Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR036690

Comparative Analysis of Supervised Classification Algorithms for Residential Water End Uses

Zahra Heydari[1] and Ashlynn S. Stillwell[1]

[1]Civil and Environmental Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

Key Points:

- We generated synthetic labeled data representing residential water end uses with a Conditional Tabular Generative Adversarial Network
- In the context of water end-use data, decision tree-based models emerge as the optimal choice for classification tasks
- Logistic function-based models, such as logistic regression, are computationally efficient alternatives for classifying specific end uses

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. S. Stillwell,
ashlynn@illinois.edu

Citation:

Heydari, Z., & Stillwell, A. S. (2024). Comparative analysis of supervised classification algorithms for residential water end uses. *Water Resources Research*, *60*, e2023WR036690. https://doi.org/10.1029/2023WR036690

Received 9 NOV 2023
Accepted 6 MAY 2024

Author Contributions:

Conceptualization: Zahra Heydari, Ashlynn S. Stillwell
Data curation: Zahra Heydari, Ashlynn S. Stillwell
Formal analysis: Zahra Heydari
Funding acquisition: Ashlynn S. Stillwell
Methodology: Zahra Heydari, Ashlynn S. Stillwell
Project administration: Ashlynn S. Stillwell
Resources: Ashlynn S. Stillwell
Supervision: Ashlynn S. Stillwell
Visualization: Zahra Heydari
Writing – original draft: Zahra Heydari

Abstract Water sustainability in the built environment requires an accurate estimation of residential water end uses (e.g., showers, toilets, faucets, etc.). In this study, we evaluate the performance of four models (Random Forest, RF; Support Vector Machines, SVM; Logistic Regression, Log-reg; and Neural Networks, NN) for residential water end-use classification using actual (measured) and synthetic labeled data sets. We generated synthetic labeled data using Conditional Tabular Generative Adversarial Networks. We then utilized grid search to train each model on their respective optimized hyperparameters. The RF model exhibited the best model performance overall, while the Log-reg model had the shortest execution times under different balanced and imbalanced (based on number of events per class) synthetic data scenarios, demonstrating a computationally efficient alternative for RF for specific end uses. The NN model exhibited high performance with the tradeoff of longer execution times compared to the other classification models. In the balanced data set scenario, all models achieved closely aligned F1-scores, ranging from 0.83 to 0.90. However, when faced with imbalanced data reflective of actual conditions, both the SVM and Log-reg models showed inferior performance compared to the RF and NN models. Overall, we concluded that decision tree-based models emerge as the optimal choice for classification tasks in the context of water end-use data. Our study advances residential smart water metering systems through creating synthetic labeled end-use data and providing insight into the strengths and weaknesses of various supervised machine learning classifiers for end-use identification.

Plain Language Summary We looked at how well different computer models can tell apart types of water use in homes, like identifying when someone is taking a shower versus flushing a toilet. We used real water meter data and also created fake, but realistic, water use data to test these models. Among the models we tested, the Random Forest model (a method that uses a collection of decisions to make predictions) was the most accurate. However, the Logistic Regression model, another type of model we tested, was faster in analyzing the data, making it a good option for quickly identifying specific water uses without needing as much computer power. We also found that all the models we tested were close in detecting what type of water use event had occurred when the data were evenly distributed across different types of water use. But, when the data were uneven—more like real-life situations—the Random Forest and Neural Network models were better than the others. This research helps improve systems that monitor how humans use water in homes, making it easier to identify where water is used and how we can save more of it, contributing to more sustainable living environments.

1. Introduction

Increasing population and urbanization combined with climate and land-use change have increased attention to water scarcity and environmental sustainability issues (R. McDonald et al., 2011). Water demand management strategies are more cost-effective and less practically constrained than water supply expansion in the context of scarcity (Inman & Jeffrey, 2006; R. I. McDonald et al., 2014). Consequently, water conservation and efficiency in urban areas is an ongoing pertinent challenge, depending on human behavior and water consumption patterns (Willis et al., 2011). Accordingly, the deployment of smart water metering systems, from add-on data loggers to digital meters, has expanded across new locations over time, offering a sophisticated means to capture detailed water usage data and support analysis (Koop et al., 2019; Mazzoni et al., 2022). These systems inform water demand management strategies to reduce costs and energy associated with water treatment and distribution (Di Mauro et al., 2022).

Several studies have investigated water consumption characteristics at different spatiotemporal scales. In many cases, water consumption data have been gathered on monthly to yearly resolutions, read manually by water

Writing – review & editing: Ashlynn S. Stillwell

utility technicians for billing purposes (Danielson, 1979; Tanverakul & Lee, 2013). However, as reported by Cominola et al. (2015), billed water data generally only allow extracting information to evaluate aggregate water consumption at a coarse spatial resolution, for example, entire city or districts, and on a coarse temporal scale, for example, monthly or seasonal. To overcome this limitation, water consumption has been investigated at finer spatiotemporal resolutions (P. W. Mayer et al., 1999; P. Mayer et al., 2004; Roberts, 2005; Mead, 2008; González et al., 2008; Willis et al., 2010; Cominola et al., 2018; Bethke et al., 2021). Progress in smart water metering technology has improved the availability of water consumption data at fine resolutions (up to seconds), revealing considerable benefits for water demand modeling (Attallah et al., 2023; Cominola et al., 2015). Yet, despite the advantages of end-use water consumption data, collecting and efficiently processing residential water consumption data remains challenging (Fagiani et al., 2015; Mazzoni et al., 2022). Intrusive monitoring on an end-use level can be both costly and time-consuming (Mazzoni et al., 2021). However, non-intrusive approaches for end-use disaggregation and classification generally require ground-truth observations from real-world data that are often unavailable (Mazzoni et al., 2021) or limited (Mauro et al., 2020) due to meter malfunctions, lack of resident participation, privacy considerations, and other challenges (Corridon, 2022; Schafer & Graham, 2002). Therefore, the availability of a reliable and realistic synthetic data set, with labeled end-use data for supervised learning (Lu et al., 2023), is highly beneficial for non-intrusive monitoring (Hosseini et al., 2017) and data-driven demand management (Kofinas et al., 2018).

Synthetic data generation is a method of generating data by models that provide accurate statistical representations of real-world observations (Barse et al., 2003), and is a powerful approach to overcome issues with ground-truth data (El Emam et al., 2020). Because synthetic data are a generalized representation of real-world data, they must be appropriately generated and follow the underlying distribution of the original data. Therefore, the algorithms that generate synthetic data must be robust and capture the patterns in the actual data (Figueira & Vaz, 2022). Synthetic data generation methods have evolved significantly over time, from simple resampling and augmentation techniques to sophisticated generative models emerging from Deep Learning such as Variational AutoEncoders (Kingma & Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), which have undergone several modifications since they were first proposed to adapt to different data structures in different domains. The main focus of models like GANs has been on computer vision tasks where the domain consists generically of images; therefore, modeling tabular data, such as water end-use consumption, can pose challenges for GANs. The generation of synthetic tabular data is a noted research need (Figueira & Vaz, 2022), and recent studies have proposed adaptive methods like Conditional Tabular GANs (CTGANs) for synthetic tabular data generation (Xu et al., 2019), enabling new opportunities for developing models to analyze smart water meter data for water end-use consumption estimation in residential households, where ground-truth labeled data are limited (Di Mauro et al., 2021; Gilbertson et al., 2011).

Several methods of analyzing fine temporal resolution data (e.g., 5 s or 10 s) have been proposed to extract water event information, such as flow trace analysis (DeOreo et al., 1996), derivative signals (Bethke et al., 2021), pressure sensing (Froehlich et al., 2011; Larson et al., 2012), and classification techniques (Heydari et al., 2022; Meyer et al., 2020; Vitter & Webber, 2018). Given the significance and potential of classification techniques, they form the main focus of this paper. Literature reviews from studies such as Chen et al. (2017) highlight that various machine learning (ML) classification algorithms possess distinct strengths and weaknesses based on the data characteristics of different study areas. Although numerous ML algorithms have been developed, their efficacy in addressing practical tasks is highly contingent upon these data characteristics, thus underscoring the need for comparative analyses. While researchers have extensively studied appliance classification in the electricity sector (Shafiq et al., 2016; Wei et al., 2015), the water sector remains relatively underexplored in this context.

Limited research has been conducted to compare residential water end-use classification algorithms. Nguyen et al. (2013) and Levasseur (2023) have shed light on the strengths and weaknesses of commercial tools such as Trace Wizard (2003), Identiflow® (Kowalski & Marshallsay, 2005), and HydroSense®. These tools were designed to address the limitations of existing methods by developing proprietary models for categorizing residential water end-use events. In a different vein, Wonders et al. (2016) conducted a comparative analysis of the classification efficiencies of three ML techniques: artificial neural networks (ANN), support vector machines (SVM), and K-nearest neighbors (KNN). Their study evaluated the implications of enlarging the training database by generating synthetic data. The data under scrutiny represented a single bathroom from a two-resident household, chosen to minimize classification errors. The primary focus of their research, however, was on the generation of synthetic data. Meanwhile, Gourmelon et al. (2021) assessed the efficacy of various supervised and

unsupervised machine learning techniques in predicting water end-use classes. Their methodology involved simulating smart meter data with the assistance of the STochastic Residential water End-use Model (STREaM), a model introduced by Cominola et al. (2018). Despite these efforts, determining an optimal algorithm based on performance remains a complex, context-dependent endeavor (Kirasich et al., 2018).

To complement this body of work, we explored supervised ML classifiers with comparative analysis of model performance under different balanced and imbalanced synthetic data scenarios. We generated synthetic data, modeled after actual labeled data obtained through a 4-week data collection process, and then assessed the performance of different models based on classification precision score, recall score, and execution time. Our approach assessed model performances under different conditions, highlighting advantages and disadvantages of different techniques for advancing residential smart water metering systems.

2. Materials and Methods

2.1. Disaggregation and Labeling

We collected fine-resolution data over a 4-week study period from September 3 to October 1, 2021, using a single-point smart water metering system installed at a fully detached, single-family residence in the Midwest United States, as documented in Heydari et al. (2022). Equipped with a custom ally® electromagnetic flow meter from Sensus, the system measured flow rate (gal/min), temperature (K), and pressure (psi) at 1-s intervals. Our analysis focused on the flow rate time series, given its significance in identifying residential water end uses. Pressure data were excluded due to potential external influences within the distribution system that introduce noise by affecting readings in ways unrelated to the household's water use (Lee et al., 2012). Furthermore, feature importance analysis by Heydari et al. (2022) indicated that the temperature recorded by the meter did not significantly determine water end uses.

During the study period, the household occupants manually recorded a water diary of labeled end uses, which were integral to our data alignment and labeling process. This diary documented six types of indoor water end uses contributing to the total household water demand: faucets, toilets, showers, refrigerator faucet, dishwasher, and washing machine. For more detailed information on the labeling process, refer to Heydari et al. (2022).

To align these water diary events with the fine-resolution time series data, we used the disaggregation model by Bethke et al. (2021) to first separate concurrent water use events and then label the individual events based on the water diary. The model isolates significant increases and decreases in water flow, represented as positive and negative derivative signals, from the vector gradients within each water use event. Through methodical iteration over these gradients, it compiles lists of consecutive non-zero gradients, which are then summed and adjusted by the duration of the event to quantify the flow rate changes in liters per minute. The collected positive and negative values are organized into separate lists for comparative analysis, facilitating the identification and segmentation of individual water use events from concurrent events in the time series. The model outputs "time of day" for each event, which was used for matching and labeling events against water diary entries. Other outputs include "duration (s)" and "average flow (gal/min)", which served as key features in subsequent stages of our study.

Despite diligent efforts in data collection and processing, we encountered challenges in accurately matching all events due to various factors, including limitations in disaggregation accuracy, omissions in the water diary by residents, and uncertainties in data representation. To preserve the integrity and specificity of our data set, we refrained from creating an "other" category for these unmatched events. We made this decision to avoid the risk of conflating distinct end-uses that, while not recorded in the diary, still belong to one of our six identified categories. We successfully labeled a total of 675 events (from a total of 965 water diary recordings), creating an imbalanced data set of 349 faucet, 161 toilet, 68 shower, 50 refrigerator faucet, 29 washing machine, and 18 dishwasher events, ensuring that each labeled event accurately reflects a specific water end-use.

2.2. Synthetic Data

Acknowledging the limitations posed by the constrained size of our data set for training robust classification models, we created additional synthetic data. This strategic augmentation is crucial for overcoming the challenges associated with the limited quantity of labeled events, enabling the exploration of various balanced and imbalanced scenarios. Such enhancement is instrumental in conducting a comprehensive comparative analysis of

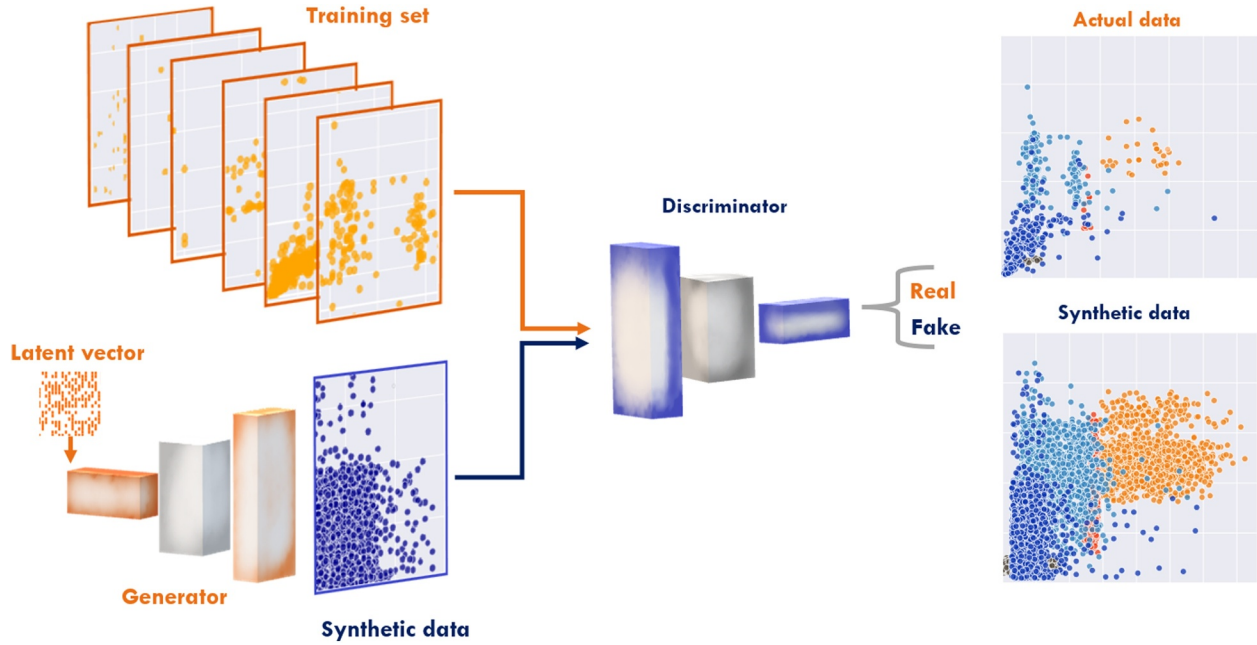

Figure 1. Generative Adversarial Network (GAN) diagram; the generator and the discriminator are both neural networks. The generator output is connected directly to the discriminator input. Through backpropagation, the discriminator's classification provides a signal that the generator uses to update its weights.

different classification models across diverse data distributions. We used CTGANs as an advanced form of GANs to generate high-quality synthetic data.

2.2.1. CTGANs: Rationale and Overview

Synthetic data generation is highly beneficial for augmenting limited data sets to robustly train machine learning models, particularly when seeking to enhance the generalizability of these models. Among the methods available, CTGANs stand out for their ability to generate synthetic tabular data that closely mirror the intricate statistical properties of original data sets. This approach is delineated in Figure 1, showcasing the adversarial interaction between the generator and discriminator components of GANs.

CTGANs were our chosen approach due to their ability to circumvent multiple issues inherent in other methods. First, our data set is characterized by non-standard distributions, including bimodal patterns in specific end uses like toilets and showers, which conventional data augmentation techniques cannot replicate accurately. In addition, the potential for correlations between key features—such as duration and average flow rate—presents a complex challenge. Traditional synthetic data generation methods typically assume independence among features, which would render them less effective in addressing the complex interdependencies that might exist in our data. CTGANs, leveraging their generative adversarial framework, excel at generating synthetic data that maintain the potential complexities in the original distributions and respect any potential inter-feature correlations, adeptly navigating these challenges. This capability was particularly beneficial for our data set, aiming to enrich the training process with a varied spectrum of synthetic examples. By effectively capturing and replicating the data's nuanced relationships, including bimodality and feature correlations, CTGANs enhance the generalizability and robustness of our predictive models, making them a superior choice for our synthetic data generation needs.

2.2.2. Application of CTGANs: Data Generation and Evaluation

We applied a CTGAN model to create synthetic data sets based on our labeled actual end uses. A limited feature set might not fully encapsulate the intricacies of residential water consumption data sets in a broader context. In certain scenarios, incorporating additional features could significantly enhance the generalizability of synthetic data.

After training the CTGAN model, we generated three data sets based on the actual data set (referred to as actual data throughout the paper).

1. Imbalanced data set: 3,800 data points with 2,000 faucet, 920 toilet, 400 shower, 300 refrigerator, 100 washing machine, and 80 dishwasher events, keeping the original proportions from the actual data.
2. 12,000 imbalanced data set: 12,000 data points with 6,316 faucet, 2,905 toilet, 1,263 shower, 947 refrigerator, 315 washing machine, and 253 dishwasher events, keeping the original proportions from the actual data.
3. 12,000 balanced data set: 12,000 data points with each end-use having 2,000 data points.

The synthetic data generation process produces a data set larger than the actual data set, making a direct one-to-one comparison with actual data impractical. Kernel density curves, which are particularly suited for non-parametric distributions, offer an insightful means to visualize and compare the distributional properties of both actual and synthetic data sets. To quantitatively assess the goodness-of-fit between the distributions of the synthetic data and the actual data for each end use, we employed the Kolmogorov-Smirnov (K-S) statistic (Massey Jr, 1951). Specifically, we utilized the K-S statistics from the `stats.ks_2samp()` method of the SciPy Python library. For evaluation, we compared the computed K-S statistics to critical values; if the K-S statistic exceeds the critical value, the null hypothesis is rejected, suggesting the two distributions are not from the same population. Due to the extensive size of our data set, the traditional p-value approach for the test has been proven less insightful (Vermeesch, 2013), prompting our emphasis on the comparison using critical values. Further details on the test methodology and relevant equations are provided in Text S5 in Supporting Information S1.

2.3. Classification Models

Supervised learning methods take an input vector comprised of n-features and map it to an associated target value or class label. The term "supervised" describes data sets (e.g., x) that contain a response label (e.g., y) and algorithms that predict y given x (Goodfellow et al., 2016). In this analysis, we consider four supervised classification models: Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (Log-reg), and a feed-forward Neural Network (NN). We describe each model in terms of its underlying mathematical structure, hyperparameters, and unique strengths and weaknesses.

2.3.1. Random Forest

RF is an ensemble-based learning algorithm that is comprised of n collections of decision trees that are de-correlated either through Bootstrap aggregation or random feature selection to reduce variance (Kirasich et al., 2018). RF models use multiple trees to compute majority votes for classification purposes in the terminal leaf nodes when making a prediction. Built off the idea of decision trees, RF models have improved prediction accuracy compared to a single tree by growing n number of trees, where each tree in the training set is sampled randomly without replacement (Breiman, 2001). The number of trees is a hyperparameter that determines how many decision trees to include in the RF ensemble. Typically, a higher number of trees leads to better performance, with the cost of increased computational complexity and memory usage. In practice, the number of trees is often chosen based on cross-validation performance or using rules of thumb.

The maximum depth is a hyperparameter that limits the depth of the decision trees in the RF model. Setting a maximum depth prevents overfitting and improves the generalization of the model. If a tree is allowed to grow too deep, it can fit noise in the training data and perform poorly on new data. However, setting the maximum depth too low can lead to underfitting and poor performance on both the training and test data. In RF, the splitting criterion is one of the most important hyperparameters, with Gini impurity and cross-entropy as the two most commonly used splitting criteria. Details on the model, including the Gini impurity and cross-entropy calculations, and grid search hyperparameter combinations can be found in Text S1 in Supporting Information S1.

2.3.2. Support Vector Machines

SVM is a machine learning algorithm used for classification and regression analysis. The goal of SVM is to find the hyperplane that separates data points into different classes as cleanly as possible. SVMs are highly accurate for classification problems and work well in complicated domains where there is a clear separation margin between classes, including both linear and non-linear data, using a kernel function to transform the data into higher

dimensions. SVMs are robust to overfitting, meaning that they can handle noise and outliers in the data well, and produce a clear decision boundary, which can help in understanding how the model is making predictions. SVMs use a subset of training points called support vectors to build the model, making the models memory-efficient and capable of handling large data sets. However, SVMs can be computationally expensive, especially when dealing with large data sets, due to the transformation of each data point into a higher-dimensional space for the kernel function. SVMs also have several hyperparameters that need to be tuned, which can make the optimization process intensive.

A comprehensive discussion on kernels, their underlying equations, and grid search is provided in Text S2 in Supporting Information S1.

2.3.3. Logistic Regression

Logistic regression operates on the principle of estimating probabilities through a logistic or sigmoid function, effectively constraining the outcome between 0 and 1 (Cessie & Houwelingen, 1992). This method meticulously analyzes the relationship between a categorical dependent variable and one or more independent variables, providing a robust framework for binary classification. To accommodate multi-class scenarios, logistic regression can be extended through strategies such as 'One-vs-Rest' (OvR) and 'One-vs-One' (OvO). These approaches allow logistic regression to navigate the complexities of multiple classes by decomposing the multi-class problem into several binary classification tasks, thereby enhancing its versatility and application scope.

Further insights into the application of logistic regression for binary classification, alongside its OvR and OvO extensions, are discussed in detail in Text S3 in Supporting Information S1. This discussion includes an exploration of the distinct loss functions associated with each strategy, the underlying mathematical formulations that govern their operations, and the implementation of grid search techniques for optimal parameter selection.

2.3.4. Neural Networks

Neural networks are generally formed by a large number of information processing units, called artificial neurons, connected with each other. The construction of artificial neurons originates from the structure of biological neurons and is mainly composed of three parts: multiple connection weights, a summation term, and a non-linear activation function. The overall structure of neural networks used in this study includes an input layer, multiple hidden layers, and an output layer. The input of each layer network is the output of the previous layer network, and the mapping process from input to output is non-linear. Through the information transmission of each layer network, the final result output is achieved. In this mode, to calculate the output of the neural network, it is necessary to carry out forward propagation step by step, input the initial vector into the input layer, and calculate all the activation values of the next layer one by one until the output layer produces the results. Such structures with one-direction information flow from the input layer through one or more hidden layers to the output layer without any feedback loops are feedforward neural networks, also known as a Multi-Layer Perceptron.

In a feedforward neural network, the input layer receives the raw data, which is then transformed through the activation functions in the hidden layers to produce the output in the final layer. Each hidden layer contains a set of neurons that are connected to the neurons in the previous and subsequent layers through weights that are learned during the training process. The output of each neuron in a hidden layer is determined by applying an activation function. In this study, a fully connected feedforward neural network trained by error–backpropagation was optimized through grid search with a number of scenarios. More detail on the choice of activation functions, optimizers, and grid search hyperparameters is provided in Text S4 in Supporting Information S1.

2.4. Experimental Environment and Design

We used the Google Colaboratory iPython notebook development environment for our experiment comparing the selected machine learning classification approaches. This environment supports TensorFlow and Keras (Chollet, 2021), and allows the implementation and training of networks using GPUs and TPUs in Google Cloud. To obtain prediction times in the Google Colaboratory environment, we used 2.12.0 and 2.12.0 versions of TensorFlow and Keras, respectively. Predictions using Google Colaboratory notebooks used an Intel® Xeon® CPU @2.20 GHz using a single core with two threads.

To compare the performance of the four models (RF, SVM, Log-reg, and NN) in all data scenarios, we recorded the execution time for both grid search optimization and training. To minimize the effect of external factors affecting execution time, we used the CPU time reported from the environment as a measure of the time spent processing a particular task. We employed a 10-fold stratified cross-validation method for model training and evaluation. This approach involves dividing the data set into 10 equal parts, or "folds," in a way that maintains the same proportion of each class label in every fold as in the original data set. This stratification ensures that each fold is a reliable representative of the whole. For each iteration of the cross-validation process, we used 9 folds (90% of the data) for training the model and the remaining fold (10% of the data) for testing. This process was repeated 10 times, with each fold serving as the testing set once.

We used three widely used metrics in classification tasks to measure the effectiveness of these models: precision, recall, and F1-score. Precision is the ratio of true positive (TP) predictions to the total number of positive predictions (i.e., true positives and false positives), measuring the fraction of the positive predictions that are correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Recall is the ratio of TP predictions to the total number of actual positive events in the data (i.e., true positives and false negatives). It measures the fraction of actual positive events that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

The F1-score is the harmonic mean of the precision and recall, providing a balanced measure of both precision and recall.

$$\text{F1-score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{3}$$

where TP represents the true positive predictions, FP represents the false positive predictions, and FN represents the false negative predictions.

3. Results

3.1. Data Generation

In this study, we used a CTGAN to generate synthetic data for each water consumption end use of shower, faucet, toilet, washing machine, dishwasher, and refrigerator faucet. We generated synthetic data, as described in Section 2.2, to increase the available data for evaluating model performance on four different data scenarios: (a) actual imbalanced data, (b) 3,800 imbalanced synthetic data, (c) 12,000 imbalanced synthetic data, and (d) 12,000 balanced synthetic data. The results of the synthetic data generation process for the 12,000 balanced synthetic data are presented in Figure 2, comparing the synthetic data generated by the CTGAN model (teal) to the actual data (gray) for each residential water end use. Additionally, Figure 3 includes density curves for the 12,000 balanced synthetic data scenario with the actual data to illustrate the goodness-of-fit of the CTGAN model. The remaining visualizations for the other data scenarios can be found in Figures S1–S7 in Supporting Information S1.

The results depicted in Figures 2 and 3 illustrate the capability of the CTGAN model to generate synthetic data that align closely with the patterns and distribution of the actual data for each end use. When the K-S statistic is smaller than the critical value, we fail to reject the null hypothesis. In the context of our study, failing to reject the null hypothesis means there is no statistical difference between the distribution of our generated synthetic data and the actual water end-use data. This result affirms the efficacy of the CTGAN model in replicating the data distribution of real-world water end-use scenarios.

3.2. Grid Search

We used these realistic synthetic data representing each water end use to evaluate the performance of machine learning models on data sets of different sizes. To optimize the performance of each model, we used grid search to

10.1029/2023WR036690

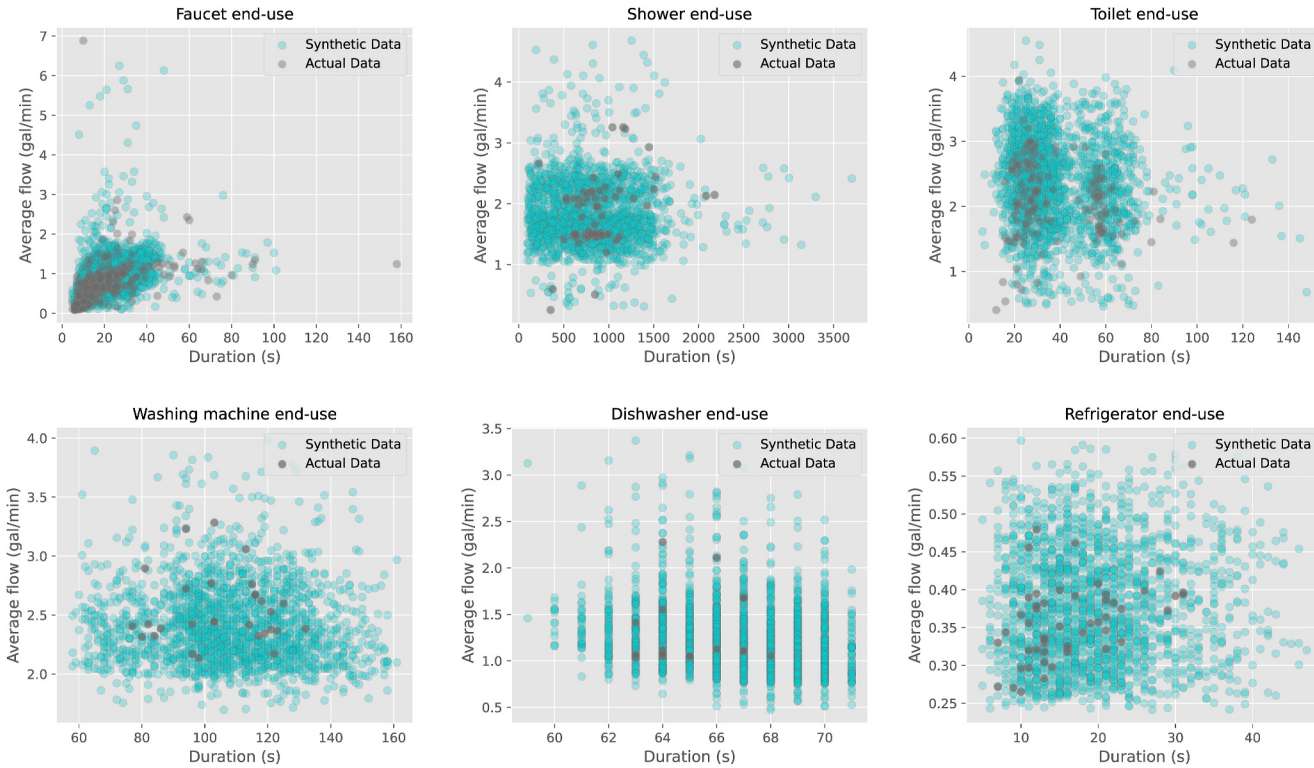

Figure 2. Comparison of 12,000 balanced synthetic data (teal) with actual data (gray) for residential water end uses: faucets, shower, toilet, washing machine, dishwasher, and refrigerator faucet.

find the best hyperparameters for each data set scenario. The best hyperparameters for the RF, SVM, and Log-reg models are shown in Table 1, based on the grid search process optimized on the four data set scenarios. For the RF model, most scenarios favored a maximum tree depth of 10, with the exception of the data set with 3,800 imbalanced entries, which preferred a depth of 5. The criterion varied between "gini" and "entropy", dependent on data set balance. The SVM model predominantly utilized the rbf kernel across all data sets, with variations in the regularization parameter "C". Lastly, the Log-reg model consistently yielded the best results with a "multinomial" multi-class strategy and a balanced class weight. The chosen solver alternated between "newton-cg" and "lbfgs" based on the specific data set. Details on these hyperparameters are available in Supporting Information S1.

We did not include the NN model in Table 1 due to its different architectural nature compared to the other models. The optimal configuration for the NN model was determined to be a 6-layer network with 30 neurons, using the Rectified Linear Unit (ReLU) activation function for the hidden layers and Softmax for the output layer, and the Adam optimizer to minimize the loss function during training.

We further evaluated the models from the perspective of execution time under the four data set scenarios as shown in Figure 4. In Figure 4, the vertical axis is CPU time in seconds, scaled logarithmically for ease of visual comparison with considerable time range differences. From left to right, the bars represent the actual data set, the 3,800 imbalanced data set, the 12,000 imbalanced data set, and the 12,000 balanced data set. Overall, the RF and Log-reg models run the fastest, with execution times of less than 12 s in all conditions. The SVM and NN models are comparatively slower, with execution times ranging from 45 s to over 50 min for the SVM model, and 17–26 min for the NN model.

Our analysis showed that data set size predominantly influences model execution time. Notably, the SVM model exhibited a marked increase in execution time as the data set size expanded, more so than the RF, Log-reg, and NN models. Although all models showed increased execution times with larger data sets, a minor but interesting increase was also observed when comparing execution times between the 12,000 imbalanced to the 12,000 balanced data sets. This finding suggests that while data set size is the primary factor affecting execution time, data imbalance also contributes to variations in execution times, albeit to a lesser extent.

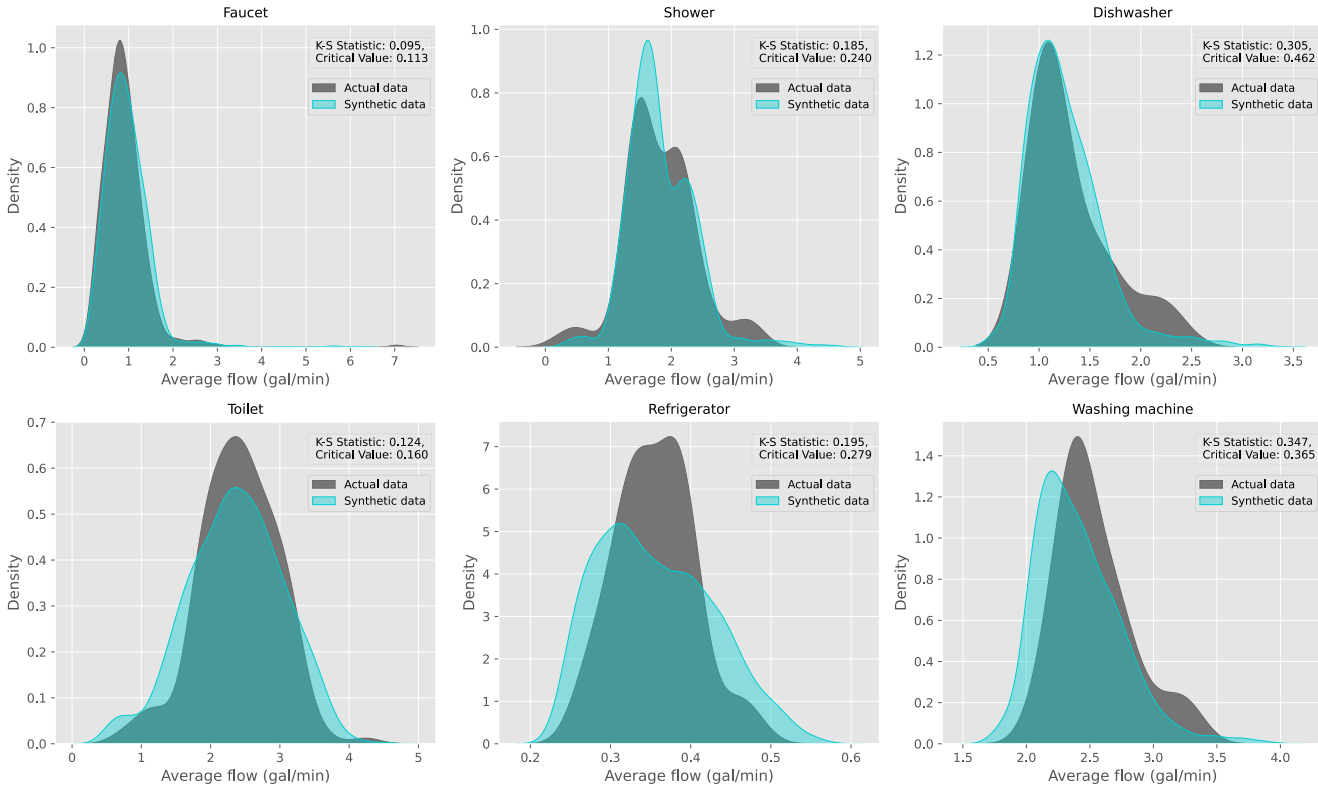

Figure 3. Kernel density distributions of average flow (gal/min) comparing the 12,000 balanced synthetic data (teal) generated by the CTGAN model with the actual data (gray). The displayed K-S statistics and critical value quantify the similarity between the distributions, quantitatively assessing the goodness-of-fit.

The SVM approach is a powerful machine learning algorithm that works well with complex data and has a high level of accuracy, but it involves solving a complex optimization problem involving quadratic programming that can be computationally intensive, as shown in the results in Figure 4. Similarly, the NN approach is a deep learning algorithm that involves training multiple layers of artificial neural networks, which requires significant

Table 1
Best Hyperparameters for Each Data Set for the Considered Machine Learning Models

RF	n_estimators	max_depth	min_split	min_leaf	criterion
Actual data	100	10	2	3	gini
12,000 balanced	50	10	2	3	gini
12,000 imbalanced	200	10	2	3	entropy
3,800 imbalanced	200	5	2	3	gini
SVM	***C***	**d_function_shape**	**degree**	**gamma**	**kernel**
Actual data	50	ovr	-	auto	rbf
12,000 balanced	10	ovr	-	auto	rbf
12,000 imbalanced	50	ovr	-	auto	rbf
3,800 imbalanced	50	ovr	-	auto	rbf
Log-reg	***C***	**multi_class**	**fit_intercept**	**class_weight**	**solver**
Actual data	50	multinomial	True	balanced	newton-cg
12,000 balanced	100	multinomial	True	balanced	lbfgs
12,000 imbalanced	50	multinomial	True	balanced	lbfgs
3,800 imbalanced	50	multinomial	True	balanced	newton-cg

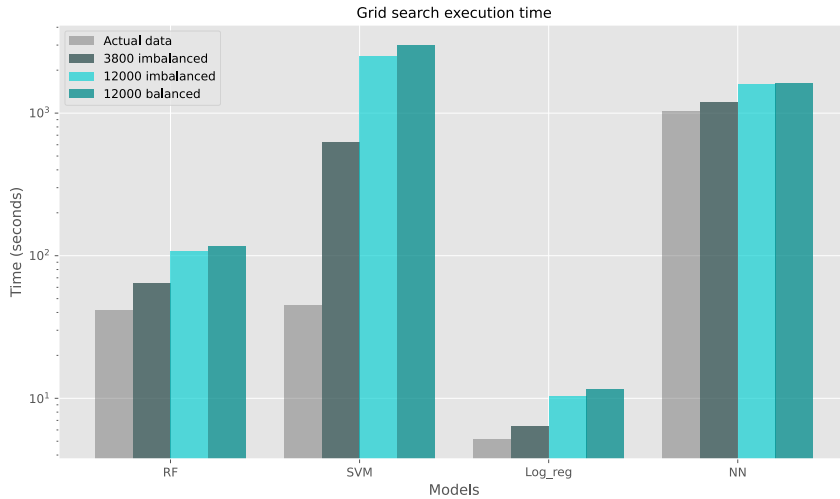

Figure 4. Grid search optimization execution time varies by scenario for the random forest (RF), support vector machine (SVM), logistic regression (Log-reg), and neural network (NN).

computational resources. Drawing insights from previous research, a more efficient approach to SVM would be to not use all of the data, and instead keep the data near the decision boundaries and omit redundant data points since SVM models mostly benefit from learning data features near the decision boundaries (Kumar & Gopal, 2010). Another approach would be to avoid using grid search for SVM model optimization and use faster techniques such as randomized search (Mantovani et al., 2015).

In contrast, RF and Log-reg are simpler models that are relatively efficient to optimize. The RF approach is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. Therefore, the optimization of RF involves tuning a few hyperparameters, such as the number of trees and the maximum depth of each tree, which can be done efficiently using grid search. Similarly, the Log-reg approach is a simple linear classification algorithm that involves fitting a logistic function to the input data. The optimization of Log-reg involves tuning a small number of hyperparameters, which can be done efficiently using grid search or other optimization techniques. These findings align with previous studies that have shown RF models to be relatively robust against parameter specifications during optimization compared to other machine learning algorithms (Couronné et al., 2018; Probst et al., 2018).

Another possible reason for longer optimization times for the SVM and NN models is sensitivity to the choice of hyperparameters. SVMs and NNs have several hyperparameters that need to be tuned carefully to achieve optimal performance, such as the kernel function, regularization parameter, learning rate, and the number of hidden layers. For instance, when we omitted the 'poly' kernel option from the original options for the SVM model once we observed that the polynomial kernel was not the best kernel in either scenario, the execution time for the SVM model decreased significantly (i.e., from the initial 50 min to 17 min). With only two features, using an SVM model with kernel functions might not be necessary or might not lead to significant improvements in performance since the main advantage of kernel functions is to implicitly map the data to a higher-dimensional space for separation. However, for two-dimensional data, it can be relatively straightforward to find a separating hyperplane in the original feature space. Using kernel functions in SVM models can still be useful for two-dimensional data when the relationship between the features and the target variable is non-linear or when the classes are not linearly separable. In these cases, a non-linear kernel function such as the 'rbf' kernels used in this study can implicitly map the data to a higher-dimensional space where it can be easier to find a separating hyperplane.

3.3. Model Performance

After optimizing each model, we examined model performances based on precision, recall, and F1-score. Figure 5 shows the precision of the models for shower, toilet, faucet, washing machine, dishwasher, and refrigerator faucet events. The precision was measured on all four data set scenarios. Overall, the RF model performed better than other models on most of the residential water end uses, and all of the evaluated models showed high precision for

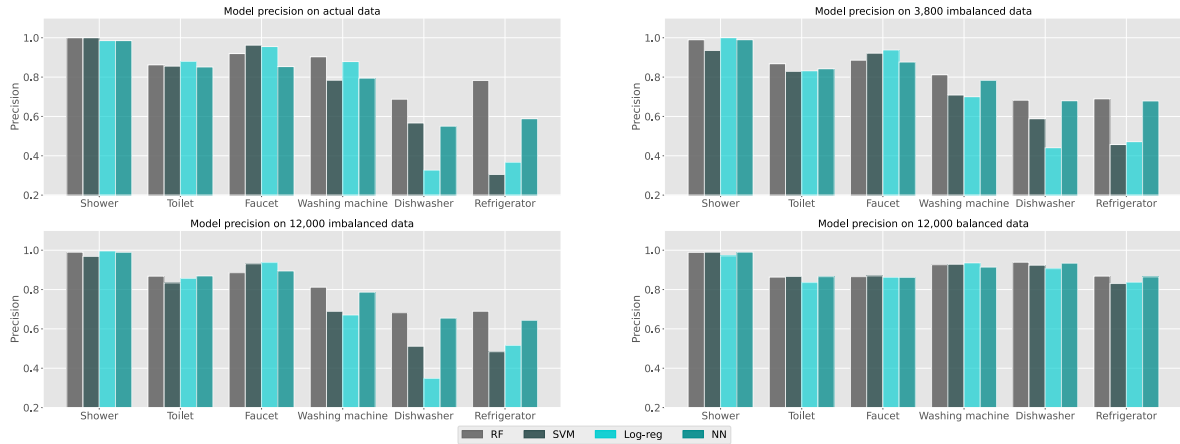

Figure 5. Precision score for each model in different scenarios for the random forest (RF), support vector machine (SVM), logistic regression (Log-reg), and neural network (NN).

shower events across all four data set scenarios, ranging from 0.93 to 1. The lowest precision for all evaluated models was for the dishwasher and refrigerator end-use events, ranging from 0.30 to 0.78, with a considerable improvement to 0.83 to 0.93 with the 12,000 balanced data set scenario.

Using the actual data, the RF and SVM models performed better on most of the water end uses, with highest precision on shower and faucet events. All models perform consistently high in detecting toilet events with precision scores near 0.86. The Log-reg model had the lowest precision values for dishwasher and refrigerator events, but outperformed the SVM and NN models on toilet and washing machine events. The SVM model precision also dropped considerably from 0.78 on washing machine events to 0.56 and 0.30 for dishwasher and refrigerator events, respectively. The overall results show that all the evaluated models had relatively similar performance for the more frequent residential water end uses (shower, toilet, and faucet events). The Log-reg model performed best with the 3,800 imbalanced data set, particularly for shower and faucet events. It surpassed all other models in these categories but had lower precision for other uses compared to its results on the full data set, except for showers. Precision scores for all evaluated models improved in the 12,000 balanced data set scenario, indicating that the class imbalance negatively affected the models' performance. However, the RF model was less affected by the imbalanced data sets compared to the SVM, Log-reg, and NN models. Moreover, balancing the data set had the most effect on the Log-reg and NN models, whose precision values increased considerably with the balanced data set scenario. Overall, the RF and SVM models were the top-performing models considering the overall precision with different data set scenarios.

Figure 6 shows the recall values for the evaluated classification models on water end uses. While all models showed relatively similar recall scores on the 12,000 balanced data set scenario, the Log-reg and SVM models outperformed the RF and NN models in the other data set scenarios. For example, on the actual data, both the Log-reg and SVM models achieved a perfect recall of 1 on washing machine events, the highest among all models. However, for toilet events, the RF and NN models achieved higher recall scores on all data sets. One noticeable result was the NN model recall of 0.2 for refrigerator events on actual data, which implied the amount of data provided on the actual set was not sufficient for the model to distinguish refrigerator events from faucet events. Looking at the effect of data size on model performance, we show that the recall of all evaluated models generally improved as the size of the data set increased, even for the NN model with refrigerator end-use events. However, the RF and SVM models were less affected by data size changes than the Log-reg and NN models in terms of performance.

In the 3,800 imbalanced scenario, the RF model exhibited robust recall scores across various water end uses, with values spanning from 0.81 for toilet events to a high of 0.98 for shower events. The SVM model was particularly proficient for the refrigerator end use, achieving an impressive recall score of 0.99. The performances of Log-reg and NN models were more diverse; while they showcased commendable recall scores for specific end uses, their efficacy was less consistent across end uses.

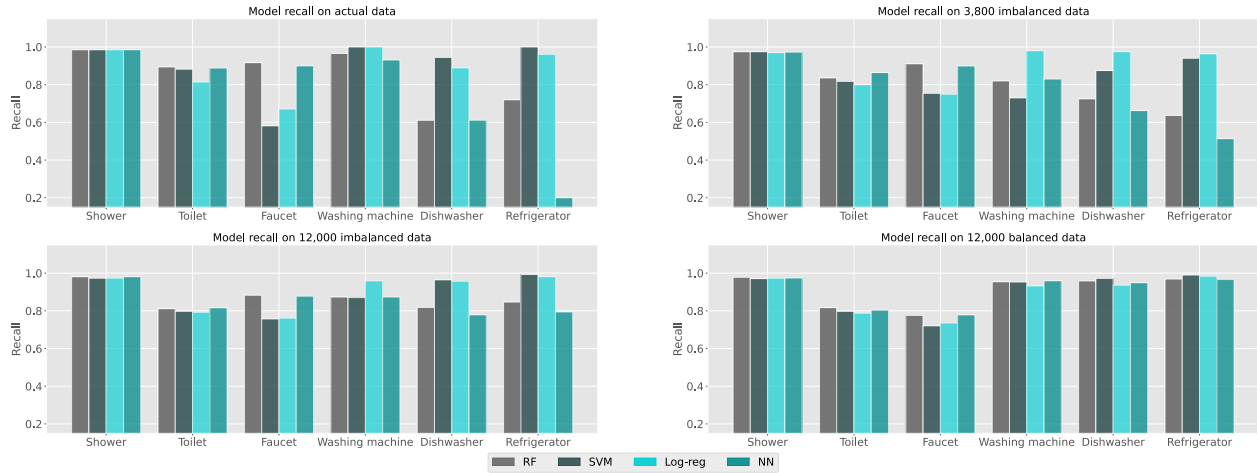

Figure 6. Recall score for each model in different scenarios for the random forest (RF), support vector machine (SVM), logistic regression (Log-reg), and neural network (NN).

Upon transitioning to the balanced scenario, there was a marked enhancement in the overall performance metrics of all models. RF consistently demonstrated strong recall values, ranging between 0.78 for faucet events and 0.98 for shower events. SVM maintained high performances, peaking with a recall score of 0.99 for refrigerator events. However, the Log-reg and NN models exhibited heterogeneous results across the various end uses, highlighting the pivotal role of model selection and data balancing techniques in determining success. Our observations stress the significance of mitigating class imbalance through data set balancing strategies, thereby augmenting the robustness and reliability of machine learning models in the realm of residential water end use categorization tasks.

High recall or precision scores alone do not necessarily indicate good model performance. For instance, the high scores of recall for the Log-reg and SVM models on refrigerator events mean that the model is able to identify almost all instances of refrigerator end uses (i.e., very few false negatives), but the low precision score on the same end use suggests that the models are also identifying other activities as refrigerator events (i.e., a high number of false positives). In this case, it is important to consider the tradeoff between precision and recall, which is captured in the metric of F1-score. Figure 7 illustrates the weighted F1-score values and model execution time, with each color representing a data set scenario. In the model performance (Figure 7, left), the RF model performed better overall than other models on all data sets, with an F1-score of 0.90 on actual data, 0.91 on the 12,000 balanced data set, 0.87 on the 12,000 imbalanced data set, and 0.87 on the 3,800 imbalanced data set. However, the NN model performs competitively with an F1-score of 0.91 for the 12,000 balanced data set. Additionally, for the 12,000 imbalanced data set, the Log-reg model shows better performance than the SVM model, with F1-scores of 0.83 versus 0.81. The SVM model performance on all data sets is lower compared to other models, with F1-scores

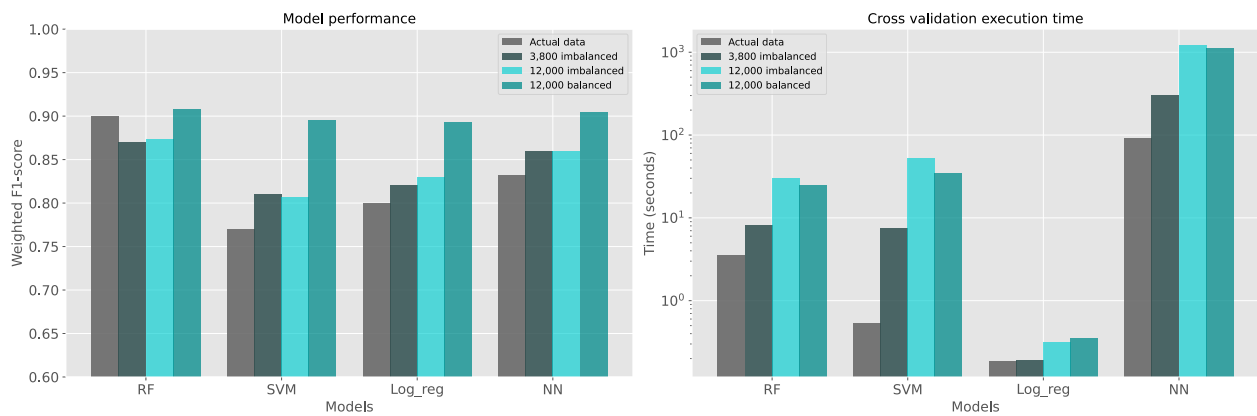

Figure 7. Model performances and execution time for the random forest (RF), support vector machine (SVM), logistic regression (Log-reg), and neural network (NN).

ranging from 0.77 to 0.90. Furthermore, the F1-scores of all models show a decline in performance when applied to imbalanced data sets, as shown in the 12,000 imbalanced and 3,800 imbalanced data set scenarios.

Next, we compared the computational costs of the models in terms of execution times, shown in Figure 7 (right), with the vertical axis scaled logarithmically for ease of visual comparison with considerable time range differences. The execution time reported in Figure 7 is the CPU time for the models to perform a 10-fold stratified cross-validation on the four data set scenarios. The Log-reg model was the fastest model, taking milliseconds to perform 10 rounds of training and testing, followed by the RF model, with execution times ranging from 3.48 to 29.7 s. One noticeable observation was the substantial difference between the optimization and cross-validation execution times of the SVM model. While the SVM model had orders of magnitude time difference in the optimization process, it had a comparable training and testing time range to the Log-reg and RF models. The NN model, however, took orders of magnitude longer than the other models to train, ranging from 1 min and 31 s to 20 min and 24 s. This marked difference in execution times can be attributed to the inherent complexity of NN models, which require iterative optimization over a large number of parameters and batches of data.

In summary, our comparative analysis and evaluation of four classification models for residential water end uses reveals that the RF model exhibits the best overall performance across all data set scenarios considering both performance and computational efficiency. Although all models show competitive performance on the balanced data set, specifically the NN model, the RF model surpasses the NN model with a notable difference in execution time for our specific context and application. Similar comparative studies in other fields have found neural networks to take longer to train (Ahmad et al., 2017; Nitze et al., 2012; Siroky, 2009) with performance superior to other classifiers once trained on large data sets. However, in contexts with simpler feature sets or smaller data sets, the superiority of tree-based models over neural networks has been consistently demonstrated in other disciplines (Sadri et al., 2018; Sarker et al., 2019). Our analysis extends these findings to the domain of water end-use classification, a context characterized by low-complexity feature spaces.

4. Discussion

4.1. Implications of a Comparative Analysis

The superior performance of the RF model across various data set scenarios underscores its robustness as an essential tool for residential water end-use classification. This outcome is notably consistent with findings from broader studies, such as the one conducted by Attallah et al. (2023), which assessed the efficacy of multiple machine learning models across a wider range of households, further implying the generalizability of RF's superior performance in the context of residential water end-use data. The decision tree-based nature of the RF model contributes significantly to its adaptability and accuracy in classifying diverse water end uses, making it a dependable choice for utility managers and researchers aiming for precision in water use monitoring and conservation strategies. Additionally, the observed improvements in NN model performance with larger, more balanced data sets underscore the critical role of data set quality and composition in securing optimal classification results, reinforcing the importance of thoughtful data preparation in the application of machine learning to water end-use classification.

However, despite the potential of NNs for nuanced classification tasks, our analysis indicates that they are not the most suitable choice for residential water end-use classification due to their high computational costs. We observed significant trade-offs between model complexity, execution time, and classification accuracy throughout our study. For instance, the Log-reg model has demonstrated itself as a computationally efficient alternative for specific end uses, such as shower detection, without notably sacrificing accuracy. This efficiency is particularly crucial in scenarios where rapid processing is a priority or when computational resources are constrained. Such findings underscore the necessity of adopting a balanced approach in model selection, carefully considering the unique requirements and constraints of water end-use classification tasks. This strategy ensures that the selected models not only achieve the desired accuracy but also align with the operational and resource considerations of the application context.

The realistic representation of water end uses in the synthetic data sets generated by CTGANs underscores the significant potential of synthetic data generation in addressing the limitations posed by real-world data. Gathering labeled data for water end-use classification poses substantial challenges—it is not only a labor-intensive effort but can also be intrusive and costly, with privacy concerns further complicating the data collection process. This

approach mitigates the challenges associated with acquiring extensive labeled data sets, helping to broaden the applicability of water end-use classification algorithms to a wider array of contexts. Such expansion includes regions or households with unique water use patterns not captured in the original data set, thereby overcoming traditional barriers to data collection and ensuring models are robust and adaptable to diverse scenarios.

We separately evaluated precision, recall, and the F1-score, each offering unique insights into model efficacy in water end-use classification. The importance of precision and recall varies depending on the specific operational goals and needs of utilities. High recall is essential for minimizing missed detections of water end-use events, crucial for tasks like leak detection or water conservation where every instance is critical. Conversely, high precision is imperative to ensure the accuracy of classifications, reducing the risk of false positives that could lead to unnecessary interventions or operational inefficiencies. Evaluating these metrics separately, alongside the F1-score—a balanced measure of precision and recall—provides a nuanced understanding of model performance, enabling utility managers to make informed decisions that align with their specific objectives and constraints.

4.2. Data and Model Limitations

The strengths and findings of this study must be contextualized within its limitations. Notably, the granularity and scope of the data present potential constraints in the broader application of our findings. The key limitations of our data set and model focus around context and scope of the actual data.

Our data are derived from a specific context, originating from a single-family household in the Midwest United States. This geographic specificity introduces inherent climatic, cultural, and infrastructural nuances, which limit the universal applicability of our findings. While offering valuable insights into residential water use within this region, our study does not fully represent the water consumption dynamics that may be present in areas with differing climates or infrastructures, such as regions with increased outdoor water use. Furthermore, the appliances and fixtures (end uses) in the study home, characterized by their unique manufacturing, design, and installation attributes, underscore the potential for varied water consumption patterns across different regions. For instance, the design of a faucet in accordance with typical U.S. premise plumbing sizes may exhibit significant differences in water flow or use duration when compared to faucet fixtures in other countries. While our modeling approach has broad implications, these end-use disparities highlight the challenges in extrapolating our specific results to diverse settings. Although indoor water consumption patterns maintain a degree of internal consistency (Abdallah & Rosenberg, 2014), local behaviors, infrastructures, and access conditions contribute additional complexity in cross-location comparisons. External factors such as regional water scarcity, cultural norms, and government regulations further influence water usage dynamics, underscoring the nuanced nature of applying our findings beyond the immediate study context.

Moreover, the temporal boundary of our data set, spanning from mid-September to mid-October 2021, captures only the water use patterns of a specific season. Factors like warmer summers or colder winters can introduce seasonal variations in water consumption behavior, potentially influencing the features on which we trained our models. Furthermore, the nuances of individual households introduce another layer of variability. Each household exhibits unique behavioral dynamics driven by habits, the number of residents, or specific needs. Transient factors like hosting guests or sudden changes in routines, which our study did not account for, can also influence water consumption features collected through our ground-truth data collection.

Lastly, from a modeling perspective, given the limited scope of our data collection, there exists a tangible risk of model overfitting. Models that excel within the confines of a study's data set might falter when faced with new or diverse data (Levasseur, 2023), emphasizing the need for diversified data sets for comprehensive training and validation. Additionally, while our selective approach to labeling—excluding data that could not be matched due to challenges stated in Section 2.1—avoids training on potentially misleading or noisy data, these exclusions can introduce sample bias. In this study, we avoid bias by ensuring our data set captures a comprehensive range of appliance types, including less frequently used appliances like dishwashers, with sufficient observations for the model to learn their distinct patterns. The disaggregation model might also incorrectly disaggregate events with overlapping characteristics or atypical signatures, suggesting areas for potential improvement in ensuring a more inclusive and comprehensive data set.

We acknowledge that while the detailed results of our study, such as the optimal hyperparameters for each model or the specific performance metrics, may not directly translate to new data sets, the overarching methodology and

the adopted approach can have broader applicability. This broader relevance stems from the inherent characteristics of residential water end-use data, which are typically structured in a tabular format with features like time of day, day of the week, duration, and average flow. Such data structure similarities across different contexts suggest that our methodological framework, focusing on the selection, application, and evaluation of machine learning models for water end-use classification, can be applied more broadly. Despite the unique attributes of the data set from a single household, the foundational aspects of our approach provide a blueprint that can guide the application of machine learning in water end-use classification across varied residential settings, underscoring the potential for broader adoption and adaptation of our research findings.

5. Summary and Conclusion

In this study, we evaluated the performance of four classification models (Random Forest, RF; Support Vector Machines, SVM; Logistic Regression, Log-reg; and Neural Networks, NN) on actual and synthetic labeled residential water consumption data sets, finding the RF model as optimal for water end-use classification. The synthetic data (generated using Conditional Tabular Generative Adversarial Networks, CTGANs) were found to be realistic and capture the patterns of the actual data for each water end use. We used grid search to identify the best hyperparameters for each model on four different data set scenarios: (a) actual imbalanced data with 675 labeled observations, (b) 3,800 imbalanced synthetic data, (c) 12,000 imbalanced synthetic data, and (d) 12,000 balanced synthetic data. The models were then trained based on the optimized hyperparameters.

Our findings show that the NN model does not have a competitive performance compared to the RF model when the data set of interest is small (actual data set; $n = 675$), even though the NN model marginally outperforms SVM and Log-reg in all data set scenarios based on an overall macro F1-score. The NN model performance improved to nearly the RF model performance with additional data observations and balanced classes. However, the NN model high performance comes with the tradeoff of notably longer execution times compared to the RF model, both in optimization and training processes. Comparing the NN model performance on the 3,800 and 12,000 imbalanced data sets, our study shows that an increase in data size does not necessarily improve model performance in this context with heterogeneous classes.

In addition, our analysis indicated that SVM and Log-reg models can compete with RF and NN in performance on balanced data sets but tend to underperform when data are imbalanced. Despite this lower performance, their shorter training times compared to NN models suggest computational efficiency for specific classification goals. Notably, while SVM models offer competitive performance, their optimization can be time-prohibitive due to the complex nature of solving quadratic programming problems and the necessity of tuning multiple hyperparameters. Conversely, the Log-reg model stands out for its computational efficiency and practical utility in projects focused on singular end-uses, such as shower events, making it an attractive option for targeted water end-use classification tasks.

Overall, the RF model outperformed other models, considering ease of optimization, performance metrics (precision, recall, F1-score), and execution time across all data set scenarios. Additionally, the RF model was less dependent on class imbalancement. Despite the widespread use of artificial neural networks, our study found that the NN model did not demonstrate substantial performance gains compared to the tree-based model in the context of water end-use classification tasks. This study contributes to the existing literature on residential water end-use classification algorithms with a comprehensive analysis of supervised machine learning classifiers. We provide synthetic end-use data representative of actual water use events and reveal insights into the strengths and weaknesses of various supervised machine learning classifiers for residential water end-use classification to help guide future research in this area.

Acknowledgments
We thank the study home occupants for their participation in recording water consumption events. The custom ally® water meter used in this analysis was provided by Sensus. This work was supported by the National Science Foundation, Grant CBET-1847404; the opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data Availability Statement

The data that support the findings of this study are openly available at the following URL/DOI: https://stillwell.cee.illinois.edu/data/. All codes developed for this study are available upon request.

References

Abdallah, A. M., & Rosenberg, D. E. (2014). Heterogeneous residential water and energy linkages and implications for conservation and management. *Journal of Water Resources Planning and Management*, *140*(3), 288–297. https://doi.org/10.1061/(asce)wr.1943-5452.0000340

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Attallah, N. A., Horsburgh, J. S., & Bastidas Pacheco, C. J. (2023). An open-source, semisupervised water end-use disaggregation and classification tool. *Journal of Water Resources Planning and Management*, *149*(7), 04023024. https://doi.org/10.1061/jwrmd5.wreng-5444

Barse, E. L., Kvarnstrom, H., & Jonsson, E. (2003). Synthesizing test data for fraud detection systems. In *19th Annual Computer Security applications conference, 2003. proceedings* (pp. 384–394).

Bethke, G. M., Cohen, A. R., & Stillwell, A. S. (2021). Emerging investigator series: Disaggregating residential sector high-resolution smart water meter data into appliance end-uses with unsupervised machine learning. *Environmental Sciences: Water Research & Technology*, *7*(3), 487–503. https://doi.org/10.1039/d0ew00724b

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Cessie, S. L., & Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society - Series C: Applied Statistics*, *41*(1), 191–201. https://doi.org/10.2307/2347628

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., et al. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, *151*, 147–160. https://doi.org/10.1016/j.catena.2016.11.032

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D. E., & Abdallah, A. M. (2018). Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environmental Modelling and Software*, *102*, 199–212. https://doi.org/10.1016/j.envsoft.2017.11.022

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., & Rizzoli, A. E. (2015). Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling and Software*, *72*, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

Corridon, P. R. (2022). Intravital microscopy datasets examining key nephron segments of transplanted decellularized kidneys. *Scientific Data*, *9*(1), 561. https://doi.org/10.1038/s41597-022-01685-9

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*, 1–14. https://doi.org/10.1186/s12859-018-2264-5

Danielson, L. E. (1979). An analysis of residential demand for water using micro time-series data. *Water Resources Research*, *15*(4), 763–767. https://doi.org/10.1029/wr015i004p00763

DeOreo, W. B., Heaney, J. P., & Mayer, P. W. (1996). Flow trace analysis to access water use. *Journal - American Water Works Association*, *88*(1), 79–90. https://doi.org/10.1002/j.1551-8833.1996.tb06487.x

Di Mauro, A., Cominola, A., Castelletti, A., & Di Nardo, A. (2021). Urban water consumption at multiple spatial and temporal scales. a review of existing datasets. *Water*, *13*(1), 36. https://doi.org/10.3390/w13010036

Di Mauro, A., Venticinque, S., Santonastaso, G. F., & Di Nardo, A. (2022). WEUSEDTO — Water End USE Dataset and TOols: An open water end use consumption dataset and data analytics tools. *SoftwareX*, *20*, 101214. https://doi.org/10.1016/j.softx.2022.101214

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). *Practical synthetic data generation: Balancing privacy and the broad availability of data*. O'Reilly Media.

Fagiani, M., Squartini, S., Gabrielli, L., Spinsante, S., & Piazza, F. (2015). A review of datasets and load forecasting techniques for smart natural gas and water grids: Analysis and experiments. *Neurocomputing*, *170*, 448–465. https://doi.org/10.1016/j.neucom.2015.04.098

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Froehlich, J., Larson, E., Saba, E., Campbell, T., Atlas, L., Fogarty, J., & Patel, S. (2011). A longitudinal study of pressure sensing to infer real-world water usage events in the home. *Proceedings Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15*. (Vol. 9), 50–69.

Gilbertson, M., Hurlimann, A., & Dolnicar, S. (2011). Does water context influence behaviour and attitudes to water conservation? *Australasian Journal of Environmental Management*, *18*(1), 47–60. https://doi.org/10.1080/14486563.2011.566160

González, F. C., Rueda, T. M., & Les, S. O. (2008). *Microcomponentes y factores explicativos del consumo doméstico de agua en la comunidad de Madrid*. Canal de Isabel II.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Gourmelon, N., Bayer, S., Mayle, M., Bach, G., Bebber, C., Munck, C., et al. (2021). Implications of experiment set-ups for residential water end-use classification. *Water*, *13*(2), 236. https://doi.org/10.3390/w13020236

Heydari, Z., Cominola, A., & Stillwell, A. S. (2022). Is smart water meter temporal resolution a limiting factor to residential water end-use classification? A quantitative experimental analysis. *Environmental Research: Infrastructure and Sustainability*, *2*(4), 045004. https://doi.org/10.1088/2634-4505/ac8a6b

Hosseini, S., Kelouwani, S., Agbossou, K., Cardenas, A., & Henao, N. (2017). A semi-synthetic dataset development tool for household energy consumption analysis. In *2017 IEEE International Conference on Industrial technology (ICIT)* (pp. 564–569). https://doi.org/10.1109/ICIT.2017.7915420

Inman, D., & Jeffrey, P. (2006). A review of residential water conservation tool performance and influences on implementation effectiveness. *Urban Water Journal*, *3*(3), 127–143. https://doi.org/10.1080/15730620600961288

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), 9.

Kofinas, D. T., Spyropoulou, A., & Laspidou, C. S. (2018). A methodology for synthetic household water consumption data generation. *Environmental Modelling & Software*, *100*, 48–66. https://doi.org/10.1016/j.envsoft.2017.11.021[

Koop, S., Van Dorssen, A., & Brouwer, S. (2019). Enhancing domestic water conservation behaviour: A review of empirical studies on influencing tactics. *Journal of Environmental Management*, *247*, 867–876. https://doi.org/10.1016/j.jenvman.2019.06.126

Kowalski, M., & Marshallsay, D. (2005). Using measured microcomponent data to model the impact of water conservation strategies on the diurnal consumption profile. *Water Science and Technology: Water Supply*, *5*(3–4), 145–150. https://doi.org/10.2166/ws.2005.0094

Kumar, M. A., & Gopal, M. (2010). A hybrid SVM based decision tree. *Pattern Recognition*, *43*(12), 3977–3987. https://doi.org/10.1016/j.patcog.2010.06.010

Larson, E., Froehlich, J., Campbell, T., Haggerty, C., Atlas, L., Fogarty, J., & Patel, S. N. (2012). Disaggregated water sensing from a single, pressure-based sensor: An extended analysis of HydroSense using staged experiments. *Pervasive and Mobile Computing*, *8*(1), 82–102. https://doi.org/10.1016/j.pmcj.2010.08.008

Lee, J., Lohani, V. K., Dietrich, A. M., & Loganathan, G. (2012). Hydraulic transients in plumbing systems. *Water Science and Technology: Water Supply*, *12*(5), 619–629. https://doi.org/10.2166/ws.2012.036

Levasseur, G. (2023). Human activity recognition in water meter data.

Lu, Y., Wang, H., & Wei, W. (2023). Machine learning for synthetic data generation: A review. *arXiv preprint arXiv:2302.04062*.

Mantovani, R. G., Rossi, A. L., Vanschoren, J., Bischl, B., & De Carvalho, A. C. (2015). Effectiveness of random search in SVM hyper-parameter tuning. In *2015 International Joint Conference on Neural Networks (ijcnn)* (pp. 1–8).

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.2307/2280095

Mauro, A. D., Nardo, A. D., Santonastaso, G. F., & Venticinque, S. (2020). Development of an IoT system for the generation of a database of residential water end-use consumption time series. *Environmental Sciences Proceedings*, *2*(1). https://doi.org/10.3390/environsciproc2020002020

Mayer, P., DeOreo, W., Towler, E., Martien, L., & Lewis, D. (2004). *Tampa water department residential water conservation study: The impacts of high efficiency plumbing fixture retrofits in single-family homes*. A Report Prepared for Tampa Water Department and the United States Environmental Protection Agency.

Mayer, P. W., DeOreo, W. B., Opitz, E. M., Kiefer, J. C., Davis, W. Y., Dziegielewski, B., & Nelson, J. O. (1999). Residential end uses of water.

Mazzoni, F., Alvisi, S., Blokker, M., Buchberger, S. G., Castelletti, A., Cominola, A., et al. (2022). Investigating the characteristics of residential end uses of water: A worldwide review. *Water Research*, *230*, 119500. https://doi.org/10.1016/j.watres.2022.119500

Mazzoni, F., Alvisi, S., Franchini, M., Ferraris, M., & Kapelan, Z. (2021). Automated household water end-use disaggregation through rule-based methodology. *Journal of Water Resources Planning and Management*, *147*(6), 04021024. https://doi.org/10.1061/(asce)wr.1943-5452.0001379

McDonald, R., Douglas, I., Grimm, N., Hale, R., Revenga, C., Gronwall, J., & Fekete, B. (2011). Implications of fast urban growth for freshwater provision. *Ambio*, *40*(5), 437–446. https://doi.org/10.1007/s13280-011-0152-6

McDonald, R. I., Weber, K., Padowski, J., Flörke, M., Schneider, C., Green, P. A., et al., (2014). Water on an urban planet: Urbanization and the reach of urban water infrastructure. *Global Environmental Change*, *27*, 96–105. https://doi.org/10.1016/j.gloenvcha.2014.04.022

Mead, N. (2008). Investigation of domestic water end use.

Meyer, B. E., Jacobs, H. E., & Ilemobade, A. (2020). Extracting household water use event characteristics from rudimentary data. *Journal of Water Supply: Research & Technology - Aqua*, *69*(4), 387–397. https://doi.org/10.2166/aqua.2020.153

Nguyen, K. A., Zhang, H., & Stewart, R. A. (2013). Development of an intelligent model to categorise residential water end use events. *Journal of Hydro-Environment Research*, *7*(3), 182–201. https://doi.org/10.1016/j.jher.2013.02.004

Nitze, I., Schulthess, U., & Asche, H. (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. *Proceedings of the 4th GEOBIA*(Vol. 79, p. 3540).

Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). Tunability: Importance of hyperparameters of machine learning algorithms. *arXiv preprint arXiv:1802.09596*.

Roberts, P. (2005). Yarra Valley water: 2004 residential end use measurement study. *Yarra Valley Water Melbourne*.

Sadri, A., Salim, F. D., Ren, Y., Shao, W., Krumm, J. C., & Mascolo, C. (2018). What will you do for the rest of the day? An approach to continuous trajectory prediction. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, *2*(4), 1–26. https://doi.org/10.1145/3287064

Sarker, I. H., Kayes, A., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, *6*(1), 1–28. https://doi.org/10.1186/s40537-019-0219-y

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989x.7.2.147

Shafiq, M., Yu, X., Laghari, A. A., Yao, L., Karn, N. K., & Abdessamia, F. (2016). Network traffic classification techniques and comparative analysis using machine learning algorithms. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)* (pp. 2451–2455).

Siroky, D. S. (2009). Navigating random forests and related advances in algorithmic modeling.

Tanverakul, S. A., & Lee, J. (2013). Residential water demand analysis due to water meter installation in California. In *World Environmental and Water Resources Congress 2013: Showcasing the future* (pp. 936–945).

Vermeesch, P. (2013). Multi-sample comparison of detrital age distributions. *Chemical Geology*, *341*, 140–146. https://doi.org/10.1016/j.chemgeo.2013.01.010

Vitter, J. S., & Webber, M. (2018). Water event categorization using sub-metered water and coincident electricity data. *Water*, *10*(6), 714. https://doi.org/10.3390/w10060714

Wei, L., Tian, W., Silva, E. A., Choudhary, R., Meng, Q., & Yang, S. (2015). Comparative study on machine learning for urban building energy analysis. *Procedia engineering (the 9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) joint with the 3rd International Conference on Building Energy and Environment (COBEE), 12-15 july 2015, Tianjin, China)*, (Vol. 121), 285–292. https://doi.org/10.1016/j.proeng.2015.08.1070

Willis, R. M., Stewart, R. A., Panuwatwanich, K., Jones, S., & Kyriakides, A. (2010). Alarming visual display monitors affecting shower end use water and energy conservation in Australian residential households. *Resources, Conservation and Recycling*, *54*(12), 1117–1127. https://doi.org/10.1016/j.resconrec.2010.03.004

Willis, R. M., Stewart, R. A., Panuwatwanich, K., Williams, P. R., & Hollingsworth, A. L. (2011). Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *Journal of Environmental Management*, *92*(8), 1996–2009. https://doi.org/10.1016/j.jenvman.2011.03.023

Wizard, T. (2003). Trace Wizard water use analysis tool. *Users Manual*.

Wonders, M., Ghassemlooy, Z., & Hossain, M. A. (2016). Training with synthesised data for disaggregated event classification at the water meter. *Expert Systems with Applications*, *43*, 15–22. https://doi.org/10.1016/j.eswa.2015.08.033

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, *32*.