Journal of Building Engineering 74 (2023) 106867

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Full length article

Identifying hybrid heating systems in the residential sector from smart meter data

Araavind Sridhar [a,b,*], Nadezda Belonogova [a], Samuli Honkapuro [a], Hannu Huuki [c,d], Maria Kopsakangas-Savolainen [c,d], Enni Ruokamo [c,d]

[a] LUT University, Lappeenranta, Finland
[b] Polytechnic University of Milan, Italy
[c] Finnish Environment Institute, Finland
[d] Department of Economics, Accounting and Finance, University of Oulu Business School, Finland

ARTICLE INFO

Keywords:
Data-driven
Flexibility
Machine learning
Smart meter data
Residential

ABSTRACT

In this paper, we identify hybrid heating systems on a single residential customer's premises using smart meter data. A comprehensive methodology is developed at a generic level for residential sector buildings to identify the type of primary and support heating systems. The methodology includes the use of unsupervised and supervised learning algorithms both separately and combined. It is applied to two datasets that vary in size, quality of data, and availability and reliability of background information. The datasets contain hourly electricity consumption profiles of residential customers together with the outdoor temperature. The validation metrics for the developed algorithms are elaborated to provide a probabilistic evaluation of the model. The results show that it is possible to identify the types of both primary and support heating systems in the form of probability of having electric- or non-electric type of heating. The results obtained help estimate the flexibility domain of the residential building sector and thereby generate a high value for the energy system as a whole.

1. Introduction

Decarbonization goals have caused revolutionary changes to the power system operation and planning. Within the EU, the initial 2030 targets of 40% renewable energy production have been revamped to 45% following the current trend in the EUcountries [1]. The increasing renewable production can mainly be attributed to the increasing intermittent renewable energy resources resulting from the advancements in technology [2], cost reduction [3], and subsidies [4] provided to utilize these resources. With such ambitious targets and the increasing intermittent energy production, the role of flexibility has never been more important than in the current times.

Renewable energy sources are penetrating all voltage levels of the grid from centralized to distributed solutions, thus changing generation from centralized and controllable to intermittent and weather-dependent profiles [5]. The introduction of bidirectional energy flows brings many challenges and requires proper management in the distribution network [6,7]. With the increasing intermittent renewable energy production, demand flexibility is vital for a smooth transition to the sustainable energy system of the future [8,9]. In addition to this, small-scale low-carbon technologies are gaining popularity among residential customers [10], thereby increasing the potential usage of demand flexibility.

[*] Corresponding author at: LUT University, Lappeenranta, Finland.
E-mail address: araavind.sridhar@lut.fi (A. Sridhar).

https://doi.org/10.1016/j.jobe.2023.106867
Received 24 March 2023; Received in revised form 15 May 2023; Accepted 16 May 2023
Available online 20 May 2023

Nomenclature

Abbreviations

AMR	Automatic Meter Reading
BOP	Bag of Patterns
BOSS	Bag of Symbolic Fourier Approximation Symbols
DR	Demand Response
DSO	Distribution System Operator
EU	European Union
EV	Electric Vehicle
GDPR	General Data Protection Regulation
PCA	Principal Component Analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
TSC	Time Series Classification
WEASEL	Word ExtrAction for Time Series Classification

Nomenclature

$\bar{y}_i$	Mean of all values used in R-square metric
$\hat{y}_i$	Predicted value from the regression at point i used in R-square metric
acc_j	Accuracy of the algorithm j
c_i	Coefficient of quadratic regression for $i = 0, 1, 2$
$L(T)$	Residential loads based on outdoor temperature
$P(x_i)$	Hybrid probability of consumer x being identified in class i
$P(x_{i,j})$	Probability of consumer x being identified in class i by algorithm j
R^2	R-square metric
T	Outdoor temperature
y_i	Actual values at point i used in R-square metric
a	Average intra-cluster distance
b	Average shortest distance to another cluster
FN	False Negatives
FP	False Positives
TP	True Negatives
TP	True Positives

The residential sector accounts for roughly 28% and 20% of the final energy consumption within the EU and Finland, respectively [11,12]. Individual residential households have low flexible loads, but when coupled together with multiple households, can raise the flexibility potential and help the energy system during peak hours [13]. Within the residential sector, the main form of flexibility can be achieved through large flexible loads, such as washing machines, dishwashers, dryers, EVs, and heating. Heating of residential houses is an ideal source of flexibility because of the opportunity for automated control within a comfortable temperature range [14]. In 2021, the ratio of annual electrical heating to total electrical consumption was 47.1% [12] in Finland, making the electrical heating an ideal form of DR flexibility. Electricity consumption patterns are changing with the electrification of the heating sector, contributing to an increase in the weather-dependent electricity load in the residential building stock within the EU [15]. This, in turn, represents a significant flexibility potential that remains untapped mainly owing to the uncertainty of flexibility and the lack of a regulatory framework for both consumers and energy stakeholders to provide and access flexibility resources, respectively. Furthermore, in recent years, the GDPR regulations and increasing concerns regarding data privacy among customers have led to consumers not disclosing private data to third party organizations. Thus, it would be essential to develop a model to identify the possible heating type based on smart meter data. The identification of heating type would directly facilitate identification of the theoretical flexibility potential that can be achieved in residential households. The extracted flexibility can increase the cost efficiency of the energy system, reducing imports and dependence from outside, assist in decarbonization, and provide a step forward towards a sustainable energy system.

Finland is among the leaders in the world in renewable energy. Finland has a very high proportion of renewable energy in the electricity mix and a very high prevalence of smart meters in residential households [16]. In addition to this, the current trends in the electrification of the heating sector in the EU and the option of having hybrid heating systems within residential households make Finland an ideal country to identify heating types based on loads.

1.1. Characteristics of finnish heating systems

In cold regions, residential households typically have a primary form of heating, which can be, e.g., district heating, direct electric heating, a heat pump, or oil heating. The primary heating mode is used, in particular, during cold months for space heating. In addition, households have an opportunity to increase their heating-related efficiency by installing an additional source of heating to complement their primary heating. This additional source of heating is called support or secondary heating. Households install secondary heating for varying reasons, such as to reduce the operating costs of heating and to increase the heating capacity to be equipped for extreme conditions or availability of alternative fuels [17,18]. Households that employ both primary and secondary types of heating can be classified as hybrid heating households. The heating types can also be categorized based on their source as electric (heat pumps, direct electric, and storage electric) and non-electric.

Finland is considered a cold country, located in the northern part of the EUand belonging to Dfc in the Köppen–Geiger climate classification [19], which indicates the presence of subarctic climatic conditions dominant in the major part of the country. In Finland, it is common to have more than one form of heating in a residential household [17]. Approximately 80% of the Finnish annual within-household energy use is related to heating space and water [12]. Recently, it has also been found that there is a trend of more residents installing secondary heating [20].

In addition to these points, with the increasing awareness regarding climate change and the high emissions from the fossil fuels heating sources, it is a possibility that many residential consumers would be switching to electric heating. The increased efficiency of heat pump based heating systems which provides additional cooling during the summers on top of the heating requirement would be vastly adopted in the coming years. Also, the possibility of having heating turned on automatically through the usage of electricity based heating solutions is an interesting option for residential consumers who prefer smart home automation in their households [21]. As a result, the usage of hybrid systems for heating and having at least one electricity based heating source is highly likely in the future of the residential heating sector. The identification of households employing electricity-based heating sources can be a useful step in assessing the flexibility potential of a household.

1.2. Literature review

There are numerous studies using different methodologies to identify patterns in residential energy. One of the most common ways is to develop different machine learning models, which can extract patterns from smart meter data. The usage of machine learning models for identifying patterns can broadly be divided into two types: supervised and unsupervised learning.

Unsupervised learning is a branch of machine learning in which the models do not have any labels, and the model is used to find groups of data that have similar patterns. Unsupervised learning has been used widely in the field of energy to identify different consumer subgroups, and this approach can also be called clustering. Wang et al. (2018) conducted a thorough review of data analytics on smart meter data [22]. Several applications including load forecasting, consumer clustering, load management, and other subcategories were explored in their study along with the respective techniques, including supervised and unsupervised learning. The existing literature in the field of residential energy consumption can be observed in Table 1.

From this table, it can be observed that there have been fewer studies focused on unsupervised learning based on heating sector than when compared to the electricity sector. Additionally, the results from the majority of the previous research highlights the significance of flexibility based on similar consumption patterns but have failed to consider the effect of heating consumption which plays a major role in residential household demand and is important extracting maximum flexibility through DR in residential sector.

Supervised learning is another branch of machine learning in which the models use known labeled datasets to train and extract significant patterns unique to the specific category. Supervised learning to identify consumer subgroups can also be called classification. Currently, several TSC techniques are used. One main drawback of the classification techniques when compared with clustering is the unavailability of training datasets to develop classification models [33]. Classification has predominantly been used to anomaly detection in smart meter data. Himeur et al. (2021) conducted a thorough review of the applications of different machine learning algorithms used for anomaly detection in residential buildings [34]. Convolutional neural networks were used by Opera et al. (2021) to detect anomaly within consumption data [35]. Zhang et al. (2019) used Gaussian Mixture Models to identify anomalies among consumers [36]. Classification has also been used to extract building features based on smart meter data. Carroll et al. (2018) used neural networks along with elastic net machine learning techniques to identify household demographic extraction based on smart meter data of Irish households [37]. Neale et al. (2022) used linear discriminant analysis to extract demographic features based on smart meter data [33]. Though classification is a well established and well researched methodology having numerous applications in different fields, it has not attracted significant attention in the field of energy to identify residential heating types.

Different methodologies have been used to identify specific load characteristics based on smart meter data. Wang et al. (2020) and Nankeolyar and Ray (2022) used a hidden Markov model approach to identify EVloads [38,39]. Weigert et al. (2020) employed several machine learning models to identify the presence of heat pumps based on smart meter data [40]. Similar studies have been performed by Hopf (2019) and Fei et al. (2013) [41,42]. Similar to the identification of heat pumps, smart meter data have also been used to identify air conditioning units as explored by Chen et al. (2019) [43]. In contrast, no significant research has been conducted related to the identification of heating systems in the residential sector.

Table 1
Existing literature on unsupervised learning in residential energy sector.

Source	Algorithm	Findings
[23]	Hierarchical clustering	This study identified specific time periods which are distinctive among consumer consumption patterns and converted the consumption data within these time periods into symbols. The results from this study highlight the effect of reducing data dimensionality through approximation technique and emphasizes the differences in the consumer consumption patterns.
[24]	K-means clustering	This study clustered 4000 residential consumers of Pacific Gas & Electricity company and identified 40 different consumer subgroups based on their consumption pattern.
[25]	Model based clustering	The study proposed two different approaches which included PCA to reduce the feature size and converting the time series data into small signature objects on which neural regression was applied. The results from this study identified 12 different consumer clusters having an average of 61% distinction between them.
[26]	K-means, Fuzzy k-means, Agglomerate hierarchical clustering	This study clustered the consumers into four subgroups and k-means clustering proved to be the best performing algorithm, which were then used for further analysis related to load prediction and building simulation.
[27]	Fuzzy c means clustering	This study identified five clusters, which were then used for short-term load prediction.
[28]	K-means, K-medoid, self-organizing map clustering	The results of this study were used to identify different profile classes to which a multinomial logistic regression was applied to attribute household characteristics to the profile.
[29]	K-means clustering	Clustering approach used to identify consumer clusters, which were then used to determine significant features affecting each cluster based on sociodemographic parameters of the consumers
[30]	K-means clustering	The results show that the daily raw profiles for a household varies from the average profile for that household emphasizes the fact that averaging datasets bypasses the diversity of electricity consumption among households.
[31]	K-means clustering	Clustering based on residential consumers district heating dataset. The results identified four different consumer categories and analyzed the effect of consumer building and occupant related characteristics.
[32]	K-means clustering	This study identified patterns among space heating of residential households. The study identified two main clusters with results emphasizing the effect of consumer behaviors affected the clusters.

1.2.1. Research gap

Based on the previous research and to the authors' best knowledge, there is little research on the identification of hybrid heating systems based on smart meter data in the residential sector. With existing literate mainly focusing on consumer subgroup identification based on overall consumption within unsupervised learning and anomaly detection within supervised learning, there is a research gap in using machine learning algorithms to identify the heating source based on electricity consumption. Such research would bring high value to the research domain as hybrid heating presents a significant flexibility potential for the power and energy system. Additionally, due to the scarcity of research in identification of heating systems, the use of both unsupervised and supervised learning can facilitate the understanding of the identification of heating problem and can facilitate the determination of the data value and requirements for the identification of hybrid heating systems. The correct identification of the hybrid heating systems would aid in extracting the theoretical flexibility potential among residential sector which is of significance as we move towards an energy system dominated by stochastic renewable energy sources.

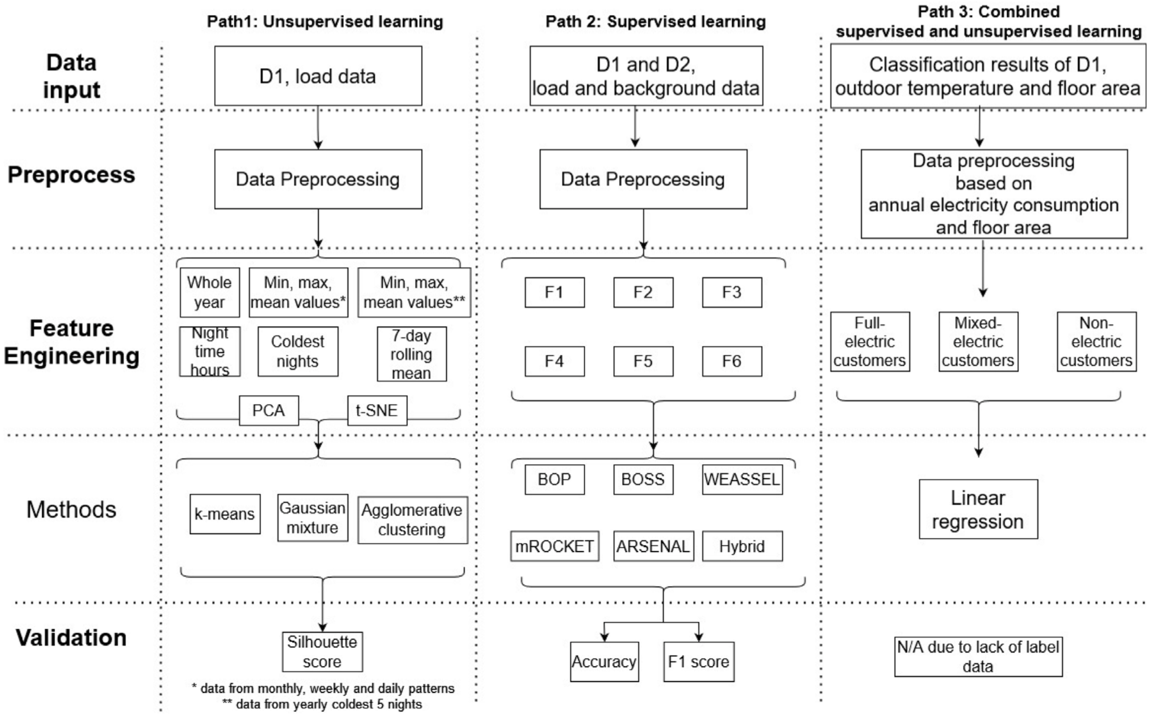


Fig. 1. Flowchart of the methodology.

1.2.2. Research objective and contributions

To address the above issues, the objective of this study is:

To analyze the residential consumers' consumption dataset to identify hybrid heating by using different machine learning approaches.

In order to bridge the gap in the literature and address the research objective, this study aims to contribute in the following way:

- Establish a methodology to detect hybrid heating, i.e., a system with both primary and supplementary heating on single residential premises, using smart meter and outdoor temperature data.
- Compare the performance of unsupervised and supervised machine learning algorithms to understand the value of consumer background information.
- Quantify the value of the background data by applying supervised machine learning algorithms to two datasets of different background data quality.
- Explore additional methods to tackle the complexity of the problem of identification of hybrid heating systems.

In order to address the above contributions, a methodology based on different algorithms was applied to two different datasets. The results of this study provide a step forward in identifying residential flexibility, which is essential for the sustainable future involving increased intermittent renewable generation.

The structure of the paper is as follows: Section 2 presents the methodology employed in this paper, containing unsupervised learning in Section 2.1, supervised learning in Section 2.2 and their combination in Section 2.3. The datasets are described in Section 3. The methodology is applied, and the results are presented in Section 4. A discussion on the key-takeaway points and a conclusion are presented in Sections 5 and 6 respectively.

2. Methodology

This study builds a methodology based on three paths. The first path contains the use of unsupervised learning, in which the consumer loads are used to identify consumer groups that can help identify the heating types. The second path is the use of supervised learning, in which, in addition to consumer loads, also background data are used, for instance consumer heating type and gross floor area. Finally, path three combines classification and regression from the two paths and uses outdoor temperatures to identify the hybrid heating system. A graphical representation of the methodology is shown in Fig. 1. Comparing the performance of the first two paths aids in deducing the impact and value of consumer background data in the identification of heating systems.

2.1. Path 1: Unsupervised learning

This path uses a dataset with no information on building or heating systems, i.e., only hourly consumption and outdoor temperature data are used in the analyses. The unsupervised learning includes feature engineering, selection of the most suitable clustering algorithm(s), and assessment of the performance of the model using silhouette score as a validation metric (Fig. 1).

2.1.1. Feature engineering

Feature engineering plays an essential role in the clustering process. The main objective of this part is to find the shortest feature vector that eliminates unnecessary noise, such as sauna peaks and other daily activities of inhabitants, at the same time preserving all the load pattern characteristics that are important from the perspective of distinguishing hybrid heating systems (e.g., peak powers caused by electric heating loads). Such a feature vector building process not only requires mathematical processing and transformation of the dataset but also background knowledge of various customer types as well as expected load patterns of various hybrid heating systems. The original time-series vector consisting of 8760 hourly values was converted into a feature vector by extracting only night-time hours, identifying minimum, maximum and mean values from daily, weekly and monthly consumption, extracting only the coldest nights of the year, and rolling the time-series data with various sizes of rolling window. After that, the obtained time-series vector was transformed using two-dimensionality reduction algorithms: PCA and (t-SNE) on top of the PCA values. As a result, numerous feature vectors (see Fig. 1) were constructed fed into the clustering algorithms.

2.1.2. Clustering

The clustering algorithms used were K-Means, Gaussian Mixture, and Agglomerative Clustering. All three algorithms require a predefined number of clusters as an input, that was estimated after applying dimension reduction to the time-series feature vectors using PCA and t-SNE methods and visualizing the results. However, there was a challenge of identifying separate cluster groups from the dataset after dimension reduction technique, since all samples were close to each other for all feature vectors. Therefore, several numbers of clusters were tried for all the three clustering algorithms and the Silhouette score was calculated for each number of clusters.

K-means algorithm belongs to centroid-based clustering algorithms that run iteratively to find the local optima based on the closeness of a data point to the cluster center. Agglomerative clustering is a data mining technique that groups data points based on their similarity, using a distance metric such as Euclidean distance. In contrast to K-means and Agglomerative clustering, Gaussian Mixture represents a soft clustering technique, where the results are generated with probability of each sample belonging to a certain cluster.

2.1.3. Regression

The objective of the regression analysis is to identify dependency of electricity consumption on the outdoor temperature reflecting various hybrid heating systems and use them as an input feature for supervised learning.

Only nighttime hours from 00:00 to 06:00 of the consumption data of each customer were used to avoid the noise caused by daytime activities of the customers, especially electric saunas. The total consumption over the six nighttime hours was calculated and regressed over the mean nighttime temperature. The regression coefficients, slope, and regression model fitting score were calculated for each month of the year for every customer. To illustrate the regression approach, slope results are shown for two different customers for December and the whole year in Fig. 2.

The monthly slopes and scores were calculated for each customer and analyzed with the objective of recognizing whether primary and/or support heating systems have outdoor temperature dependence or not.

2.1.4. Performance assessment

For clustering, the silhouette scores were calculated for each of the feature vector and clustering algorithms given in Fig. 1. The silhouette score is defined as the ratio of the difference between the average shortest distance to another cluster and the average intra-cluster distance to the maximum value among them [44]. This is also shown in Eq. (1).

$$silhouette\ score = \frac{b-a}{max(a,b)} \tag{1}$$

where:

a = average intra-cluster distance

b = average shortest distance to another cluster

2.2. Path 2: Supervised learning

Classification of individual residential customers is performed using the hourly load data together with the background information obtained partly from the Population Information System of Finland and partly from distribution system operators. The background data contain the information concerning the primary heating system used within the household and the gross floor area.

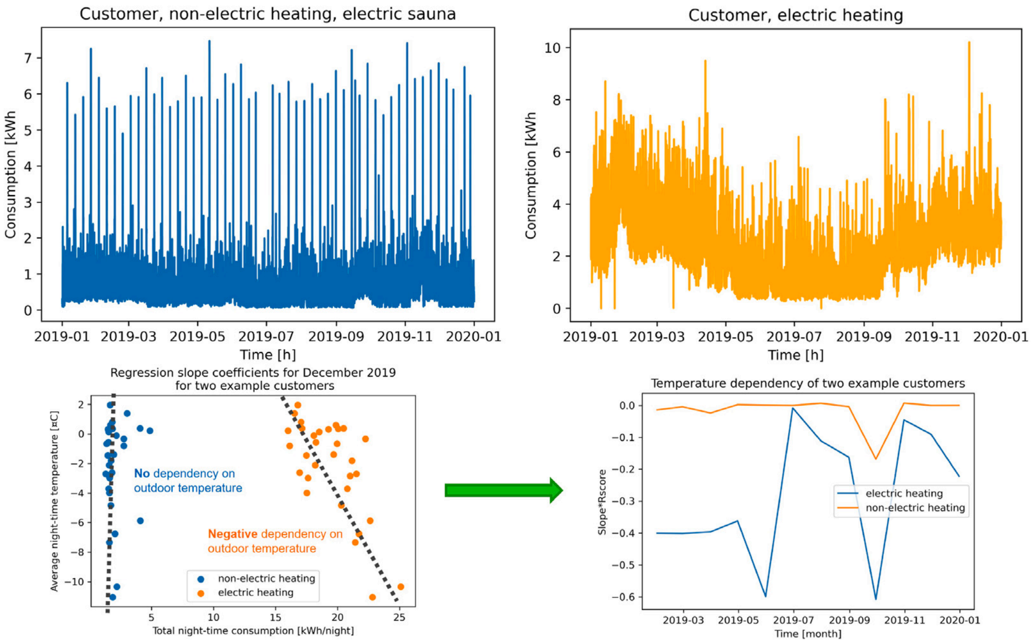


Fig. 2. Regression example results for two customers.

2.2.1. Algorithms

TSC is a specific domain of classification that handles time series data. TSC is extremely vital and has been studied extensively in recent years. One of the main issues regarding TSC is the extraction and transformation of time series data, which can then be grouped into different classes. Although classification algorithms have reached saturation in recent years, there have been significant developments in terms of different transformations that help in extracting the striking data among different datasets, which can then be classified later. The following algorithms were considered to be able to identify the consumers' heating type:

- BOP: This is a type of transformation algorithm that uses a sliding window to extract subsequences from the time series data and converts this subsequence to a word by using symbolic aggregate approximation and piecewise aggregate approximation algorithms [45].
- BOSS: This is a type of transformation algorithm that uses a structure-based similarity measure to classify different time series together. It also helps in reducing noise and is computationally less demanding, making it suitable for huge data analysis [46].
- WEASEL: This is a type of transformation algorithm that is similar to BOP. It converts the time series data into words and builds features representing the frequency of occurrences for each time series data [47].
- Mini-ROCKET: This is a type of transformation algorithm that uses random convolutional kernels. This algorithm computes two features from the resulting feature maps: the maximum and proportion of positive values. The transformed features are then used to train a classifier [48].
- Arsenal: This is a type of kernel-based classification algorithm similar to the ROCKET algorithm. It is an ensemble of the ROCKET transformer with Ridge Classification as the base classifier. The Arsenal algorithm weights each classifier using the accuracy from the ridge cross-validation, thereby allowing a generation of probability estimates at the expense of scalability compared with ROCKET [49].

The algorithms BOP, BOSS, and WEASEL are dictionary-based transformation algorithms, which are used to transform the input data into alphabets, and then, a classifier is used on top of these transformed data, whereas Mini-ROCKET and Arsenal are kernel-based algorithms. All the dictionary-based algorithms and Mini-ROCKET are made to use a Random forest classifier on top of the transformed data, while Arsenal has its own Ridge regression for its transformed data.

2.2.2. Performance assessment

There are four types of performance assessment metrics used in classification in this study. They are accuracy, precision, recall, and F1 score. All these metrics are achieved based on the obtained true positives (TP: an outcome where the model correctly predicts the positive class), true negatives (TN: an outcome where the model correctly predicts the negative class), false positives (FP: an outcome where the model incorrectly predicts the positive class), and false negatives (FN: an outcome where the model incorrectly predicts the negative class). The equations for the performance assessment metrics are the following [50]:

$$Accuracy = (TP + NP)/(TP + TN + FP + FN) \tag{2}$$

$$Recall = TP/(TP + FP) \tag{3}$$

$$Precision = TP/(TP + FN) \tag{4}$$

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5}$$

As it is difficult to compare algorithms having a low precision and a high recall, or vice versa, it is better to compare F1 scores between different algorithms. The F1 score is defined in Eq. (5). The weighted F1 score is obtained if weighted recall and weighted precision are used in Eq. (5).

Hybrid probability. As TSC is quite complicated, and no single algorithm is always better than the other when the accuracy and the F1 scores are almost equal, we incorporate hybrid probability to classify the consumers based on all the above-mentioned algorithms based on [49]. By doing so, the probability that a consumer is being assigned to a specific class by one algorithm is reduced, and the consumer is only placed in a specific class when the majority of all the other algorithms predict the same, while addressing the accuracy of each algorithm. The equation for the hybrid probability is written as

$$P(x_i) = \sum_{j=1}^{n} (acc_j)^n * P(x_{i,j}) \tag{6}$$

where $P(x_i)$ is the probability of a consumer x being identified in class i, acc_j is the accuracy of algorithm j, n is the total number of algorithms used in this methodology, and $P(x_{i,j})$ is the probability of a consumer x being assigned to class i by algorithm j.

2.3. Path 3: Classification and regression combined

The main objective of this section is to show how the two described paths can support each other and thus form a combined approach that outweighs in performance each of the individual paths. The path consists of two main phases, i.e., preprocessing of the classification results and a linear regression analysis.

The preprocessing of the classification results was carried out to take into account the false classes resulting from poor background information and hence a possibility of using wrong labels. For this, non-electric heating and electric heating customers were filtered out of the two classes using two parameters, annual energy consumption and floor area of the house. As a result, three subclasses were obtained: non-electric, full-electric, and mixed-electric customers (see Fig. 3).

After the preprocessing, linear regression was applied to all three groups to observe the dependence of electricity consumption on the outdoor temperature throughout the one-year period. The monthly regression coefficients help identify whether electric heating is used as a primary source throughout the year or as a support heating source during the coldest and/or special periods, like holidays or weekends. The procedure for the regression analysis follows the one described in Section 2.1.3.

2.3.1. Performance assessment

The classification results were assessed based on the metrics described in Section 2.2.2. For regression coefficients, the coefficient of determination of the prediction, or R-squared was calculated for each slope value obtained. R-square is defined as the ratio of variation of data points provided by the regression line and total variation of data points from the mean. This is also shown in Eq. (7) [51].

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

where:

y_i is the actual values at point i

$\hat{y}_i$ is the predicted value from the regression at point i

$\bar{y}_i$ is the mean of all values

R-squared contains values between 0 and 1, 0 representing a poor fit and 1 reflecting the best possible fit of the curve. For the analysis and illustration of regression results, the multiplication of the slope coefficient and R-squared value was used.

2.4. Methodological limitation

The limitations of the established methodology are formulated to understand the limitations of the results obtained:

- The results of a machine learning-based analysis are always influenced by the quality of the data.
- Feature engineering technique is data-specific and should be adapted to characteristic features of consumption data [52]. In this study, the features considered for different machine learning techniques are specifically adapted to fit the Finnish consumers. Thus, it would not be ideal to use the same feature selection for consumers living in completely different climatic conditions to compare the accuracy of the algorithms.
- We have limited our study to only a handful of algorithms to investigate the viability of the identification of hybrid heating. There could be other algorithms, e.g., deep learning, which might yield better results but can be considered a topic of future studies.

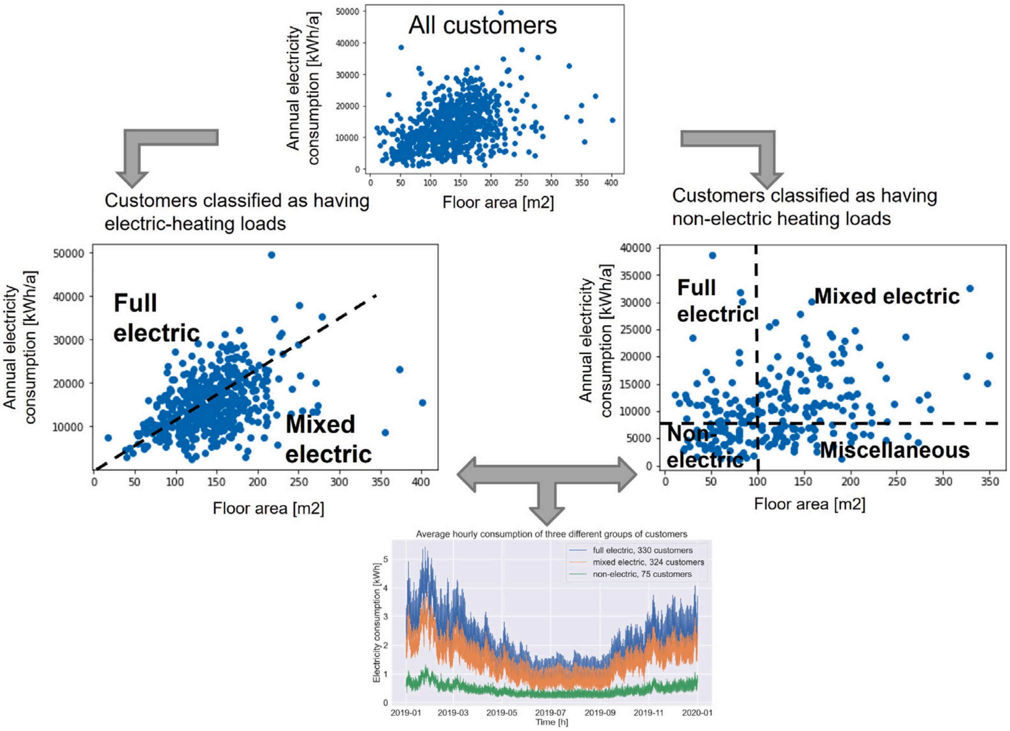


Fig. 3. Preprocessing of the classification results for the regression analysis.

3. Data description

The first dataset contains AMR data of residential customers collected from four different DSOs. The overall dataset contains hourly consumption values of approx. 2300 consumers in the years 2015, 2016, 2018, and 2019. In this paper, the AMR data for the year 2018 were used, as there were no significantly different results available when the AMR data from different years were analyzed.

The second dataset used in this study was obtained through an online survey from the customers of a Finnish DSO Caruna Ltd. It is the largest DSO in Finland, and it has residential customers in southern, western, eastern, and northern parts of the country. The online survey was initially distributed to 10,670 consumers in May 2018. A total of 1554 consumers responded to the survey focusing on different energy issues, and 1361 of the respondents gave permission to use their AMR data from years 2014 to 2017 for research purposes. In this paper, we use a subsample of 1361 consumers and their AMR data from 2017 in the analyses.

Additionally, customers from both datasets live in detached houses. The above-mentioned information together with the background data and quality of datasets is summarized in Table 2.

Both datasets contained building-specific background information. In addition to that, dataset 2 also included sociodemographic characteristics. The quality of the background data differed between the datasets. First, for dataset 1, the main factor degrading the quality was a risk for outdated information about the main heating system type in case when the end customers do not update their local DSO about changes in their heating system. Secondly, the building-specific data were purchased from the Population Information System of Finland and linked to the AMR data based on geographical coordinates. There was a risk of human errors during the coordinate linkage phase, for instance in the case when two buildings of the same size were located next to each other, e.g., one for living purposes and the other one for storage purposes. Thus, the gross floor area and heat source may have been identified wrongly in marginal cases. Thirdly, the main heating system may have been erroneously determined as a support heating system. For instance, the direct electric heating system may have been turned into a support heating after installation of a heat pump.

In contrast to this, the quality of background data on dataset 2 was much better, because it was obtained through a survey and directly from the customers. The information on the support heating system was missing from both datasets.

For the dataset 2, AMR data are of a good quality, including, e.g., explanations for missing hourly observations (if such exist) and whether the hourly observation is metered or estimated by the DSO. Moreover, the information gathered in the survey covers several sociodemographic characteristics for each consumer that can be further used to understand the value of such information in machine learning.

Table 2
Dataset description.

	Dataset 1	Dataset 2
Source of load data	Several DSOs operating in Finland mainly in rural areas	Caruna Ltd. – a DSO operating in Finland
Source of background data	Population Information System of Finland and DSO Customer Information System	Caruna Ltd. consumer responses to survey
Number of consumers	2300	1361
Data timeframe	2015, 2016, 2018, 2019	2014–2017
Background data collected	Year of construction, gross floor area, house type, presence of electric sauna, heat source	Year of construction, heated floor area, household size, house type, presence of electric sauna, primary heating type, and supplementary heating type
Quality of background data	Outdated DSO information; errors when linking coordinates to consumption points; lack of information on supplementary heating; unreliable information on primary heating	Good, background data obtained directly from the customers

3.1. Data preprocessing

In order to identify residential customers from the AMR data, three main preprocessing steps are implemented in this study:

(a) Annual loads: Based on the literature study, the annual residential consumption should be between 1000 kWh and 60,000 kWh [53]. Any consumers having an annual load outside these limits were removed from this study.
(b) Peak load: Based on the literature study, the peak power for a residential consumption at a given hour should not be more than 24 kWh (typically, the maximum fuse size for residential customers is 3 × 35 A in Finland). Any consumers having a peak power more than 24 kWh were removed from this study.
(c) Vacation homes: To remove the vacation homes (summer cottages), any consumers having zero consumption for more than 500 h in one year were removed from this study.

3.2. Feature selection

Feature selection is defined as the process of isolating the most consistent and relevant features to be used in training the model to increase accuracy and reduce computational burden [54]. As this study focuses on evaluating high-dimensional time series data for classification, different features were selected to identify the most accurate classification. The different features selected to be analyzed further are:

(a) F_1: The whole time series data are used without any specific features selected.
(b) F_2: The time series data are selected only for the nighttime hours (00:00 to 06:00). During this specific time frame, most of the residential loads other than heating loads are minimized, which can help in identifying a better classification.
(c) F_3: The whole time series data are divided by the gross floor area of the specific consumer. This feature is to analyze the effect of the household size on the electricity consumption loads.
(d) F_4: This is similar to F3, but it uses only the nighttime hours as of F2.
(e) F_5: This type of feature selection uses linear regression coefficients as features. The coefficients are obtained for the linear relationship between the outdoor average temperature and the average load consumption in the nighttime (see Eq. (8) [55]).
(f) F_6: This type of feature selection is similar to F5, but uses a nonlinear function to find the relationship within the average loads L and the average outdoor temperature T during the nighttime period using a quadratic Eq. (9) [55].

$$L(T) = c_0 + c_1 * T \tag{8}$$

$$L(T) = c_0 + c_1 * T + c_2 * T^2 \tag{9}$$

Table 3
Primary heating distribution within datasets.

Dataset	Heating share (%)	
	Electric	Non-Electric
D_1	66.1	33.9
D_2	71.6	28.4

Table 4
Silhouette scores from Path 1.

Number of clusters	Feature reduction	k-Means	Agglomerative clustering	Gaussian mixture
2 clusters	–	0.45	0.44	0.45
	PCA + t-SNE	0.58	0.58	0.57
3 clusters	–	0.38	0.34	0.38
	PCA + t-SNE	0.53	0.49	0.53
4 clusters	–	0.33	0.28	0.33
	PCA + t-SNE	0.51	0.5	0.5

3.3. Final input datasets

After data preprocessing, dataset 1 (D_1) resulted in 2281 residential consumers. Similarly, for dataset 2 (D_2), the data preprocessing resulted in 978 residential consumers. Based on the different datasets and different feature selection methodologies, a total of 12 different input datasets were obtained. All these datasets are denoted by $D_i F_j$, where D_i is dataset i, and F_j is feature method j used on the dataset. The primary heating distribution within the different datasets can be observed in Table 3.

4. Results

4.1. Unsupervised learning results: clustering

The best performance and the highest silhouette score were obtained for one-month consumption data with nighttime hours only. The results from unsupervised learning can be observed in Table 4.

From Table 4, it can be observed that the different clusters tested on different algorithms with and without feature reduction techniques. The dimension reduction techniques PCA and t-SNE significantly improved the score whereas the difference between different algorithms were not significant. The best silhouette score was attained for 2 clusters with feature reduction techniques used on k-means and agglomerative clustering techniques with a score of 0.58. As there have not been studies identifying the type of heating source based on consumption data, it is hard to compare the results of this study with the existing literature, but similar studies have been undertaken. Motlagh et al. (2019) had converted consumer consumption data into map-models to cluster them into similar consumer groups with a distinction ($silhouettescores/inertia$) of 61% among clusters [25]. Similarly, Czétány et al. (2021) used different unsupervised learning methods to cluster similar consumer groups based on consumption data with an average silhouette score of 0.2 [26]. Increasing the number of clusters led to a poorer performance for all the algorithms.

Another outcome of the clustering is that the division of the customers into fixed clusters turned out to be challenging because of the presence of hybrid heating. Thus, a customer may be allocated to one cluster based on the primary heating system and to another cluster based on the support heating system. To illustrate this, two example customers are selected, both having electric-based support heating but different primary heating systems (see Fig. 4).

Therefore, it was concluded that a traditional way of clustering the whole feature vector is inappropriate for the purpose of identification of hybrid heating systems.

4.2. Supervised learning results

The accuracy and F1 score of implementing all the algorithms on the different datasets are shown in Table 5.

The results show that in comparison with the different datasets, D_2 have a significantly higher accuracy than D_1 irrespective of the different features selected. In addition to this, in comparison with the different feature selection, it can be seen that F_6 has the best result for both the datasets. F_6 represents the nonlinear regression feature selections explained in Section 3.2. The best-performing results can be viewed as a confusion matrix as depicted in Figs. 5(a) and 5(b). Fig. 5(a) demonstrates that there are a significant number of consumers being classified as electric-type consumers, although they have stated that they have non-electric type of primary heating. Whereas in comparison with D_1, D_2 had a very high accuracy with far fewer consumers being wrongly classified. One of the main reasons for this is the quality of the dataset as explained in Section 3. Currently, there have been little research which uses different supervised learning approach to identify heating source based on consumer consumption data. Weigert et al. (2020) has used consumption, solar, weather, grid and survey data for their analysis to identify heat pumps using KNN, random forest, support vector mechanism and artificial neural network. The results from their study has an F1 score of 0.74 [40]. Additionally, a similar research has been performed by Fei et al. (2013) where they have used biased support vector machine to identify the presence of heat pumps. The results from their study has a similar F1 score of 0.86 [42]. Based on existing studies, it can be seen that the results from Path 2 has comparable results to the current literature and is on par with the accuracy.

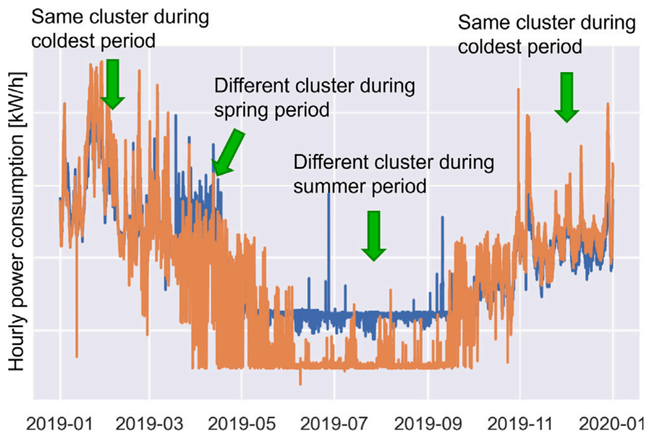
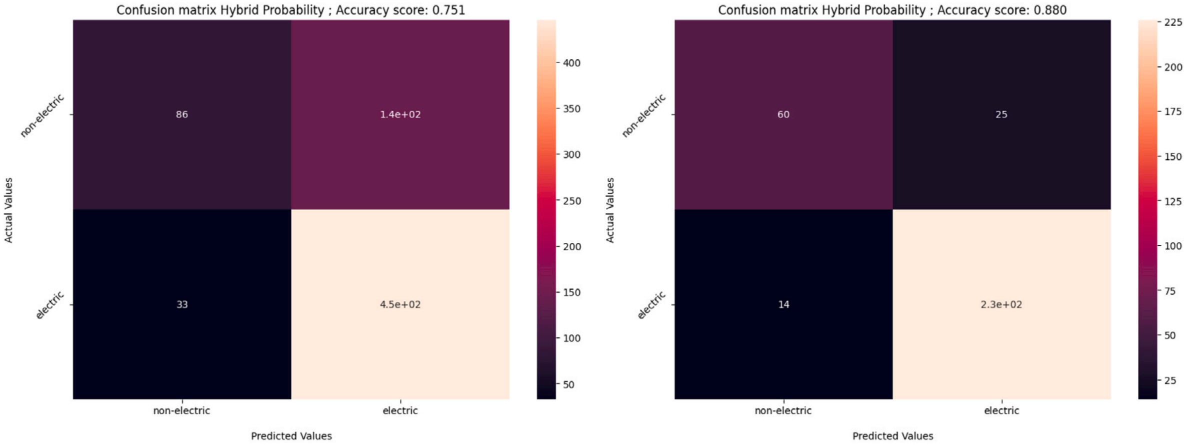


Fig. 4. Dynamic clusters of two example customers.



(a) Confusion matrix of D_1F_6

(b) Confusion matrix of D_2F_6

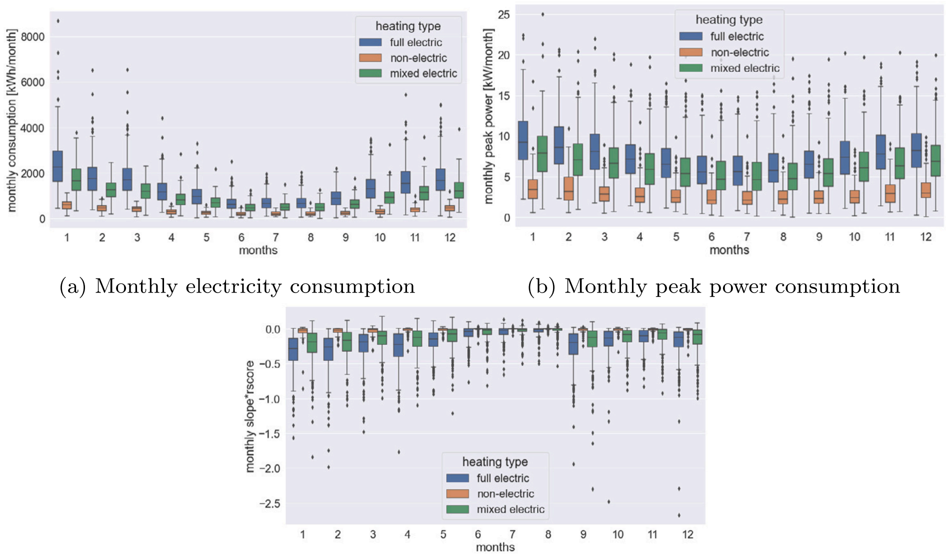
Fig. 5. Best results from supervised learning.

Table 5
Results of classification.

Input dataset	Accuracy						F1 score					
	BOP	BOSS	WEASEL	Mini-R	Arsenal	Hybrid	BOP	BOSS	WEASEL	Mini-R	Arsenal	Hybrid
D_1F_1	0.68	0.71	0.69	0.71	0.71	0.71	0.65	0.68	0.66	0.70	0.67	0.68
D_1F_2	0.69	0.69	0.69	0.73	0.69	0.71	0.65	0.66	0.66	0.72	0.65	0.68
D_1F_3	0.72	0.72	0.72	0.74	0.74	0.74	0.69	0.70	0.70	0.72	0.71	0.73
D_1F_4	0.68	0.69	0.70	0.70	0.73	0.72	0.64	0.65	0.67	0.69	0.69	0.68
D_1F_5	0.68	0.67	0.67	0.70	0.71	0.71	0.62	0.62	0.61	0.68	0.68	0.67
D_1F_6	0.69	0.76	0.73	0.74	0.75	0.75	0.66	0.74	0.70	0.72	0.72	0.72
D_2F_1	0.79	0.87	0.82	0.87	0.81	0.86	0.75	0.87	0.81	0.87	0.79	0.85
D_2F_2	0.82	0.88	0.85	0.88	0.82	0.88	0.79	0.88	0.84	0.88	0.79	0.87
D_2F_3	0.79	0.86	0.85	0.86	0.82	0.86	0.77	0.86	0.84	0.86	0.81	0.85
D_2F_4	0.76	0.85	0.82	0.87	0.79	0.86	0.70	0.84	0.79	0.87	0.76	0.84
D_2F_5	0.78	0.79	0.73	0.86	0.85	0.84	0.75	0.78	0.71	0.86	0.85	0.85
D_2F_6	0.80	0.88	0.85	0.88	0.88	0.88	0.80	0.88	0.85	0.88	0.88	0.88

4.3. Results of path 3: classification and regression

After classification, the two obtained classes were further split into three subclasses, i.e., non-electric, full-electric, and mixed-electric loads based on the annual electricity consumption and the gross floor area. The threshold level for the floor area was assumed



(a) Monthly electricity consumption

(b) Monthly peak power consumption

(c) Monthly distribution of regression coefficients

Fig. 6. Monthly consumption analysis.

to be 100 m^2 and the annual consumption 6000 kWh/a. The assumptions are based on the heuristic experience of electric-based loads typically exceeding 6000 kWh/a considering Finnish weather conditions and building characteristics.

The full-electric customers are expected to have both primary and support heating systems based on electricity, because they have a large annual consumption and a moderate floor area. The mixed-electric customers having a lower annual electricity consumption but the same level of monthly peak powers as the full-electric customers tend to have either primary or support heating to be based on electricity.

However, it is worth remembering that the outcomes are sensitive to the selected thresholds, and therefore, the results can be only in a probabilistic form.

To illustrate how the consumption changes throughout the year for each of the three subclasses, monthly electricity consumption and peak powers were drawn in the form of a box plot for every subclass (see Figs. 6(a) and 6(b), respectively).

It can be seen that the monthly electricity consumption is highest during the winter months for both full- and mixed-electric customers, although the mixed-electric customers have a lower monthly consumption than the full-electric customers because of the presence of non-electric-based heating. However, the consumption decreases in summer months (outside the heating period) for the mixed-electric customer group, which indicates that this group has electric-based heating only as either the primary or the support heating system. Instead, the monthly peak power levels stay at approx. the same level for both full- and mixed-electric customers also during the off-heating period, which makes it challenging to identify whether the primary or support heating is electric.

Therefore, regression analyses are required to observe the temperature dependence of the two groups throughout the year. The slope and the model fitting score were calculated for every customer and every month, and the results are illustrated in Fig. 6(c). The vertical axis represents the multiplied slope by the fitting score. The more negative the value is, the stronger is the dependence of the consumption on the outdoor temperature.

It can be seen from the figure that the dependence is minor outside the heating period, being close to zero for all the three subclasses. During the heating period, i.e., months 1,2,3,4 and 9,10,11,12, the dependence is stronger for the full-electric than for the mixed-electric customers. Again, this can be explained by the fact that mixed-electric customers may have non-electric heating loads as either the primary or support heating source. However, it is still challenging to say which of the two heating sources, primary or support, is an electric or non-electric one, because there is no clear change in the temperature dependence on the coldest (months 12, 1, 2) to the moderately cold months (3, 4, 9, 10, 11). The reason for no clear change is the mixture of customers with electric-based primary heating system (stronger dependence during the heating period) and customers with electric-based support heating systems (stronger dependence during the off-heating period) in the same group of mixed-electric customers.

In terms of flexibility potential, this uncertainty can lead to either over- or underestimation of flexibility, which can have implications on the energy system as a whole.

5. Discussion

The increasing electrification of the heating sector and the usage of multiple types of heating among residential households have led to a situation where identification of the heating types is essential to extract the flexibility potential within residential sectors. The usage of multiple heating sources in the Nordic countries hinders the approach to identify the theoretical flexibility potential which is gaining more influence in a sustainable energy system dominated by variable renewable energy sources. With raising concerns among residential users regarding their data usage and privacy, the need to identify the heating types based on the minimum data requirement is important. This study aimed to use multiple machine learning and data analysis techniques to identify the different heating types among Finnish residential households based on varying datasets having different data quality and quantity. The study started with the minimal consumption data without background information and started to incorporate more features of the datasets based on different methods. Based on the results of the study, there are five key takeaways:

(a) Among the different unsupervised learning algorithms, the best performing algorithm was the k-means clustering with feature reducing techniques which clustered consumers into two groups. The results of the unsupervised learning generated clusters that did not indicate any specific heating profiles among the consumers. Therefore, it is essential to highlight that it was not possible to identify any form of heating based on cluster analysis only, which highlights the need for consumer background information to perform any further analysis.

(b) The results of the supervised learning helped in identifying the consumer primary heating type with good accuracy which can be comparable to the similar studies using classification techniques to identify heat pumps [40,42]. Among the different algorithms and preprocessing techniques, the kernel based algorithms outperformed other types and the preprocessing with regression coefficients based on outdoor temperature and nighttime loads provided the best results. Overall, through the usage of classification for primary heating source identification an accuracy of 75%–88% was observed.

(c) Neither the unsupervised nor the supervised approach were able to identify the hybrid heating solutions. Within the unsupervised learning, the subgroups did not show any distinctive patterns to be clustered based on heating. The results from classification highlight the need for secondary heating solution labels which are a prerequisite for classification methodology and hence could not be used for identifying the hybrid energy systems. As a result, a hybrid approach is recommended to identify hybrid heating solutions among residential consumers.

(d) The quality of datasets significantly impacts the performance. As observed in this study, dataset D_2 was of a better quality and naturally yielded a higher accuracy than the results from dataset D_1. The collection of datasets plays a significant role in any form of analysis. With the datasets having varying quantity and quality, the machine learning algorithms are susceptible to over or under fitting. With the increasing privacy concerns regarding consumers sharing their personal data on their residential household, electricity suppliers and aggregators who are expanding their competence in flexibility sector need to provide additional assurances over data security to make consumers comfortable with their data shared. From this study, the effect of data quality on the results is evident and significant.

(e) With both unsupervised and supervised machine learning techniques failing to identify the hybrid heating solutions, this study proposed a hybrid approach. In this approach, the results from classification were used to then identify the secondary heating source. The regression analysis applied on top of the classification results showed that it is possible to identify groups of customers with fully electric hybrid heating systems and mixed hybrid heating systems, using the floor area and the outdoor temperature apart from the consumption data. However, owing to the lack of information on the support heating system and unreliable information on the primary heating system type, regression is still not sufficient to identify whether a primary or a supplementary heating system is based on electricity and dependent on the outdoor temperature.

Based on the results from the path 3, the identified heating source could potentially lead to the flexibility potential being either over- or underestimated for a group of customers. If the primary heating system is erroneously identified to be electric based and the supplementary heating system to be non-electric based, the flexibility potential will be overestimated. This is because in reality, electric-based heating is used only during the coldest periods of the year and not throughout the whole year as assumed. The consequences are a poor accuracy of the flexibility estimation, emphasizing the implications on both power state estimation and electricity market needs, which are of utmost importance with the increasing rate of renewable energy systems and electrification of the heating sector.

The practical implications of this study highlights the need for high quality of hourly consumption data and background data of residential consumers to be able to identify the type of heating source which is essential as we move towards a sustainable electricity system. The theoretical flexibility potential can be extracted from such identification of electric heating systems which can help the aggregators and retailers to leverage consumers heating up to a certain pre-agreed limit with the consumer to be changed during the time of need for the electricity system. In addition to this, such theoretical flexibility potential from households can provide an overall understanding of the current residential flexibility limits and how policies should be targeted to make consumers understand the need for flexibility and choose a flexible resource for the future.

Overall, this study takes a step forward towards residential flexibility through identification of heating systems in residential households and highlights the current complexity of this approach while providing additional suggestions for future studies.

6. Conclusion

The increasing electrification of heating sector and the rapid increase of renewable energy share in recent years had resulted in an additional need for flexibility in the energy system. The flexibility provided by individual houses is relatively low but when aggregated together can provide significant support to the energy system during peak hours. Therefore, the identification of the type of heating used by the residential consumers plays an important role to provide flexibility. In this paper, we have utilized unsupervised and supervised machine learning approach to identify the consumer heating source within residential houses. We

Table A.1
Monthly regression coefficient for 5 consumers based on linear regression.

Month	Consumer 1	Consumer 2	Consumer 3	Consumer 4	Consumer 5
31/01/2015	−0.58317	−0.77336	−1.54315	−4.10039	−1.35553
28/02/2015	−0.57133	−0.8544	−1.60412	−3.47005	−0.95166
31/03/2015	−0.25494	−0.55678	−0.38131	−4.24878	−0.65731
30/04/2015	−0.26943	−0.19153	−0.46044	−0.25743	−0.53131
31/05/2015	−0.78262	−0.70557	−0.76945	−2.78333	−0.62179
30/06/2015	−0.38738	0.251974	0.157553	−0.77265	−0.63618
31/07/2015	−0.98894	−0.19759	−0.57586	−0.21333	−0.81265
31/08/2015	−0.35159	0.172607	1.508849	0.911506	−0.54981
30/09/2015	−0.28815	−0.37602	0.253875	−0.00064	−0.93707
31/10/2015	−0.45381	−0.39094	−0.11376	−1.48801	−0.555
30/11/2015	−0.30114	−0.27699	−0.3595	−2.34802	−0.56225
31/12/2015	−0.2686	−0.54686	−1.449	−3.36953	−1.44502
31/01/2016	−0.52754	−1.05223	−1.61973	−0.7009	−1.21232
29/02/2016	−0.47034	−0.36173	−0.73394	−0.75701	−0.75566
31/03/2016	−0.5687	−0.8444	−0.68036	−0.54844	−1.37042
30/04/2016	−0.1559	−0.35808	−0.12292	−2.03687	−1.13022
31/05/2016	0.410812	−0.64839	−0.00861	−0.81234	−1.20455
30/06/2016	−2.47682	−0.78851	46.53562	5.157724	−1.69487
31/07/2016	−0.64981	0.630933	−1.23969	2.790874	−1.52691
31/08/2016	−1.07319	−0.08563	1.812381	4.940445	−0.66741
30/09/2016	−0.54774	−0.85944	−0.48442	−3.30523	−0.54813
31/10/2016	−0.21968	−0.46348	0.058413	−2.4629	−0.85947
30/11/2016	−0.52528	−0.55898	−0.755	−3.32751	−1.26349
31/12/2016	−0.50575	−0.61532	−0.24642	−2.46072	−0.47935

Table A.2
Monthly regression coefficient of quadratic term (c_2 term in Eq. (9)) for 5 consumers based on quadratic regression.

Month	Consumer 1	Consumer 2	Consumer 3	Consumer 4	Consumer 5
31/01/2015	−2.20588	−0.85052	−0.59679	−0.15095	−0.47671
28/02/2015	−0.44046	−0.10002	−0.18822	−0.1632	0.096011
31/03/2015	−1.2428	−0.71955	−0.07591	−0.10964	−0.22813
30/04/2015	−3.3738	1.41972	−1.11656	0.414892	−0.11977
31/05/2015	0.976299	−0.15336	−0.39642	−0.01284	−0.97698
30/06/2015	−0.16716	1.209836	2.076117	0.086338	−1.77961
31/07/2015	0.615493	−0.98055	−0.35374	0.092999	0.556039
31/08/2015	0.445346	0.397696	0.134524	−0.24745	−1.5364
30/09/2015	0.890058	−0.45213	0.124234	0.652484	0.065577
31/10/2015	−0.16645	−0.42705	0.177142	0.001461	0.273809
30/11/2015	−3.0781	−0.18109	0.628266	−0.29358	−0.38308
31/12/2015	−1.1712	−0.64044	−0.47639	−0.18355	−0.35871
31/01/2016	−2.45628	−0.78766	−0.73065	−0.37345	−0.10353
29/02/2016	−0.8707	−0.03352	0.165971	−0.15452	−0.48361
31/03/2016	−1.63412	−0.9805	−0.47845	−0.0985	−0.43459
30/04/2016	−2.59064	−0.73386	−0.18372	−0.1934	−0.61134
31/05/2016	−0.07718	−2.01263	0.930564	−0.21297	−0.68394
30/06/2016	−0.25777	−0.46695	0.000889	−0.1144	−0.66489
31/07/2016	0.076851	0.050006	−0.39256	0.071044	−0.78999
31/08/2016	0.327464	−0.15317	0.448289	0.515248	0.449519
30/09/2016	−0.13344	0.042894	0.434042	−0.00247	−0.945
31/10/2016	−1.64176	−1.63913	−0.65522	−0.21485	−0.15544
30/11/2016	−0.95714	−0.66235	−0.28256	−0.10303	−0.48573
31/12/2016	−1.51067	−1.03469	0.309183	−0.17219	−0.32685

have implemented different algorithms in different approaches and the methodology was tested on two different datasets with varying data quality. Within the unsupervised learning, the best clustering was obtained for two clusters with PCA and t-SNE dimension reduction techniques used to transform the input dataset. The results from the unsupervised learning failed to extract the heating source and an unsupervised machine learning approach to identifying heating based on consumption data did not yield any significant result. Following this, a supervised approach was performed within the two different datasets. The results from the supervised learning suggested that the quality of the dataset played a significant role in classification accuracy and the kernel-based algorithms outperformed the dictionary-based algorithms. Though the results from supervised learning were able to identify the primary source of heating with 75%–88% accuracy, due to the lack of labeled secondary heating source, classification could not be further utilized. In order to overcome this, a hybrid approach (Path 3) of combined supervised and unsupervised learning was proposed. This path utilizes the results from classification of primary heating source and then performs unsupervised clustering to identify the secondary source of heating. The results from this path were able to identify the consumer groups with fully electric hybrid heating systems and mixed electric hybrid heating systems using the floor area of the household and the outdoor temperature

Table A.3
Monthly regression coefficient of linear term (c_1 term in Eq. (9)) for 5 consumers based on quadratic regression.

Month	Consumer 1	Consumer 2	Consumer 3	Consumer 4	Consumer 5
31/01/2015	26.21102	16.24146	10.33679	3.282997	9.090652
28/02/2015	24.07055	16.29011	9.154742	3.249457	8.770713
31/03/2015	24.13184	15.94446	7.436576	2.959253	8.674551
30/04/2015	23.83155	11.31643	6.547384	2.19584	7.07209
31/05/2015	6.960188	7.118329	2.774852	2.420104	8.007174
30/06/2015	8.722195	−4.77696	−13.35696	0.79288	13.57936
31/07/2015	−0.19549	10.93489	3.808392	0.661919	−1.96692
31/08/2015	0.838589	−0.18975	−0.929896	3.651571	13.89854
30/09/2015	0.434274	7.14543	−0.555412	−1.52624	2.911974
31/10/2015	14.24354	12.04177	1.303008	2.806971	7.059272
30/11/2015	20.53429	13.74328	5.89371	3.938136	8.278529
31/12/2015	20.39816	15.00128	7.11752	3.284139	8.439871
31/01/2016	26.35198	16.21414	10.919943	5.178785	16.63002
29/02/2016	26.20037	16.33375	9.602956	3.873129	9.132754
31/03/2016	26.4214	16.4136	6.772355	3.757497	9.232852
30/04/2016	16.03512	13.01945	2.08302	3.446924	8.553543
31/05/2016	3.887138	18.20385	−4.450399	2.933228	8.094636
30/06/2016	5.302039	7.841989	0.269848	1.812785	9.944784
31/07/2016	1.817042	2.570353	4.835429	0.087037	10.03112
31/08/2016	1.465666	4.766377	−3.226439	−2.24014	−0.48137
30/09/2016	4.310097	5.509684	−0.859323	2.290138	8.599605
31/10/2016	15.26341	14.15773	1.607445	3.201598	8.86859
30/11/2016	24.95043	13.99034	5.318031	2.95217	9.787157
31/12/2016	27.6448	13.76223	7.980636	3.878595	14.042

as additional features in addition to the consumption data. Due to data unavailability on the secondary source of heating, and the unreliability of information regarding primary source of heating, this path fails to identify whether a primary or secondary heating system is based on electricity and outdoor temperature. Overall, this path highlights the complexity of the current hybrid systems and the issues with the datasets while providing suggestions for future studies.

CRediT authorship contribution statement

Araavind Sridhar: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Nadezda Belonogova:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Samuli Honkapuro:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing. **Hannu Huuki:** Methodology, Writing – review & editing. **Maria Kopsakangas-Savolainen:** Methodology, Writing – review & editing. **Enni Ruokamo:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This research has been funded by Business Finland in project “Highly Optimized Energy Systems HOPE”, funding decision 22693/31/2020. The authors would like to thank Hanna Niemelä for English language proofreading.

Appendix. Regression coefficients

See Tables A.1–A.3

References

[1] European Commission, REPowerEU: A plan to rapidly reduce dependence on Russian fossil fuels and fast forward the green transition, 2022.

[2] S. Sen, S. Ganguly, Opportunities, barriers and issues with renewable energy development – a discussion, Renew. Sustain. Energy Rev. 69 (2017) 1170–1181, http://dx.doi.org/10.1016/j.rser.2016.09.137, URL https://www.sciencedirect.com/science/article/pii/S1364032116306487.

[3] B.N. Stram, Key challenges to expanding renewable energy, Energy Policy 96 (2016) 728–734.

[4] A. Olabi, M.A. Abdelkareem, Renewable energy and climate change, Renew. Sustain. Energy Rev. 158 (2022) 112111, http://dx.doi.org/10.1016/j.rser.2022.112111, URL https://www.sciencedirect.com/science/article/pii/S1364032122000405.

[5] C.D. Iweh, S. Gyamfi, E. Tanyi, E. Effah-Donyina, Distributed generation and renewable energy integration into the grid: Prerequisites, push factors, practical options, issues and merits, Energies 14 (17) (2021) http://dx.doi.org/10.3390/en14175375, URL https://www.mdpi.com/1996-1073/14/17/5375.

[6] J. Salehi, A. Abdolahi, Optimal scheduling of active distribution networks with penetration of PHEV considering congestion and air pollution using DR program, Sustainable Cities Soc. 51 (2019) 101709.

[7] W. van Westering, H. Hellendoorn, Low voltage power grid congestion reduction using a community battery: Design principles, control and experimental validation, Int. J. Electr. Power Energy Syst. 114 (2020) 105349.

[8] P. Pinson, H. Madsen, et al., Benefits and challenges of electrical demand response: A critical review, Renew. Sustain. Energy Rev. 39 (2014) 686–699.

[9] S. Nolan, M. O'Malley, Challenges and barriers to demand response deployment and evaluation, Appl. Energy 152 (2015) 1–10.

[10] F. Geels, A. McMeekin, B. Pfluger, Socio-technical scenarios as a methodological tool to explore social and political feasibility in low-carbon transitions: Bridging computer models and the multi-level perspective in UK electricity generation (2010–2050), Technol. Forecast. Soc. Change 151 (2020) 119258, http://dx.doi.org/10.1016/j.techfore.2018.04.001, URL https://www.sciencedirect.com/science/article/pii/S0040162518305638.

[11] European Commission, Energy statistics - an overview, 2022, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview.

[12] Official Statistics of Finland (OSF), Energy consumption in households. URL http://www.stat.fi/til/asen/2020/asen_2020_2021-12-16_tie_001_en.html.

[13] A. Sridhar, S. Honkapuro, F. Ruiz, S. Annala, A. Wolff, Assessing the economic and environmental benefits of residential demand response: A finnish case study, in: 2022 18th International Conference on the European Energy Market, EEM, IEEE, 2022, pp. 1–6.

[14] M. Beccali, L. Bellia, F. Fragliasso, M. Bonomolo, G. Zizzo, G. Spada, Assessing the lighting systems flexibility for reducing and managing the power peaks in smart grids, Appl. Energy 268 (2020) 114924.

[15] G. Thomaßen, K. Kavvadias, J.P.J. Navarro, The decarbonisation of the EU heating sector through electrification: A parametric analysis, Energy Policy 148 (2021) 111929.

[16] Energy Authority, National report 2017 to the agency for the cooperation of energy regulators and to the European commission, 2017.

[17] E. Ruokamo, Household preferences of hybrid home heating systems – A choice experiment application, Energy Policy 95 (2016) 224–237, http://dx.doi.org/10.1016/j.enpol.2016.04.017, URL https://www.sciencedirect.com/science/article/pii/S0301421516301859.

[18] J. Räihä, E. Ruokamo, Determinants of supplementary heating system choices and adoption consideration in Finland, Energy Build. 251 (2021) 111366, http://dx.doi.org/10.1016/j.enbuild.2021.111366, URL https://www.sciencedirect.com/science/article/pii/S0378778821006502.

[19] M.C. Peel, B.L. Finlayson, T.A. McMahon, Updated world map of the köppen-geiger climate classification, Hydrol. Earth Syst. Sci. 11 (5) (2007) 1633–1644, http://dx.doi.org/10.5194/hess-11-1633-2007, URL https://hess.copernicus.org/articles/11/1633/2007/.

[20] J. Vihola, J. Heljo, Development of lämming methods 2000–2012. Data analysis, 2012.

[21] A. Sridhar, S. Honkapuro, F. Ruiz, J. Stoklasa, S. Annala, A. Wolff, A. Rautiainen, Residential consumer preferences to demand response: Analysis of different motivators to enroll in direct load control demand response, Energy Policy 173 (2023) 113420.

[22] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of smart meter data analytics: Applications, methodologies, and challenges, IEEE Trans. Smart Grid 10 (3) (2019) 3125–3148, http://dx.doi.org/10.1109/TSG.2018.2818167.

[23] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, S. Ghavidel, L. Li, J. Zhang, P. Siano, A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications, Energy Build. 203 (2019) 109455, http://dx.doi.org/10.1016/j.enbuild.2019.109455, URL https://www.sciencedirect.com/science/article/pii/S037877881931103X.

[24] J. Wong, R. Rajagopal, A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting, in: ACEEE Summer Study on Energy Efficiency in Buildings, 2012, pp. 374–386.

[25] O. Motlagh, A. Berry, L. O'Neil, Clustering of residential electricity customers using load time series, Appl. Energy 237 (2019) 11–24, http://dx.doi.org/10.1016/j.apenergy.2018.12.063, URL https://www.sciencedirect.com/science/article/pii/S0306261918318816.

[26] L. Czétány, V. Vámos, M. Horváth, Z. Szalay, A. Mota-Babiloni, Z. Deme-Bélafi, T. Csoknyai, Development of electricity consumption profiles of residential buildings based on smart meter data clustering, Energy Build. 252 (2021) 111376, http://dx.doi.org/10.1016/j.enbuild.2021.111376, URL https://www.sciencedirect.com/science/article/pii/S0378778821006605.

[27] X. Fu, X.-J. Zeng, P. Feng, X. Cai, Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China, Energy 165 (2018) 76–89, http://dx.doi.org/10.1016/j.energy.2018.09.156, URL https://www.sciencedirect.com/science/article/pii/S0360544218319285.

[28] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, Appl. Energy 141 (2015) 190–199.

[29] C.M.R. do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy Build. 125 (2016) 171–180, http://dx.doi.org/10.1016/j.enbuild.2016.04.079, URL https://www.sciencedirect.com/science/article/pii/S0378778816303565.

[30] S. Yilmaz, J. Chambers, M. Patel, Comparison of clustering approaches for domestic electricity load profile characterisation - implications for demand side management, Energy 180 (2019) 665–677, http://dx.doi.org/10.1016/j.energy.2019.05.124, URL https://www.sciencedirect.com/science/article/pii/S0360544219310060.

[31] P. Gianniou, X. Liu, A. Heller, P.S. Nielsen, C. Rode, Clustering-based analysis for residential district heating data, Energy Convers. Manage. 165 (2018) 840–850, http://dx.doi.org/10.1016/j.enconman.2018.03.015, URL https://www.sciencedirect.com/science/article/pii/S019689041830236X.

[32] C.M.R. do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy Build. 125 (2016) 171–180.

[33] A. Neale, M. l Kummert, M. Bernier, Discriminant analysis classification of residential electricity smart meter data, Energy Build. 258 (2022) 111823, http://dx.doi.org/10.1016/j.enbuild.2021.111823, URL https://www.sciencedirect.com/science/article/pii/S0378778821011075.

[34] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, A. Amira, Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives, Appl. Energy 287 (2021) 116601, http://dx.doi.org/10.1016/j.apenergy.2021.116601, URL https://www.sciencedirect.com/science/article/pii/S0306261921001409.

[35] S.-V. Oprea, A. Bâra, F.C. Puican, I.C. Radu, Anomaly detection with machine learning algorithms and big data in electricity consumption, Sustainability 13 (19) (2021) http://dx.doi.org/10.3390/su131910963, URL https://www.mdpi.com/2071-1050/13/19/10963.

[36] L. Zhang, L. Wan, Y. Xiao, S. Li, C. Zhu, Anomaly detection method of smart meters data based on GMM-LDA clustering feature learning and PSO support vector machine, in: 2019 IEEE Sustainable Power and Energy Conference, ISPEC, IEEE, 2019, pp. 2407–2412.

[37] P. Carroll, T. Murphy, M. Hanley, D. Dempsey, J. Dunne, Household classification using smart meter data., J. Off. Statist. (JOS) 34 (1) (2018).

[38] S. Wang, L. Du, J. Ye, D. Zhao, A deep generative model for non-intrusive identification of EV charging profiles, IEEE Trans. Smart Grid 11 (2020) 4916–4927.

[39] S. Nandkeolyar, P.K. Ray, Identifying households with electrical vehicle for demand response participation, Electr. Power Syst. Res. 208 (2022) 107909, http://dx.doi.org/10.1016/j.epsr.2022.107909, URL https://www.sciencedirect.com/science/article/pii/S0378779622001390.

[40] N.W. Andreas Weigert, T. Staake, Detection of heat pumps from smart meter and open data, Energy Inform. 3 (2020) http://dx.doi.org/10.1186/s42162-020-00124-6.

[41] K. Hopf, Predictive Analytics for Energy Efficiency and Energy Retailing, in: Contributions of the Faculty Information Systems and Applied Computer Sciences of the University of Bamberg, vol. 36, University of Bamberg Press, 2019, http://dx.doi.org/10.20378/irbo-54833.

[42] H. Fei, Y. Kim, S. Sahu, M. Naphade, S.K. Mamidipalli, J. Hutchinson, Heat pump detection from coarse grained smart meter data with positive and unlabeled learning, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1330–1338.

[43] M. Chen, K.T. Sanders, G.A. Ban-Weiss, A new method utilizing smart meter data for identifying the existence of air conditioning in residential homes, Environ. Res. Lett. 14 (9) (2019) 094004, http://dx.doi.org/10.1088/1748-9326/ab35a8.

[44] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.

[45] J. Large, A. Bagnall, S. Malinowski, R. Tavenard, From BOP to BOSS and beyond: time series classification with dictionary based classifiers, 2018, arXiv preprint arXiv:1809.06751.

[46] P. Schäfer, The BOSS is concerned with time series classification in the presence of noise, Data Min. Knowl. Discov. 29 (6) (2015) 1505–1530.

[47] P. Schäfer, U. Leser, Fast and accurate time series classification with weasel, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 637–646.

[48] A. Dempster, D.F. Schmidt, G.I. Webb, MiniRocket: A very fast (almost) deterministic transform for time series classification, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, New York, 2021, pp. 248–257.

[49] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, A. Bagnall, HIVE-COTE 2.0: a new meta ensemble for time series classification, Mach. Learn. 110 (11) (2021) 3211–3243.

[50] R. Caruana, A. Niculescu-Mizil, Data mining in metric space: an empirical analysis of supervised learning performance criteria, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 69–78.

[51] P. Vidyullatha, D.R. Rao, Machine learning techniques on multidimensional curve fitting data based on R-square and chi-square methods, Int. J. Electr. Comput. Eng. 6 (3) (2016) 974.

[52] G. Dong, H. Liu, Feature Engineering for Machine Learning and Data Analytics, CRC Press, 2018.

[53] J. Haakana, V. Tikka, J. Lassila, J. Tuunanen, J. Partanen, N. Belonogova, Power-based tariffs boosting customer-side energy storages, 2016.

[54] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.

[55] N. Fumo, M.R. Biswas, Regression analysis for prediction of residential energy consumption, Renew. Sustain. Energy Rev. 47 (2015) 332–343.