\widehat{DC}$ and DC is that $\widehat{DC}$ is the predicted class while DC is the desired class. Eight devices are represented by the



Classification Accuracy

Calculation Time (s)

FIGURE 10. Classification accuracy of electrical devices signatures with Decision Tree, two proposed MLP models and KNN.



MLP Model 1

MLP Model 2

FIGURE 11. The accuracy of the MLP configurations (model 1 and model 2) when using different numbers of hidden neurons to classify the global database of electrical device signatures.

binary vector F. The number of classes is n_{ap}.

$$\begin{cases} \forall DC, \widehat{DC} \in F = \{0, 1\} = (DC_i)_{i\in[1,n_{ap}]} \\ \widehat{DC} = \left(\widehat{DC}_i\right)^{n_{ap}}_{i\in[1,n_{ap}]} \\ D(DC, \widehat{DC}) = \neq \{i, DC \neq \widehat{DC}\} = \sum_{i=1} \left(DC_i XOR \widehat{DC}_i\right) \end{cases} \tag{8}$$

The Hamming distance is represented by Equation 8. It specifies the number of elements in the set of DC values. It differs from $\widehat{DC}$. The classification of the ECOC-MLP model yielded the following results. The classification accuracy on the training dataset is 99.74%, and it is 99.55% on the testing dataset. The average training and testing time is 52.22 seconds.

2) CLASSIFICATION USING OAA (ONE AGAINST ALL)

To analyze the effectiveness of ECOC-MLP another approach based on the classification of one class at a time is proposed. To classify eight electrical devices eight OAA-MLP classifiers with binary outputs are used. The 16 features corresponding to each electrical device serve as inputs to the OAA-MLPs. Each OAA-MLP has eight outputs, allowing it to recognize only one class at a time. The results of these eight models were averaged after being evaluated ten times. Figure 12 shows the final configuration options for eight OAA-MLP models. The classification rates of the training and test sets determine the optimal number of hidden layers in the OAA-MLPs. The class $\widehat{DC}$ of a new signature is calculated using the following equation from the outputs of the eight OAA-MLPs.

$$\widehat{DC} = \arg\max\,(\text{MLP}-OAA_i)_{i\in[1,8]} \tag{9}$$

Figure 12 shows the results with OAA-MLP models. From the results, we found that classes C8 and C3 are the easiest to categorize. The high classification rates on the training and test dataset demonstrated this. The OAA-MLP model

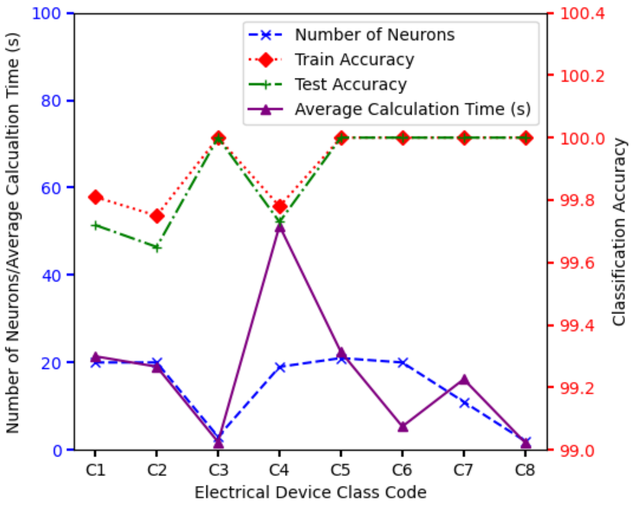


FIGURE 12. Classification accuracy and optimal number of neurons required for the classification of electrical devices with OAA-MLP.

produces a classification accuracy of 99.93% on the training dataset and 99.9% on the testing dataset. The classification process took around 400 seconds. According to the results, the OAA-MLP outperformed MLP models 1 and 2 as well as the DT and KNN. If we compare classification accuracies then the OAA-MLP achieves a classification accuracy of 99.93%, compared to the classification accuracy of 99.65% and 99.28% with the DT and MLP model 2 respectively. It should be emphasized that while DT, MLPs, and KNNs are capable of solving multi-class classification problems, their classification accuracy can vary based on various factors such as the complexity of the data, the choice of hyperparameters, and the presence of noisy or imbalanced data. These approaches thus do not always ensure classification accuracy. Whereas OAA-MLP enhances the classification performance while simplifying the complexity. ECOC is an ensemble technique that divides the multi-class classification problem into multiple binary subproblems and trains multiple binary classifiers to make predictions.

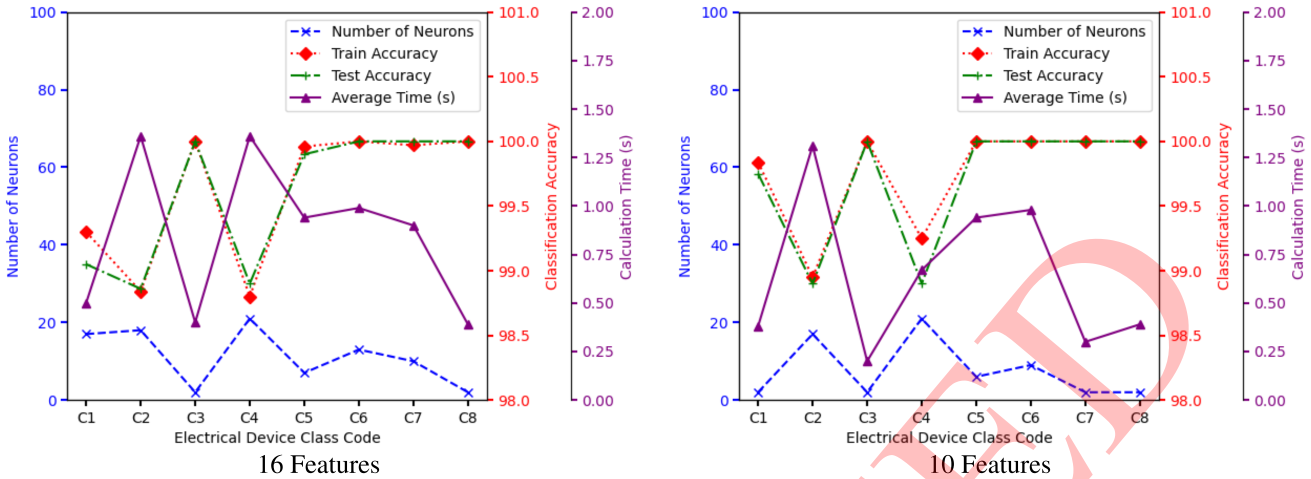
We can also demonstrate the effectiveness of the proposed approach by comparing its accuracy with the classification accuracies reported in previous research [33], [34]. In previous research, it is reported that binary MLP and RBF kernel-based SVM model classify the TV with an accuracy of 76.86% and 98.70% respectively [33]. We found both are lower than our result of 99.83%. Similarly, in previous research, the classification accuracies to classify refrigerator and lamp are 88.8% and 61.5% respecitvely [34]. While our approach achieves a classification accuracy of 100%. Furthermore, we discovered that some devices are more challenging to classify than others using the OAA-MLP classification. Thus we propose to analyze the features of electrical devices using deep learning technique [51] to reduce the number of inputs to the classification model.

TABLE 5. ECOC classifiers with the device class codes of eight electrical devices to be classified.

Classifier	Class							
	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$
h1	1	1	1	0	1	1	0	0
h2	1	1	0	1	0	1	0	0
h3	1	0	1	0	1	0	1	0
h4	1	1	1	1	0	0	1	0

E. FEATURE REDUCTION OF ELECTRICAL DEVICES SIGNATURES

Assigning importance scores to features provide information about which features are more relevant and significant in the data, and which features can be ignored or removed. These also help to understand which features have the biggest impact on the model's predictions, providing insight into the underlying relationships in the data. The scores also indicate the most important features for distinguishing one device from another and the least important features. Typically, a subject matter expert interprets this stage and uses it as a starting point for gathering more or alternative data. As a result, the categorization model is simplified and its inputs are reduced. In our application, we calculate the significance scores of each feature based on the performance of the decision tree approach. The decision trees (DTs) are used to represent a new model that uses the data in a way that emphasizes the relevance of each feature based on how frequently it is utilized to make important choices. To enable ranking and comparison of features, this relevance is determined directly for each feature. It is calculated as follows: for a single decision tree, the significance of each feature is assessed by counting the number of times that each feature plays a role in improving the performance metric, taking into account the number of observations assigned to each node. The average of all decision trees' feature significance ratings is then calculated. An ensemble model that combines a number of DTs, the set of gradient-boosting trees is the foundation for allocating significance scores for each feature [52]. Gradient Boosting is an ensemble learning technique for regression and classification problems. It combines multiple weak models, such as decision trees, to form a stronger and more accurate model. In gradient boosting, each regression tree is trained to correct the errors made by the previous tree in the sequence. The prediction from each tree is added to the overall prediction, and the process is repeated until a stopping criterion is met. The scikit-learn (Python) XGBoost library is utilized with its default settings. Separate analyses were done on the random data, tampered data, noisy data, and normal data. Key characteristics of assignment to a particular device class are extracted by thoroughly examining each subset. The proposed ECOC-MLP model serves as the foundation for this task. Table 6 provides the outcomes of adopting this methodology. Figure 13 shows the classification accuracy of electrical devices with the ECOC-MLP model



16 Features

10 Features

FIGURE 13. Classification accuracy of electrical devices with the ECOC-MLP model before and after extracting relevant features.

TABLE 6. Assigning relevance scores to the features of electrical devices with the gradient-boosting tree method.

Dataset	**Features**															
	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$	$C9$	$C10$	$C11$	$C12$	$C13$	$C14$	$C15$	$C16$
Normal	0.03	0.01		0.1		0.83										
Noisy	0.03	0.01		0.1		0.83										
Tempered								1.0								
Random	0.1	0.08	0.05	0.063	0.07	0.4	0.045	0.05	0.001	0.005		0.001		0.002		0.001

before and after extracting relevant features. Each feature of the electronic devices was given a priority score using gradient-boosting DT. We only chose the features from these scores that had a score greater than 0.005 and that allowed us to choose several features based on the subset of the data:

- Normal and Noisy datasets: For both datasets, only four features out of the 16 initiated features are chosen (1, 2, 4, 6). On both datasets, the classification accuracy was 100% and the algorithm took less than one second to classify all devices.
- Tampered data: In this dataset, only 8 automatically identified features are retained. Similar to the above datasets, the algorithm took less than one second to classify all devices with a classification accuracy of 100%.
- Random data: For this dataset, only the first 10 features are chosen. The classification duration and accuracy of the electrical device signature data were positively affected by classifying the electrical devices with fewer features. Figure 13 illustrates the accuracy of classifying electrical devices using the ECOC-MLP model, both before and after extracting 10 features.

We compared three machine-learning approaches for the classification of electrical devices. These are a KNN with two configurations, an MLP with two configurations, and a DT. According to the results of the classification, model 2 of the neural network returns the best accuracy but this model requires high computation time for training. We suggest using ECOC-MLP and OAA-MLP ensemble classification models to reduce computation time. Through these models, we were able to convert the eight classes of our multi-class classification into a binary classification. Furthermore, the models reduce the training time.

VI. CONCLUSION

The article intends to investigate consumption patterns of water and electricity that appear normal and abnormal. With proposed machine learning approaches, we can classify the daily water and electricity consumption more promptly. The experiments demonstrated that at the consumer level, changes in consumption patterns can be detected through the clustering and supervised classification of water consumption load curves (CLCs). Additionally, the proposed approaches can accurately identify abnormal water consumption classes caused by occasional events (heat waves, holidays, etc.). To identify electricity usage, both supervised and unsupervised algorithms were utilized to isolate individual devices from the aggregated load current. A non-intrusive data collection technique combined with machine learning is proposed. The results demonstrate that it can accurately identify and categorize the unique electrical signatures of different electrical devices in real-world working conditions. To enhance the classification accuracy, we evaluated two assembly methods using a group of optimal classifiers. These are based on OAA-MLP and ECOC-MLP concepts. From the experiments, we found that both OAA-MLP and ECOC-MLP perform better than MLPs, DT, and KNN. We also found that OAA-MLP outperformed all other approaches in terms of performance, but it is more complex than the ECOC-MLP.

We also improved the classification accuracy by reducing the features of the electrical devices. This involved feature extraction and analysis. A gradient-boosting approach based on decision trees is proposed to assign relevance ratings to each of the 16 features used as inputs to the classification models.

Future work in the field of water and electricity consumption prediction involves improving the accuracy of the prediction models through the use of more advanced machine learning algorithms and techniques, incorporating additional data sources such as weather data or socio-economic indicators, and developing new approaches to interpret the results of the predictions for decision-making purposes. Additionally, developing scalable and secure solutions for collecting and analyzing large amounts of smart meter data can also be an area of future work.

REFERENCES

[1] M. Issaoui, S. Jellali, A. A. Zorpas, and P. Dutournie, "Membrane technology for sustainable water resources management: Challenges and future projections," *Sustain. Chem. Pharmacy*, vol. 25, Apr. 2022, Art. no. 100590.

[2] A. Colmenar-Santos, A.-M. Muñoz-Gómez, E. Rosales-Asensio, G. Fernandez Aznar, and N. Galan-Hernandez, "Adaptive model predictive control for electricity management in the household sector," *Int. J. Electr. Power Energy Syst.*, vol. 137, May 2022, Art. no. 107831.

[3] M. S. Aliero, K. N. Qureshi, M. F. Pasha, and G. Jeon, "Smart home energy management systems in Internet of Things networks for green cities demands and services," *Environ. Technol. Innov.*, vol. 22, May 2021, Art. no. 101443.

[4] S. Tiwari, J. Rosak-Szyrocka, and J. Żywiołek, "Internet of Things as a sustainable energy management solution at tourism destinations in India," *Energies*, vol. 15, no. 7, p. 2433, Mar. 2022.

[5] M. A. Omran, B. J. Hamza, and W. K. Saad, "The design and fulfillment of a smart home (SH) material powered by the IoT using the blynk app," *Mater. Today: Proc.*, vol. 60, pp. 1199–1212, 2022.

[6] I. Szilagyi and P. Wira, "Ontologies and semantic web for the Internet of Things—A survey," in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2016, pp. 6949–6954.

[7] A. Koohang, C. S. Sargent, J. H. Nord, and J. Paliszkiewicz, "Internet of Things (IoT): From awareness to continued use," *Int. J. Inf. Manage.*, vol. 62, Feb. 2022, Art. no. 102442.

[8] R. F. Molanes, K. Amarasinghe, J. Rodriguez-Andina, and M. Manic, "Deep learning and reconfigurable platforms in the Internet of Things: Challenges and opportunities in algorithms and hardware," *IEEE Ind. Electron. Mag.*, vol. 12, no. 2, pp. 36–49, Jun. 2018.

[9] J. L. Gallardo, M. A. Ahmed, and N. Jara, "Clustering algorithm-based network planning for advanced metering infrastructure in smart grid," *IEEE Access*, vol. 9, pp. 48992–49006, 2021.

[10] S. D. T. Kelly, N. K. Suryadevara, and S. C. Mukhopadhyay, "Towards the implementation of IoT for environmental condition monitoring in homes," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3846–3853, Oct. 2013.

[11] A. Grandjean, J. Adnot, and G. Binet, "A review and an analysis of the residential electric load curve models," *Renew. Sustain. Energy Rev.*, vol. 16, no. 9, pp. 6539–6565, Dec. 2012.

[12] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5196–5206, Sep. 2018.

[13] D. Walker, E. Creaco, L. Vamvakeridou-Lyroudia, R. Farmani, Z. Kapelan, and D. Savić, "Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks," *Proc. Eng.*, vol. 119, pp. 1419–1428, Jan. 2015.

[14] A. Candelieri, D. Soldi, and F. Archetti, "Short-term forecasting of hourly water consumption by using automatic metering readers data," *Proc. Eng.*, vol. 119, pp. 844–853, Jan. 2015.

[15] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, Jul. 2017.

[16] M. Liu, D. Liu, G. Sun, Y. Zhao, D. Wang, F. Liu, X. Fang, Q. He, and D. Xu, "Deep learning detection of inaccurate smart electricity meters: A case study," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 79–90, Dec. 2020.

[17] J. Szoplik, "Forecasting of natural gas consumption with artificial neural networks," *Energy*, vol. 85, pp. 208–220, Jun. 2015.

[18] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019.

[19] M. L. Abadi, A. Same, L. Oukhellou, N. Cheifetz, P. Mandel, C. Feliers, and O. Chesneau, "Predictive classification of water consumption time series using non-homogeneous Markov models," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2017, pp. 323–331.

[20] P. Huntra and T. C. Keener, "Evaluating the impact of meteorological factors on water demand in the Las Vegas Valley using time-series analysis: 1990–2014," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 8, p. 249, 2017.

[21] S. Alvisi, M. Franchini, and A. Marinelli, "A short-term, pattern-based model for water-demand forecasting," *J. Hydroinform.*, vol. 9, no. 1, pp. 39–50, Jan. 2007.

[22] C. Peña-Guzmán, J. Melgarejo, and D. Prats, "Forecasting water demand in residential, commercial, and industrial zones in Bogotá, Colombia, using least-squares support vector machines," *Math. Problems Eng.*, vol. 2016, Dec. 2016, Art. no. 5712347.

[23] D. S. Kenney, C. Goemans, R. Klein, J. Lowrey, and K. Reidy, "Residential water demand management: Lessons from Aurora, Colorado," *JAWRA J. Amer. Water Resour. Assoc.*, vol. 44, no. 1, pp. 192–207, 2008.

[24] D. C. Fettermann, A. Borriello, A. Pellegrini, C. G. Cavalcante, J. M. Rose, and P. F. Burke, "Getting smarter about household energy: The who and what of demand for smart meters," *Building Res. Inf.*, vol. 49, no. 1, pp. 100–112, Jan. 2021.

[25] J. Y. Lee and T. Yim, "Energy and flow demand analysis of domestic hot water in an apartment complex using a smart meter," *Energy*, vol. 229, Aug. 2021, Art. no. 120678.

[26] G. W. Hart, "Nonintrusive appliance load monitoring," *Proc. IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec. 1992.

[27] J. Spiegel, P. Wira, and G. Hermann, "A comparative experimental study of lossless compression algorithms for enhancing energy efficiency in smart meters," in *Proc. IEEE 16th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2018, pp. 447–452.

[28] J. Spiegel, P. Wira, and G. Hermann, "Energy efficiency optimization in fluid flow metering," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Feb. 2018, pp. 1940–1945.

[29] A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, and B. B. Gupta, "Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings," *Future Gener. Comput. Syst.*, vol. 82, pp. 349–357, May 2018.

[30] F. Biscarri, I. Monedero, A. García, J. I. Guerrero, and C. León, "Electricity clustering framework for automatic classification of customer loads," *Expert Syst. Appl.*, vol. 86, pp. 54–63, Nov. 2017.

[31] S. Chu, E. Keogh, D. Hart, and M. Pazzani, "Iterative deepening dynamic time warping for time series," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2002, pp. 195–212.

[32] A. Boudhaouia and P. Wira, "A real-time data analysis platform for short-term water consumption forecasting with machine learning," *Forecasting*, vol. 3, no. 4, pp. 682–694, Sep. 2021.

[33] T.-M. Nguyen, "Contribution to the analysis and understanting of electricalgrid signals with signal processing and machine learning techniques," Ph.D. thesis, Mulhouse, France, 2017.

[34] D. Srinivasan, W. S. Ng, and A. C. Liew, "Neural-network-based signature recognition for harmonic source identification," *IEEE Trans. Power Del.*, vol. 21, no. 1, pp. 398–405, Jan. 2006.

[35] C. Wu, Q. Peng, J. Lee, K. Leibnitz, and Y. Xia, "Effective hierarchical clustering based on structural similarities in nearest neighbor graphs," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107295.

[36] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.

[37] J. M. Giron-Sierra, *Digital Signal Processing With MATLAB Examples*, vol. 3. Springer, 2017.

[38] A. Zimmermann, "Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 2, p. e1330, Mar. 2020.

[39] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, no. 2, pp. 33–40, Feb. 1962.

[40] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Comput. Statist. Data Anal.*, vol. 71, pp. 52–78, Mar. 2014.

[41] P. Mangiameli, S. K. Chen, and D. West, "A comparison of SOM neural network and hierarchical clustering methods," *Eur. J. Oper. Res.*, vol. 93, no. 2, pp. 402–417, Sep. 1996.

[42] S. W. Makonin, "Real-time embedded low-frequency load disaggregation," Ph.D. thesis, Appl. Sci., School Comput. Sci., 2014.

[43] N. Somu, G. Raman M R, and K. Ramamritham, "A deep learning framework for building energy consumption forecast," *Renew. Sustain. Energy Rev.*, vol. 137, Mar. 2021, Art. no. 110591.

[44] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The CART decision tree for mining data streams," *Inf. Sci.*, vol. 266, pp. 1–15, May 2014.

[45] M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, and S. Kallel, *The Impact of Digital Technologies on Public Health in Developed and Developing Countries: 18th International Conference, ICOST 2020, Hammamet, Tunisia, June 24–26, 2020, Proceedings*, vol. 12157. Springer Nature, 2020.

[46] S. Escalera and O. Pujol, "ECOC-ONE: A novel coding and decoding strategy," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 578–581.

[47] Y. Duan, B. Zou, J. Xu, F. Chen, J. Wei, and Y. Y. Tang, "OAA-SVM-MS: A fast and efficient multi-class classification algorithm," *Neurocomputing*, vol. 454, pp. 448–460, Sep. 2021.

[48] J. Sharmila Joseph, A. Vidyarthi, and V. P. Singh, "Multiclass image classification using OAA-SVM," in *Machine Intelligence and Smart Systems*. Springer, 2022, pp. 235–244.

[49] N. Hatami, R. Ebrahimpour, and R. Ghaderi, "ECOC-based training of neural networks for face recognition," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Sep. 2008, pp. 450–454.

[50] M. Hasan, A. Kotov, A. Idalski Carcone, M. Dong, S. Naar, and K. Brogan Hartlieb, "A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories," *J. Biomed. Informat.*, vol. 62, pp. 21–31, Aug. 2016.

[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[52] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Springer, 2009.

SHARIQ BASHIR received the Doctorate (Dr.-Ing.) (Ph.D.) degree in engineering—computer science from the Vienna University of Technology, Austria. In 2013, he was a Postdoctoral Researcher with New York University Abu Dhabi (NYUAD). During his Ph.D. research, he worked closely with Information Retrieval Facility (IRF), Vienna. He has more than 15 years of progressive experience in academia and research. He is currently an Assistant Professor in big data and machine learning at the Institute of Applied Data Analytics (IADA), Universiti Brunei Darussalam (UBD). He has published numerous papers in refereed journals and international conferences and served as a reviewer and a committee member for several major journals, conferences, and workshops. His research interests include the broad scope of information retrieval, data science, and machine learning. In recent years, he has been investigating techniques for private web search. He is also collaborating with researchers in several disciplines of information retrieval, machine learning, and data science.