Article

Household Electricity Consumer Classification Using Novel Clustering Approach, Review, and Case Study

Gaikwad Sachin Ramnath , Harikrishnan R., S. M. Muyeen and Ketan Kotecha

Article

Household Electricity Consumer Classification Using Novel Clustering Approach, Review, and Case Study

Gaikwad Sachin Ramnath [1] , Harikrishnan R. [1,*] , S. M. Muyeen [2,*] and Ketan Kotecha [3]

1   Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed) University, Pune 412115, India; sachin.r.gaikwad@outlook.com
2   Department of Electrical Engineering, Qatar University, Doha 2713, Qatar
3   Symbiosis Centre for Applied AI (SCAAI), Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune 412115, India; director@sitpune.edu.in
*   Correspondence: dr.rhareish@gmail.com (H.R.); sm.muyeen@qu.edu.qa (S.M.M.)

Abstract: There is an increasing demand for electricity on a global level. Thus, the utility companies are looking for the effective implementation of demand response management (DRM). For this, utility companies should know the energy demand and optimal household consumer classification (OHCC) of the end users. In this regard, data mining (DM) techniques can give better insights and support. This work proposes a DM-technique-based novel methodology for OHCC in the Indian context. This work uses the household electricity consumption (HEC) of 225 houses from three districts of Maharashtra, India. The data sets used are namely questionnaire survey (QS), monthly energy consumption (MEC), and tariff orders. This work addresses the challenges for OHCC in energy meter data sets of the conventional grid and smart grid (SG). This work uses expert classification and clustering-based classification methods for OHCC. The expert classification method provides four new classes for OHCC. The clustering method is employed to develop eight different classification models. The two-stage clustering model, using K-means (KM) and the self-organizing map (SOM), is the best fit among the eight models. The result shows that the two-stage clustering of the SOM with the KM model provides 88% of overlap-free samples and 0.532 of the silhouette score (SS) mean compared to the expert classification method. This study can be beneficial to the electricity distribution companies for OHCC and can offer better services to consumers.

Keywords: data mining; machine learning; household electricity consumption; residential consumer classification

1. Introduction

The demand for electricity consumption has been recently increased due to various reasons such as an enhanced standard of living, a greater population size, the impact of urbanization, socio-economic growth, large-scale machinery, and electronic market trading [1,2]. However, the worldwide domestic energy consumption is the second largest in terms of the overall consumption share, increasing by about 25%. The overall Carbon dioxide emissions are also approximately rising by 17% as reported by the International Energy Agency (abbreviated as IEA is shown in Table 1) in 2016 [3]. On the other hand, the IEA stated in 2011 that worldwide, the domestic sector encompassed more significant energy-saving potential of about 0.48×10^6 Ktoe per year [4–7]. Furthermore, the study of the consumption analysis of the individual domestic consumer is ignored in developing countries. This is due to various factors affecting energy consumption, dynamic consumption behavior, and consumers, the outcome of which is having diverse consumption patterns throughout the year [8,9]. Some of the factors resulting in diverse consumption patterns are installing rooftop solar panels, large batteries, and smart home devices [10].

Table 1. Abbreviations used.

Symbols	Description	Symbols	Description
DM	Data Mining	DRM	Demand Response Management
QS	Questionnaire Survey	HEC	Household Electricity Consumption
KM	K-Means	MEC	Monthly Energy Consumption
H	Hierarchical	OHCC	Optimal Household Consumer Classification
SOM	Self-Organizing Map	LEDH	Less Energy Demand House
SG	Smart Grid	MEDH	Moderate Energy Demand House
ML	Machine Learning	PEDH	Peak Energy Demand House
FE	Feature Engineering	EPEDH	Extra Peak Energy Demand House
kWh	Kilowatt-hour	DRP	Demand Response Program
IEA	International Energy Agency	PCA	Principal Component Analysis
NA	Not applicable	SVM	Support Vector Machine
HID	Household Identity	C1	Cluster 1
SS	Silhouette Score	C2	Cluster 2
Avg_kWh	Average Energy Consumption in kWh	MSEDCL	Maharashtra State Electricity Distribution Company Limited

The specified diverse consumption patterns pose a challenge for the electricity-generating sectors to provide affordable, secure, green, clean, and reliable energy. These problems can be recognized through a household load proofing system and DM techniques such as clustering and classification [7,8,10–12]. Furthermore, the essential requirement of utility companies is a practical and straightforward approach for classifying and clustering consumers. This approach may consist of fewer features, fewer optimal data samples, and less time to calculate optimal groups based on consumption [11]. The household consumer classification essentially groups the same characteristics of households under one class. This class can be of further use for energy prediction, optimization and DRM. On the other hand, installing a smart meter may help utility companies to comprehend the varying consumption demand and provide better offers to consumers [13]. This approach may assist in the preparation of the consumer mindset in the process of the adoption of innovative technologies. This can be an initial stage for developing the SG [8,14,15]. The electricity company has limitations in analyzing the individual consumer's electricity demand using monthly electricity bill data. Hence, a simple and effective consumer classification approach is necessary. However, consumer classification is mandatory because one standard consumption pattern is not compatible with everyone. Along the same line, the author of [16] has challenged Swiss Energy Norms in one common way that includes all homes.

The main purpose of the article is to guide the research community and direct their attention to the most urgent and uncovered HEC of optimal household consumer classification in the field of conventional and smart grid research.

1.1. Motivation and Significance

There is a dire need to balance the electricity business and provide reliable and quality power supply services to consumers in the present competitive power sector market. However, in the present scenario of increasing energy demand, the main objective of the energy provider company is to develop effective DRM. For DRM, the utility company should know the exact requirement of the monthly household electricity demand of each consumer. Furthermore, a conventional grid of manual meter reading and a billing-based system has reported three main challenges in taking actual energy meter readings [17].

The first challenge is the technical error due to a faulty meter. The second challenge is the occurrence of errors in taking the meter-reading photo, as well as the faulty entry of readings into the system, and the third challenge is the inability to take the reading due to an untenanted house or restrictions on movement in a pandemic situation such as COVID-19. Due to these challenges, most of the time the utility company provides the average bills to consumers, whereas the average bill is mainly calculated on the basis of an

average energy consumption of the previous three consecutive months [17]. This method of average billing is ineffective, as for example, the consumption of electricity in the previous month in the case of the winter season and the average calculation for the months in the summer season would be incompatible. This problem has been greater than before due to the COVID-19 pandemic, when the average bill of the summer months (March 2020 to May 2020) was calculated based on the previous months of the winter season in India [18–20].

The 225 houses of the three districts of Pune, Nashik, and Ahmednagar in Maharashtra, which are based in India, noted that, from March 2019 to August 2020, around 19% of the electricity bills were average bills and 81% were normal bills. Although, out of the 225 houses, the majority of houses were from the district places. Thus, the findings of the case study concluded that, in a conventional grid system the monthly consumption data set does not always reflect the actual energy consumption pattern and energy demand. Due to this, only the MEC data set may not be adequately helpful to provide the optimal household consumer classification.

Furthermore, the installation of a smart meter can provide a partial solution to the problem of optimal consumer classification by the elimination of human errors [8,14,15]. Most of the existing schemes are not adequate to achieve load stability by analyzing the historical energy consumption data collected from smart meters. However, the smart-meter data set cannot provide load stability using only the energy consumption data sets [21]. However, a smart meter has the limitation of missing other data sets such as household characteristics, socio-demographics, appliance characteristics, feedback awareness, weather, occupant consumption behavior and so on [22]. According to the literature, the effective consumer classification and consumption analysis can be performed by considering these factors [21].

The motivation of this study is to obtain the optimal household consumers classifications in both the conventional grid and the SG system. Moreover, the proposed study can be used to address the challenges of energy consumption, prediction accuracy and energy optimization. However, the optimal consumer classification may prove to be useful for utility companies to make informed decisions on DRM. In addition to this, it is possible to target specific household consumers in order to attract consumers and provide better offers and services to them [13].

1.2. Novelty and Contribution

The majority of the previous studies applied the clustering-based classification method and used the smart-meter data sets [7,8,10,23,24]. This study has addressed the limitations of optimal consumer classification in both conventional energy meter data sets and smart-meter data sets. The proposed methodology for consumer classification is different than the previous studies.

As per the literature on the household electricity consumption in the Indian context, this is the first study which will predict the individual household monthly electricity consumption for energy optimization for different locations, using the survey study and various data sets [25]. The optimal consumer classification is the first stage for the accurate prediction of energy consumption and energy optimization. Furthermore, for optimal consumer classification, this study proposes a novel methodology using classification and clustering methods. In addition, this study develops eight clustering models and finally proposes an efficient, hybrid, two-stage model using KM and SOM clustering models. The proposed model is also used for the result validation of the expert-based classification of consumers. This study focuses on the optimal classification of household consumers using different data sets such as QS, MEC, and consumption slab data from tariff orders of the top seven utility companies of India. The proposed study is unique due to the following contributions:

1.　A novel methodology for optimal household consumer classifications is proposed.
2.　An expert classification is performed and new consumer classes are formed, and a two-stage indirect clustering model for optimal consumer classification is proposed.
3.　The energy consumption data set and the QS data set are analyzed to find the history and pattern of energy consumption of the consumer.

4. The challenges, implications, and future directions in the household electricity consumption (HEC) study are addressed.

The paper consists of five sections. Section 2 presents the related works on the HEC study using DM techniques for OHCC. Section 3 is the materials and methods for OHCC that shed light on the proposal of a novel methodology. Section 4 consists of the results and discussions that show the household consumer classifications. Finally, Section 5 provides the concluding remarks with future directions.

2. Related Work

The DM techniques and neural-network-based approaches were mainly used to acquire knowledge from data sets. This knowledge is required to make informed decisions in diverse applications [8,12,26]. According to the literature, the power sector has mainly applied DM techniques to applications such as load classification, tariff structure reform, load management, anomaly data detection, consumer classification, outlier detections, and so on [8,10,12,14,16,21,24,26–28]. The DM technique includes two main methods: classification using supervised learning and clustering using unsupervised methods. The classification approach helps form the optimal tariff structure, which is necessary to balance consumer service satisfaction, reduce distribution losses, and increase company profits. On the other hand, an effective tariff structure design needs electrical behavior in different period data sets. At the same time, the proper selection of variables is crucial to improve the classification performance [7,8,10,12,26]. Furthermore, consumer classification is more challenging in smart-meter-based data [10]. In the case of smart-meter-data-based studies, the other significant data types are missing such as socio-demographics, household characteristics, appliances, and weather. This study addresses the mentioned gaps of smart-meter data by including various data types through a structured QS technique. The different data types with their significance are discussed in Section 3. There are many classification algorithms reported in the literature using different data types. The famous classification algorithms are the decision tree, logistic regression, and SVM [7,12,26]. Moreover, the papers [10,26,27] have worked on consumer classification using load demand data.

The author of [10] has shown the load-profile-based clustering using canonical variate analysis (CVA), linear discriminate analysis, and the locality sensitive hashing method classifier for the detection of abnormal energy consumption [29–31]. The result of the classification technique can be assessed through different parameters. The basic parameters used for classification result evaluation are accuracy, precision, recall, and the F1 score. Figure 1 shows the types of clustering methods: partitioning, hierarchical, density-based, grid-based, and model-based. All of the clustering methods have their specialties and potentials, meaning that none of the clustering methods is always superior. So, the correct selection of the clustering method is crucial and mainly dependent on the type of application, the specialty of the clustering methods, and the kinds of data sets. Most of the time, clustering methods are selected based on their popularity, simplicity of operation, and performance [8,10,27].
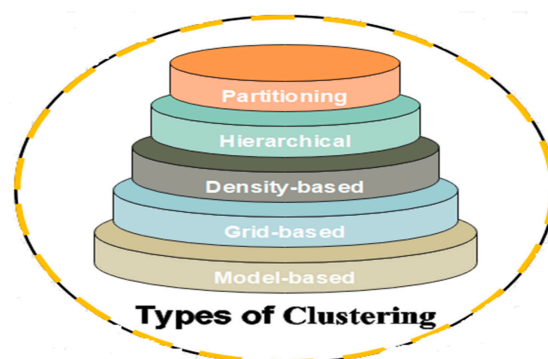

Figure 1. Types of clustering methods [29].

Furthermore, the commonly applied clustering algorithms, namely KM, H and SOM [7,8,21,26,32], SVM, faint, honey-bee-mating optimization, ant-colony-optimization algorithm, follow-the-leader, iterative refinement clustering, and ISODATA [27]. The papers [7,10] have discussed types of clustering techniques: direct or traditional clustering and indirect or advanced clustering. The direct clustering method uses the collected data directly, which performs clustering without preprocessing techniques such as correlations and feature extractions. The conventional clustering technique is considered a benchmark for further clustering development [7]. Furthermore, the authors of [7,8,10,27,29] have discussed the direct clustering methods using KM, H and SOM with smart-meter data sets. Table 2 summarizes the comparative study of the literature on HEC using DM techniques.

Table 2. HEC-based comparative study of previous work done using DM techniques.

Study	Objectives of the Study	Algorithms/Models Used	Type of Data Used	DM Techniques/Approaches Used	Applications
[7]	• DM techniques for load profiling in terms of clustering techniques and evaluation criteria for clustering results with applications to DRP were reviewed. • To highlight indirect clustering techniques for load-profiling purposes.	• Clustering- and classification-based DM techniques were reviewed.	• Review paper • NA	• Clustering techniques like direct and indirect, then clustering evaluation criteria • Clustering-based load profiling and then classification is applied.	• Clustering-based load profiling for classification in DRP applications using selected features. • For load forecasting, optimizing the electricity bill.
[8]	• New research on clustering methods and applications in power sector systems for domestic consumers was reviewed.	• Cluster combination using SOM and KM was discussed.	• Review paper • NA	• Clustering	• Helpful for system operators in peak load reduction and promoting DRPs.
[9]	• To determine an optimal number of representative load profiles using clustering for each season.	• KM clustering • Regression approach for post-clustering analysis.	• 103 homes smart-meter data of duration November 2012 to October 2013	• Clustering-based load profiling • A regression method for post-clustering using correlation techniques was applied.	• Electricity bill optimization
[11]	• To compare the performance of clustering algorithms • To apply data size-reduction techniques	• Modified follow-the-leader, hierarchical clustering, KM, fuzzy KM, and SOM	• 234 non-residential consumers of medium volts	• Clustering-based classification • The PCA and CCA-based data reduction techniques were applied.	• To find the optimal number of clusters. • To design the tariff structure of a distribution company
[12]	• The basic theory of classifications reviewed • Consumer classification performance was assessed using actual data of distribution companies.	• Modified follow-the-leader algorithm and the SOM	• 234 consumer data from the Romanian national electricity distribution company, Electrica, has been collected.	• Clustering-based classification	• Applicable to industrial, services, and small-business activity types. • They applied to tariff design with rates.
[16]	• A rigorous assessment of clustering-based classification, which identified the electricity demand profiles, was proposed.	• KM	• One-year smart-meter data of 656 consumers • Socio-demographics and household characteristics • households, Switzerland.	• Five feature-based clustering • Electricity demand-based classifications	• The identified electricity demand profiles help policy makers.

Table 2. *Cont.*

Study	Objectives of the Study	Algorithms/Models Used	Type of Data Used	DM Techniques/Approaches Used	Applications
[24]	• To propose a new and robust framework using DM techniques for finding the relevant knowledge on how and when consumers use electricity	• K-Means, SOM, and decision trees with the ruleset	• Real and historical database of 165 low-voltage consumers from the Portuguese distribution company.	• Load-profile-based clustering, then classification	• Classification tools can be helpful to distribution and retail companies.
[26]	• A review of electricity load-profile-classification methods	• KM, hierarchical, fuzzy KM, follow-the-leader	• Review paper • NA	• Clustering and classification	• Tariff design at electricity companies
[27]	• Application of clustering methods for load classification were briefly reviewed, and performance was assessed using evaluation methods.	• Load classification using fuzzy C-Means (FCM). • Reviewed four commonly used clustering methods: KM, FCM, hierarchical, and SOM	• Review paper • NA	• Clustering-based classification • A five-stage process model of load classification was constructed.	• Application for inadequate data identification, correction, load forecast, designing tariff, DSM, marketing strategies, value-added services, etc.
[29]	• To establish groups of consumers using electrical load pattern data. • To fine-tune the hyper-parameter of clustering algorithms and check its impact on clustering validity indicators.	• Discussed clustering methods HS2, HS5, KM, FKM, FDL, and HA2	• 400 consumers of non-residential Medium Volts on a representative weekday of the intermediate season.	• Electric load-pattern-data-based clustering • Post-clustering technique for finding the hidden potential of clustering results and removing outliers	• Operator for electrical load pattern grouping and representative load pattern analysis
[30]	• To propose a method to utilize a more significant number of data points to establish baselines that are not limited to similar temperature days.	• This paper presents a cohort-based baseline method that utilizes event-day metered load. Simple KM, ensemble KM, and decision tree	• Real data from the control group during the event period • Demographics, weather, day of the week, and exogenous factors data.	• Clustering-based classification • Feature-selection techniques can be applied to simplify the tree.	• Estimated baselines more accurately than existing methods for residential and industrial load applications.

Furthermore, the authors of [8–10,12,26,33] reviewed the latest research work in clustering techniques for the domestic load profiling of consumers. In addition, the authors of [10,34] also suggested that the performance of the clustering model mainly depends on the selection of an optimal number of clusters and the available persons in the home. Moreover, the clustering-based classification approach is also practical for reducing the prediction error. On the same line, Seasonal-Nave and Holt-Winters algorithm results have shown that the prediction accuracy increases with the number of clusters. Furthermore, the number of clusters needed before or after depends on the selected clustering algorithm and its application. Sometimes prior knowledge of the optimal clusters is essential. Though the number of clusters can be known through different tests, all tests give slightly varied results. One test is the R-function-based NbClust for cluster-determination purposes [9].

However, the author of [24] covered the methods of the selection of an optimal number of classes. The formula for defining the optimal number of classes is 2 to $\sqrt{M}$, where M denotes the number of responses. Moreover, the paper [26] used concisely different clustering-based classification algorithms. The new consumer class formation was completed using a hierarchical algorithm with better performance. In addition, the authors of [8,16] achieved a consumption pattern study by using a clustering-based classification approach and averaging the domestic load profiles. In addition, the sensor-based or time-interval-meter-based setup option was referred for classification. However, the sensor-based option for consumer classification is costly and time-consuming. Due to this, the

author of [35] agreed to use the load-profile-based clustering and classification method. This method is less expensive for an individual's energy demand analysis [35]. Moreover, the Portuguese distribution company proposed the clustering-based classification of 165 low-voltage residential data and other consumer category types' real and historical consumption data sets [8,24]. The authors of [10,36] covered a shape-based clustering approach that works on individual consumer time-based detailed load data.

This clustering approach is similar to KM, but the distance is quantified using the DTW method. Furthermore, most clustering techniques use Euclidean distance metrics for all dimensions. These metrics are unable to reflect the actual shape of the load curve. This issue has been addressed in [7,37] using three primary subspace clustering load-profiling techniques: cell, density, and clustering with fine-tuning of the hyperparameters. Furthermore, the authors of [10,38] focused on the k-shape-based clustering approach. This approach can be used for building time-series data for prediction purposes. In addition, the high-dimensional, non-linear correlation among the consumptions of different periods was reported in [10,39]. In addition, the load curves were first normalized, and a mixture of adaptive KM and hierarchical clustering was performed. Quantity and variability were considered for the use of other suitable measures [8]. The author of [10] discussed a mixture model of clustering techniques. This technique considered the expectation–maximization (EM) method to obtain the variabilities in residential load profiles. In addition, the EM method was used as the transition matrix approach on a second-order Markov chain and Markov decision processes. In addition, in [16] the author proposed five feature-based DM techniques to create compelling representative energy demand profiles. The average-consumption approach can compromise the variations of an individual consumer [8,16].

In addition, the direct clustering technique has two fundamental limitations. The first limitation is that it cannot perform well on highly fine-grained smart-meter data ranging from 1 min to 2 h. For this issue, the results in [10] suggest that the smart-meter data of a minimum of 30 min work better than data sets with other granularities. The second limitation is poor performance on smart-meter data, as well as dynamic, time-dependent, or time-series data [10]. However, indirect clustering is further explored based on the multi-stages of clustering. According to the literature, major parts of works have applied up to two stages of clustering. This is also known as the two-fold approach. The first step performs preprocessing on the collected data, the same as the direct clustering technique. The second step is to apply various advanced data-quality-improvement techniques such as dimension reduction or data size reduction, correlation methods, and feature extraction [7,10]. Along the same line, the author of [30] discussed two-stage DM techniques that were used first for clustering-based classification and then consumption prediction purposes. Further, the author applied the KM algorithm for the load profiling of the non-event days' load and decision trees to predict energy consumption levels using socio-demographics and appliance-based data. Similarly, the authors of [26,27,32] attempted the development of a two-level methodology for the classification of electricity consumers. The author of [35] proposed a two-stage methodology to classify electricity consumers based on electricity consumption patterns, load curves, and load values.

Furthermore, the papers [10,40] applied a two-level clustering approach to reduce the computational complexity. In this approach, a load-profiling analysis was performed using the KM method. The second level further carried out the clustering process based on the cluster centers acquired in the first level. The author of [8] proposed the ensemble clustering technique with the SOM-based first level for dimension reduction, and then the KM algorithm was used at the second level to form the clusters as the output. The paper [41] captured one of the advanced indirect clustering techniques with a deep embedded clustering approach. This clustering technique is helpful for both feature-extraction and clustering purposes. In addition, the correlation method is highly beneficial when the data sets include more variables and need selected variables [42]. On the same line, the author of [42] briefly discussed the concept of correlations. In addition, correlation-based feature selection can help to prevent over-fitting and low-accuracy classification issues [43]. The

authors of [27,29] used correlation methods for load-profiling applications. In addition, the author of [9] studied the regression-analysis-based correlation using seasonal data in terms of the hourly electricity consumption of 103 consumers in Austin, TX, USA.

Furthermore, another method for indirect clustering techniques is a feature-engineering-based feature-extraction approach, mainly used to reduce the dimensions of the input data. Moreover, indirect clustering is again dimension-reduction-based, and time-series-based clustering is classified based on the feature extraction. The feature-extraction techniques have been explained in detailed in [7], whereas [16] proposed five feature-based works wherein the improved clustering results are compared to the collection of all features. The author of [10] studied how the feature-engineering technique of feature extraction improves the indirect clustering performance [10]. In addition, the author of [8] applied feature-extraction techniques using different methods such as discrete Fourier transform (DFT), harmonics-based coefficients, and discrete wavelet transform. The data-reduction step improved the performance of clustering algorithms [9]. Power systems can apply clustering-based techniques to reduce data dimensions and find customers with similar patterns [8,10]. This can be possible using methods such as PCA, Sammon mapping and symbolic aggregate approximation, curvilinear component analysis (CCA), and CVA [8,29].

After obtaining the clustering results, there is a need to assess the quality of the created clusters. The clustering technique creates the load-profiling-based energy consumption pattern, which has great potential to be explored further through the post-clustering technique [13,29]. In addition, there are different ways to evaluate clusters by measuring the similarity within the clusters and the differences among them [8]. Moreover, the clusters can be assessed in two main ways: the compactness of the data points within the cluster and the separation of the clusters between clusters. On the same line, the author of [7] discussed two methods of cluster evaluation: similarity-orientated and classification-orientated. Another aspect of post-clustering is to determine the significant features of each cluster. After the clustering technique, a correlation method can be applied to identify the essential features. Moreover, the clustering validity indicators represent the positive correlation between clustering performance and the number of outliers available [29]. In [7–10,24,44], the commonly used clustering validity indicators are discussed.

This paper has focused on the widely used clustering algorithms such as KM, H, and SOM. These algorithms have been applied to direct clustering and 1st- and 2nd-stage indirect clustering methods. The detailed discussion is incorporated under Sections 3 and 4 of this paper.

The household energy consumption is dynamic in nature and difficult to understand. This is mainly due to the fact that it depends on multiple indoor and outdoor factors [2]. Additionally, as mentioned, this is the first study in the Indian context that tried the simple, effective, and popular clustering algorithms, namely KM, hierarchical and SOM for the consumer classifications, and eight clustering models were developed.

3. Materials and Methods for Optimal Consumer Classification

This study proposes a novel methodology for OHCC using a comparison between the simple expert classification approach and its result using different clustering approaches, namely direct, indirect and multi-stage. This work requires different data sets, namely: QS, MEC, and tariff orders of seven states of utility companies in Maharashtra, India. Furthermore, before launching a survey, a pilot survey was carried out and suggestions from respondents and experts were included. For the consumer classifications, this study used the average MEC (Avg_kWh) as the target variable. Thus, this study applied a simple, average-based statistical method to 225 households for classification.

3.1. Data Collection and Data Preparation

A total of 225 houses' QS and MEC data were collected. This case study collected data from three districts, namely Pune, Nashik, and Ahmednagar of Maharashtra, India, as

shown in Figure 2. The random sampling data-collection method was applied for the QS data set. The QS data set was structured and had detailed information on the electricity consumption of individual houses. The QS consists of six parts, namely: basic information, electricity bill information, house characteristics, socio-demographic factors, appliance characteristics, feedback, and awareness, as shown in Figure 3. The major factors and the list of variables of the questionnaire are also shown in Table 3. Additionally, there are various data-collection methods that can be found in [3]. The QS data sets were collected using different modes, namely Google form, e-mails, phone calls, social media, and on-site field visits.

Moreover, the specialties of the designed questioner were structured to include sample images, examples, and inbuilt thresholds for the entry-side data validation, and also uses English and the local language. The expert validation and pilot study for the improvement of the quality and quantity of the survey study was performed. All of these best practices have helped to increase the response rate of QS data collection.
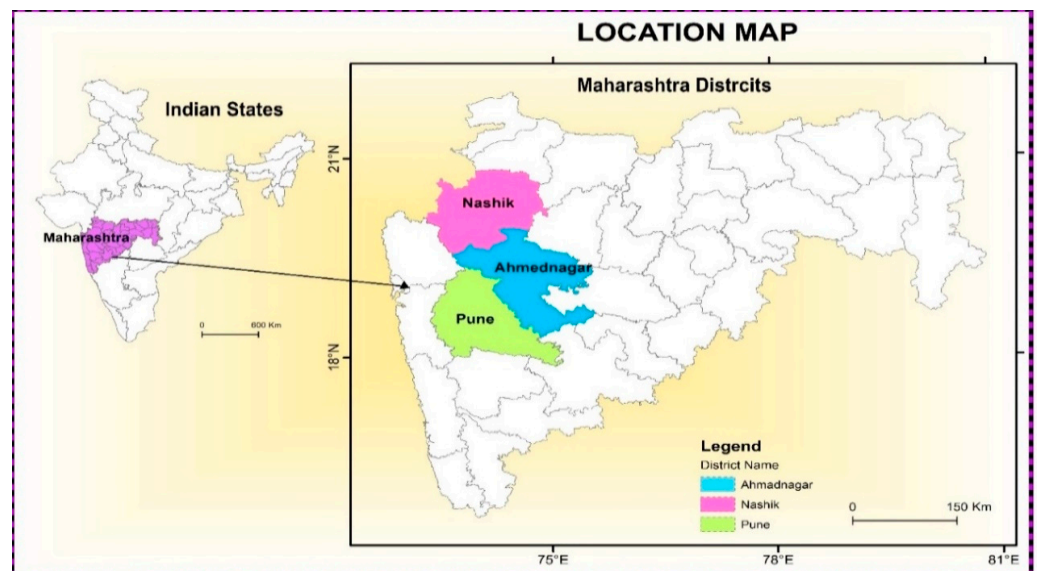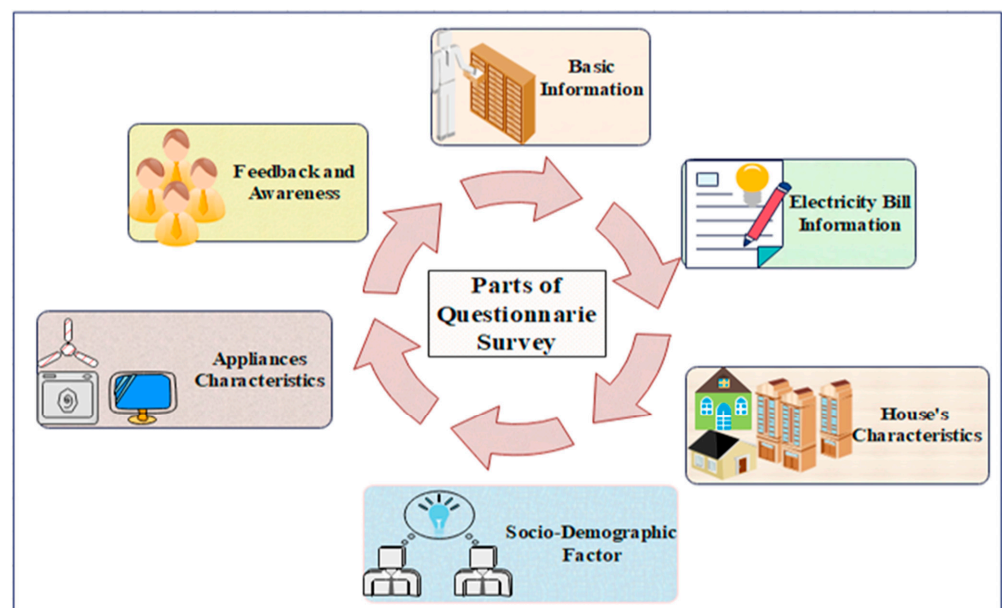

Figure 2. Field locations for sample survey.

Figure 3. Parts of a questionnaire.

Table 3. List of QS factors and variables.

Sr. No.	Factors	Total Variables	List of Variables
1	Socio-demographic	4	Districts, Members, Monthly Family Income, Education
2	Household characteristics	8	HomeOwnership, CarpetArea, Rooms, Windows, Balconies, HomeLocation, DoorDirection, VentilationSun-lighting
3	Regular Appliances	15	TV, Refrigerator, TableFansNo, CelingFansNo, Mobiles, WaterHC, Mixer, WaterPurifier, LEDBulbNo, CFLBulbNo, T-shapedLampNo, FluorTubeNo, LEDTubeNo, OtheBulbsNo, Iron
4	Lifestyle Appliances	30	AC, Actemp, AircoolerNo, ExhaustFansNo, WashingMa, MobilePB, Laptop, Geyser, InductionCooktop, MWOven, LEDBulbNo, LEDBulbChrNo, LED0bulbNo, ZeroBulbNo, OtheBulbsNo, InteriorLighting, Inverter, Motor, Desktop, Dongle/Internet USB2, WifiRouter, HomeAS, ElectroGames, EV, ELCB, ToasterNo, SteamerNo, HairDreier, AlexaD, HomeSS
5	Other Factors	08	OutageType, VoltageFluctuation, VentilationSun-lighting, MajorAppliancesOff, TreeShade, WaterBodies, HID, 5YearOldNo

Note: WaterHC = water heating coil; LED = light-emitting diode; CLF = compact fluorescent lamps; AC = air conditioner; WashingMa = washing machine; MobilePB = mobile power bank; MWOven = microwave oven; LEDBulbChrNo = no. of LED charging bulbs; HomeAS = home audio system; ElectroGames = electronic games; EV = electric vehicle; ELCB = earth leakage circuit breaker; AlexaD = Alexa device; HomeSS = home security system; HID = home identification; 5YearOldNo = five-year old appliances.

In addition, the individual consumer's MEC data set was collected by electricity bills from consumers and MSEDCL from March 2019 to August 2020, whereas the electricity tariff orders data set of different utility companies was collected from the utility company website. Besides, the Table 4 shows the poorly associated variables. Furthermore, the top seven utility companies of different states of India were considered, namely: Maharashtra, Gujrat, Punjab, Haryana, Karnataka, Andhra-Pradesh, and Assam [17,45–49] as shown in Table 5. These tariff orders were comparatively studied and the new classes for consumer classification were proposed.

Table 4. The poorly associated variables with the target variable.

Sr. No.	Variable Name	S-Value	*p*-Value	Sr. No.	Variable Name	S-Value	*p*-Value
1	District	−0.199	−0.214	13	HomeSS	−0.011	−0.071
2	HomeLocation	−0.188	−0.174	14	AlexaD	−0.011	−0.071
3	VentilationSun-lighting	−0.148	−0.209	15	Iron	0.002	−0.055
4	MajorAppliancesOff	−0.139	−0.201	16	WaterBodies	−0.091	−0.098
5	OutageType	−0.099	−0.126	17	Geyser*	−0.073	0.01
6	DoorDirection	−0.088	−0.011	18	Education	−0.04	0.042
7	HomeOwnership	−0.014	−0.098	19	Balconies	−0.029	0.057
8	TableFansNo	−0.084	−0.026	20	CeilingFan*	−0.025	−0.019
9	OtheBulbsNo	−0.064	−0.056	21	FloreTube*	−0.017	0.012
10	AC*	−0.059	0.055	22	Refrigerator*	−0.009	0.052
11	LEDLamp*	−0.026	0.038	23	VoltageFluctuation	−0.005	0.052
12	WashingMac*	−0.017	0.084				

Note: S-value = Spearman value; *p*-value = Pearson value; AC* = air conditioner star rating; LEDLamp* = light-emitting diode star rating; Geyser* = geyser star rating; CeilingFan* = ceiling fan star rating; FloreTube* = florescent tube star rating; Refrigerator* = refrigerator star rating; WashingMac* = washing machine star rating; AlexaD = Alexa device; HomeSS = home security system.

The data preparation is an essential step before applying the data sets to the data driven model. This step is required to improve the overall performance of the model. On the other hand, there are the significant challenges while collecting primary QS data, such as the presence of missing values or anomaly data points, unexpected happenings such as the COVID-19 pandemic situation, which limits an on-site study, not being able to collect the data, the inability to take an energy meter reading due to a locked home, technical

problems such as a faulty energy meter, and a lack of communication. Thus, QS data is the primary data type, which needs to be converted into the quality data type.

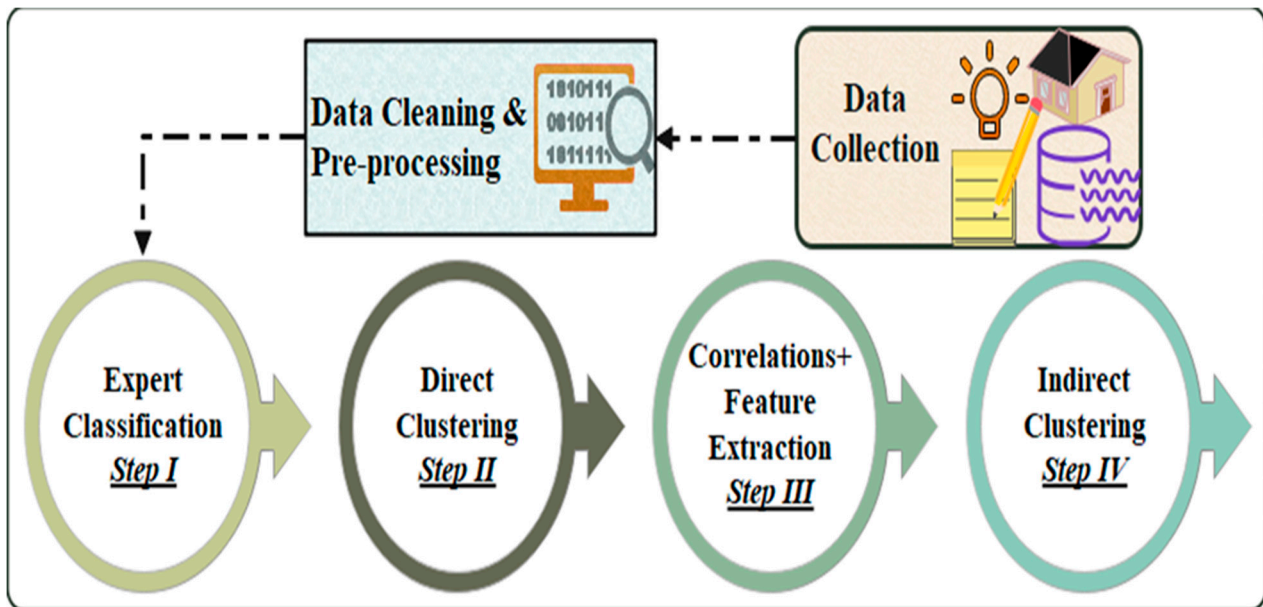**Table 5.** Consumption slab-based consumer classification using tariff orders.

Sr. No.	State Regulatory Commission	Consumer Category	Consumption Slabs (kWh)	
1	Maharashtra Electricity Regulatory Commission (MERC), Maharashtra	LT I (A)—Residential—Below Poverty Line (BPL)	0–30	
		LT I (B)—Residential (Non-BPL)	0–100 (+100) 101–300 (+200)	301–500 (+200) 501 and above
2	Gujrat Electricity Regulatory Commission (GERC), Gujrat	Urban & Rural—Residential Group (A): BPL consumers	0–50	
		Urban & Rural—Residential Group (B): Other than BPL consumers	Existing slabs 0–50 (+50) 50–100 (+50) 100–250 (+150) 250 and above	Proposed slabs 0–50 (+50) 50–200 (+150) 200–350 (+150) 350 and above
3	Punjab Electricity Regulatory Commission (PERC) Punjab	Domestic consumer category using consumption slabs	0–100 (+100) 101–300 (+200)	Above 300
4	Haryana Electricity Regulatory Commission (HERC) Haryana	Domestic supply, Category I: Total consumption up to 100 units per month	0–50 (+50) 51–100 (+50)	
		Domestic supply, Category II: Total consumption more than 100 units per month and up to 800 units per month	0–150 (+150) 151–250 (+100) 251–500 (+250)	501–800 801 and above
5	Karnataka Electricity Regulatory Commission (KERC) Karnataka	Tariff Schedule, LT-1: Under Bhagya Jyoti and Kutira Jyoti Schemes	0–40 units	
		LT-2(a) (i-Urban) (ii-Rural): Applicable to Areas under Village Panchayats	0–50 (+50) 51–100 (+50)	101–200 (+100) Above 200 units
6	Andra-Pradesh Electricity Regulatory Commission (APERC) Andra-Pradesh	Domestic-LT-I (3 Groups: A, B & C)	Group A: 0–75 (Telescopic) 0–50 & 51–75 Group B: 76–225 0–50 51–100 101–200 & 201–225	Group C: Above 225 0–50 & 51–100 101–200 201–300 301–400 401–500 & + 500
7	Assam Electricity Regulatory Commission (AERC) Assam	LT Category-I: (Below 0.5 kW) Jeevan Dhara	0–30	
		LT Category-II: (0.5–5 kW) & LT Category-III: (5–25 kW): Domestic A & B	0–120 121–240 Above 240	

For this, the study applied different techniques such as data preprocessing, data cleaning, data validation, data reduction, FE, correlation methods, identification and removal of outliers, and the formation of appropriate assumptions, as shown in Figure 4. Other clustering and classification approaches can be used for data preparation [29].

3.2. Proposed Methodology for OHCC

The household electricity consumption study in the Indian context is the first study that will predict the individual household monthly electricity consumption for the energy optimization of different locations using a survey study. For the accurate prediction of energy consumption and energy optimization, optimal consumer classification is the first stage. Furthermore, for optimal consumer classification, this study proposed a novel methodology using classification and clustering methods as shown in Figure 4. In addition, the study developed eight clustering models and finally proposed an efficient, hybrid,

two-stage model using KM and SOM clustering models. The proposed model was also used for the result validation of the expert-based classification of consumers. The proposed methodology was implemented in four steps: The first step was expert classification, the second step was direct clustering, the third step was data preprocessing by correlation and feature extraction, and the fourth step was indirect clustering in multiple stages.

Figure 4. Proposed methodology for optimal consumer classification.

The first step was based on the expert classification method by using two data sets, namely the history of electricity consumption data set and the energy consumption slab data set from different tariff orders of India, as shown in Figure 5. The tariff order provided energy consumption slabs, which can be considered as one kind of a baseline household consumer classification, as shown in Table 5. The features used for the expert classification were HID, eighteen months of average energy consumption (Avg_kWh) and consumer class threshold using the consumption slabs of the reviewed tariff orders. The Avg_kWh was a new feature generated from the secondary data set, tariff orders are available on websites of utility companies, and the HID feature was taken from QS data set. Thus, the overall data quality of the expert consumer classification method is greater than the QS-based clustering model of consumer classification.

Thus, the specialty of expert classification is its simple, generalized approach, which includes a smaller number of features, optimal data samples, and less time to calculate optimal groups, making it a practical and straightforward approach. The majority of the requirements of the utility company for consumer classification are the same as mentioned above [11]. Thus, the expert classification method may give better insights and support to utility companies for optimal consumer classification.

The second step and fourth step were based on clustering methods by using various data sets, namely the history of monthly consumption data, the consumption slab data set from different tariff orders of India, and the QS data set, as shown in Figure 5. The QS was structured in detail in order to determine the individual household consumption and to understand the significant factors. The QS consisted of six parts, namely basic information, electricity bill, house characteristics, socio-demographic factors, appliance characteristics, and feedback and awareness, as shown in Figure 3 and Table 3. Furthermore, this study explored the clustering method based on the type of data and the combination of models. In continuation, for achieving the optimal consumer classification, we used three clustering approaches, namely: direct, indirect, and multi-stage clustering approaches.
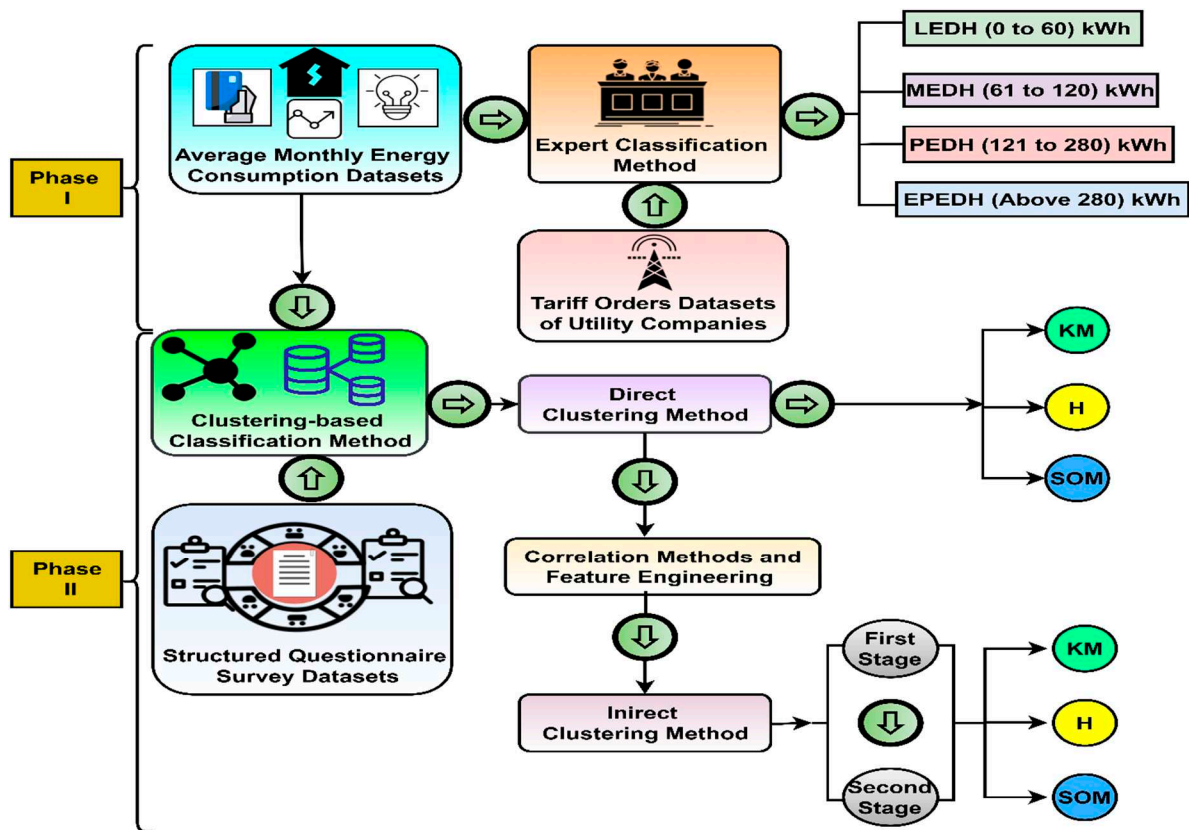
Figure 5. The input–output-based structure of the proposed work.

The direct clustering model provides the original data set as the input data in order to model without much preprocessing such as data optimization, FE, and so on. In the case of the indirect clustering model, the original data set is provided as the input data after improving the data quality through preprocessing steps. In the case of multi-stage modeling, there is a combination of the individual indirect clustering model to form a hybrid model. Furthermore, this is the proposed methodology of consumer classification, which is the first study in the Indian context. The open-source Orange tool was used to develop the clustering models. The authors of [19,50,51] and the official website https://orangedatamining.com (accessed on 27 March 2022) give the details about the Orange tool.

The required clustering algorithms were selected based on their simplicity, better performance, and different applications. The selected clustering algorithms that were used for optimal consumer classification were KM, H [13], and SOM. However, no one algorithm or technique is superior. All of them have their potential and their performance may vary based on the applied type of data, fine-tuning the parameters of the algorithms, and the application. Thus, the superiority of any DM technique and algorithm can be tested by considering the same data sets and rule sets.

The Orange tool has the option of automatic selection of the optimal number of clusters and their SS in the cluster modeling. Among all the direct and 1st-stage indirect clustering models, the KM model had better performance, as shown in Figures 6 and 7, respectively. Furthermore, Figure 7 shows that the optimal cluster number is two, and KM++ initialization was selected for the 1st-stage KM clustering model development. At the same time, the SOM algorithm result could not directly provide the number of clusters. Thus, after the SOM clustering operation, there was more of a need to obtain the number of clusters for the post-clustering step. Figure 8 shows the workflow diagram with the hyperparameters of the hybrid, 2nd-stage SOM with the KM clustering model. After comparing all the direct and indirect clustering models, the proposed model was the best

fitted, that have been discussed under section Results and Discussions. The development of clustering models is discussed in Sections 3 and 4.
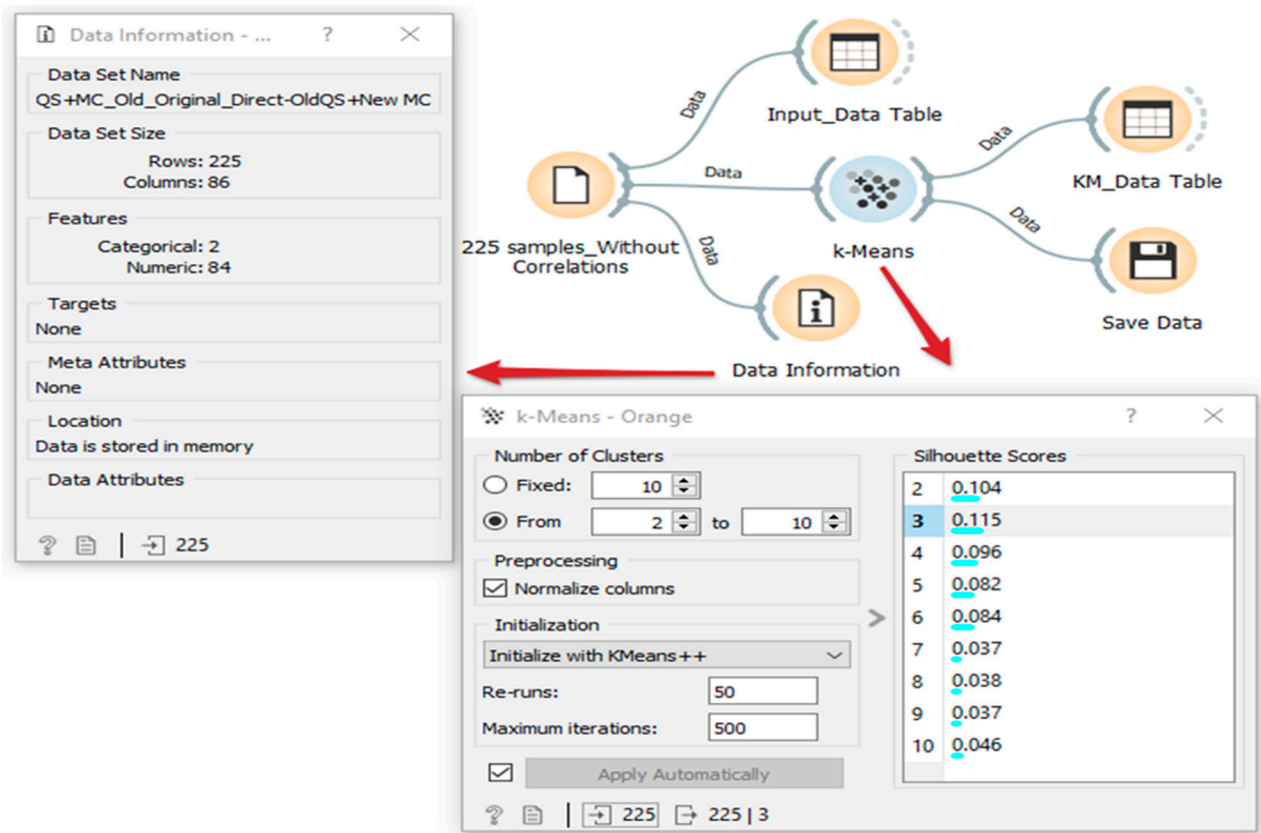

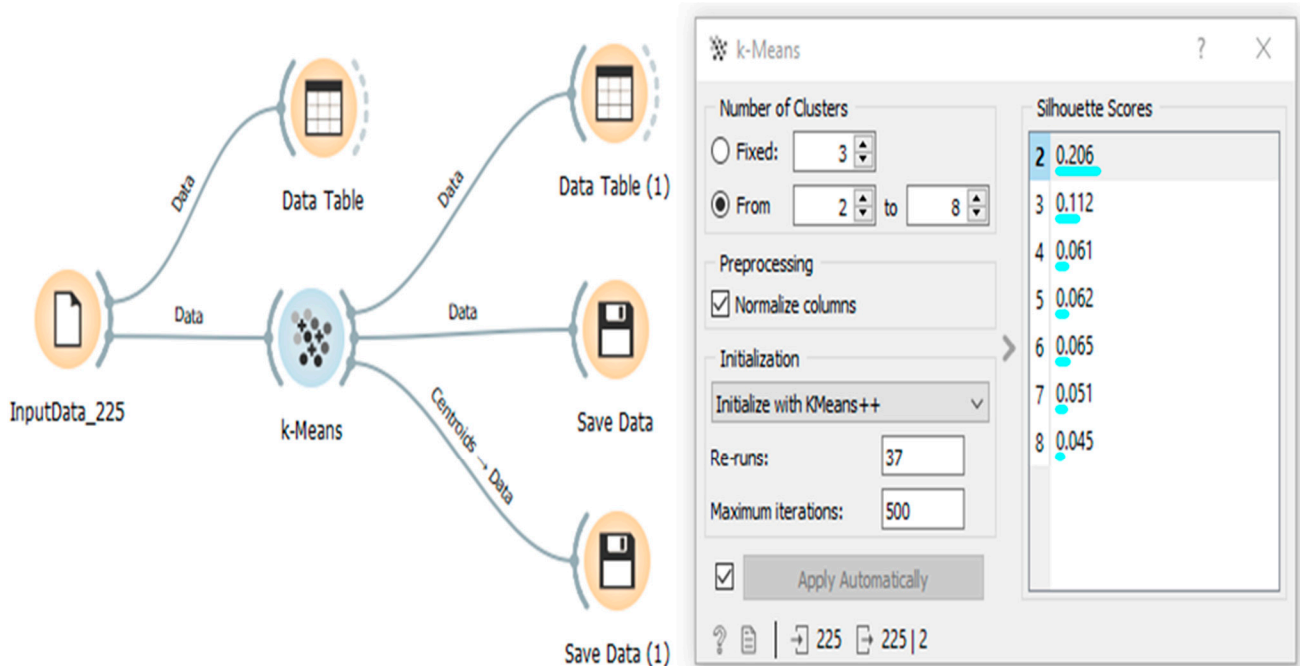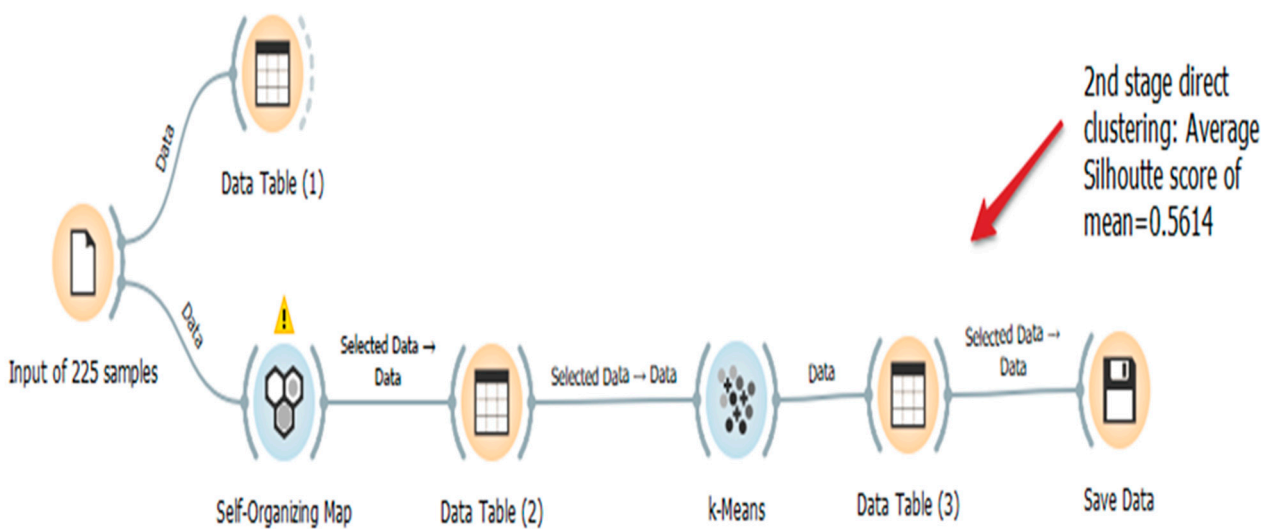**Figure 6.** Workflow of direct KM algorithm.

Figure 7. Workflow of best-fitted 1st-stage KM model.

Figure 8. Workflow of best-fitted 2nd-stage indirect clustering model.

Step three was data preprocessing using correlation methods and feature extractions, as shown in Figure 5. Initially, 77 houses with a complete and validated data set were considered for the correlation study. The purpose of the correlation study was to optimize the data features and improve data quality. In the 77 samples, the data include 1.1% missing values that were imputed using the average and most frequent method. Moreover, the Spearman (S) and Pearson (P) correlation methods identified 23 poorly associated variables with the target variable, as shown in Table 4. Furthermore, we added eight new features to the QS data set by using the FE technique, named TotalReguAppl, TotalLifeAppl, Total*Appliances, LiveDate, RespSubmission, Days, Months, and 5YearOldNo. The FE technique aims to understand the data set, improve the associations between input and output variables, increase the quality of the input data set, add new features, and enhance the overall clustering performance.

On the other hand, the correlation methods identified the six highly influencing variables with the Avg_kWh target variable, namely TotalReguAppl, Rooms, Geyser, CelingFansNo, 5YearOldNo, and WaterPurifier. Of these, TotalReguAppl and 5YearOld variables are from newly added features. These newly added significant features are one of the contributions of this study. Moreover, these features are also significant for optimal consumer classification and prediction accuracy. Further, a comparative result analysis was performed. Figure 8 shows the workflow diagram with the hyperparameters of the best-performed 2nd-stage SOM with the KM clustering model.

4. Results and Discussion

The QS data-collection-based HEC case study may produce an overlapped cluster. This overlapped cluster can be reduced after data preprocessing steps, and through clustering or classification models [20]. This case study applied DM techniques, namely expert classification and direct, indirect and multi-stage clustering methods, as shown in Figures 4 and 5. The results of the expert classification are provided in this section. This study applied KM, H, and SOM algorithms and developed eight models for OHCC. From the eight clustering models, this study focused on three models from each clustering method, as shown in Figures 6–8.

4.1. New Classes for Categorizating Household Consumers

The input–output framework of the proposed OHCC is shown in Figure 5. The categorization of the household consumer mainly depends on the requirement of the applications and the type of data set. At the same time, the proposed household consumer classes can be optimal and overlap-free [29]. According to the tariff order of the top

seven utility companies of India, the energy consumption slabs can be considered for categorizing the household consumers, as shown in Table 5 [17,45–49,52]. Table 5 reports that all the utility companies decided their own consumption slabs. The minimum energy consumption slabs are different, namely 30, 40, 50, 75, and 100. After a rigorous study of different tariff orders [17,45–49,52] and understanding the different data sets, this study proposed 0–60 kWh, which can be the initial consumption slab.

In addition, the 0–60 kWh consumption slab was compared with the average minimum consumption of SOM with the KM indirect clustering model, which was 51 kWh. The difference of 9 kWh can be further reduced by correcting data sets and improving the data quality. Table 6 shows the proposed new classes for categorizing the 225 houses. However, the consumption range of new classes is almost equal to 60% of the consumption slabs of the MERC tariff order.

Table 6. Consumers' class formation of 225 houses.

Group No.	Class Name	Class Range (kWh)	No. of Consumers
1	LEDH	0 to 60	38
2	MEDH	61 to 120	95
3	PEDH	121 to 280	86
4	EPEDH	Above 280	06

Note: LEDH = less energy demand house; MEDH = moderate energy demand house; PEDH = peak energy demand house; EPEDH = extra peak energy demand house.

### 4.2. Direct and Indirect Clustering Methods

The three clustering methods were used for OHCC, namely direct, indirect and multi-stage, as shown in Figure 5. First was the direct clustering method in which the data set is directly given to the clustering algorithms without applying the Correlation and FE techniques. Second was the indirect clustering method in which the data set goes through the correlation and FE techniques. Third was the multi-stage clustering, which is the extension of indirect clustering. This case study developed eight clustering models. Here, the three most fitted models are selected from the eight models. This study was developed in two stages: the 1st stage used one algorithm and the 2nd stage added two algorithms. Moreover, a total of eight models were developed. Figures 6–8 show the best-fitted model from each clustering method. The overall eight models with twelve parameters are evaluating. Among these, the SS mean and overlap-free sample are the most important.

The results of the direct clustering method show 0% of overlap-free samples in all of the clusters, whereas the KM direct model was the most fitted compared to H and SOM, with the highest SS median being 0.527. In the case of the 1st-stage indirect model, an improved performance was observed as compared to the direct method. The overlap-free samples have improved from 0% to maximum 87% in KM model. In addition, the SS medians of the 1st-stage indirect H and SOM models were improved to 0.403 and 0.22, respectively. This improvement in the indirect clustering models occurred due to the correlation and FE techniques. Thus, the KM algorithm was the most fitted in the direct and 1st-stage indirect clustering model. As a result, for the development of the 2nd-stage indirect clustering method, the KM algorithm was used at the output side, as shown in Figure 8. The SOM with KM indirect clustering model yielded a better performance with the highest percentage of overlap-free samples (88%) and an SS median of 0.531.

Thus, the SOM with KM indirect model was considered further for the comparison of the expert classification OHCC, as shown in Figure 9. Furthermore, Tables 7–10 shows the 1st-stage KM indirect and 2nd-stage SOM with KM indirect clustering results, respectively. The 1st-stage and 2nd-stage clustering models generated two clusters, C1 and C2, which are optimal clusters and auto-generated by the Orange tool based on the higher SS value. Moreover, the major parameters considered for the result analysis were the SS mean, average energy consumption (Avg_kWh), seasonal average energy consumption (Avg_kWh(S)),

and highly significant variables using the S and P correlation methods. The seasonal average energy consumption parameter was formed after a rigorous analysis of each cluster based on the monthly energy consumption. However, the difference between the mean and median parameters can provide information on the distribution of the data points. There were three common significant variables in C1 and C2, namely geyser, television, and ceiling fan. The SS mean of the maximum monthly consumption in C1 was the highest, which was 0.63.
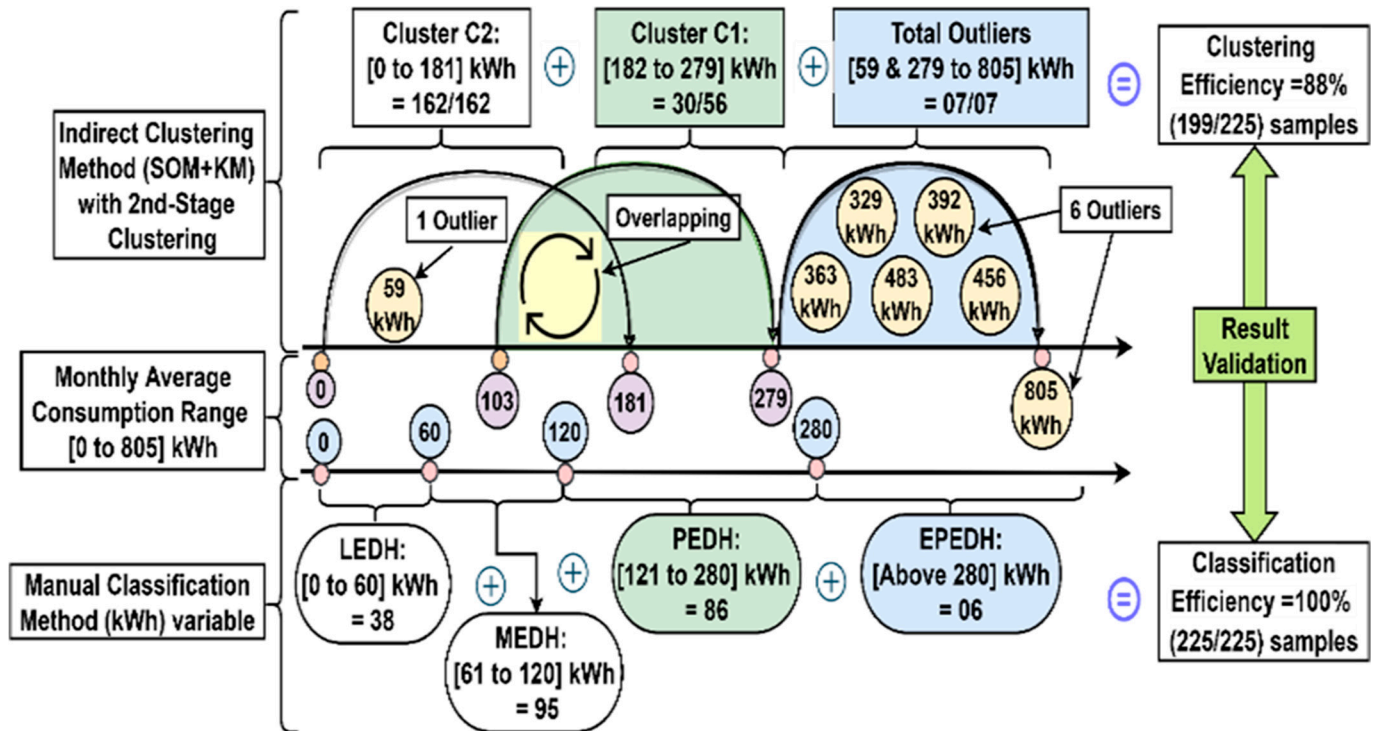
Figure 9. Proposed DM techniques' result comparison for OHCC. Note: LEDH = less energy demand house; MEDH = moderate energy demand house; PEDH = peak energy demand house; EPEDH = extra peak energy demand house; kWh = kilowatt-hour; SOM = self-organizing map, KM = K-means; Outlier = data points away from the sample mean; Overlapping = data points belongs to more than single cluster; OHCC = optimal household consumer classification.

Table 7. 1st-stage indirect KM clustering analysis of C1.

Parameters	SS Mean	Avg_kWh	Avg_kWh (S)	Influencing QS Variables						
				TV*	E	G	TV	EG	ZB	CF
Min	0.53	0	65	0	0	0	0	0	0	0
Max	0.63	181	88	6	2	2	2	2	3	6
Mean	0.6	89	79	3.7	0.8	0.3	1	0	0.2	2.5
Median	0.6	87	79	4	1	0	1	0	0	3
Mn-Md	0	2	0	0	0	0	0	0	0	0.5

Note: QS = questionnaire survey; SS = silhouette score; Avg_kWh = Average kilowatt-hours; Avg_kWh (s) = average kilowatt-hours seasonal; TV* = television with star ratings; E = earth leakage circuit breaker (ELCB); G = geyser; EG = electronic games; ZB = zero bulbs; CF = ceiling fans.

Table 8. 1st-stage indirect KM clustering analysis of C2.

Parameters	SS Mean	Avg_kWh	Avg_kWh (S)	Influencing QS Variables						
				R	EF	5O	G	WP	TV	CF
Min	0.44	103	138	0	0	0	0	0	1	2

Table 8. *Cont.*

Parameters	SS Mean	Avg_kWh	Avg_kWh (S)	Influencing QS Variables						
				R	EF	5O	G	WP	TV	CF
Max	0.51	279	191	9	3	12	3	3	3	7
Mean	0.41	186	157	2	0.55	5.1	0.6	0.8	1.1	3.5
Median	0.46	186	150	1	0	5	0	1	1	3
Mn-Md	0	0	3	1	0.55	0	0.61	0	0	0.52

Note: Note: QS = questionnaire survey; SS = silhouette score; Avg_kWh = average kilowatt-hours; Avg_kWh (s) = average kilowatt-hours seasonal; R = refrigerator; EF = exhaust fan; 5O = five-year-old appliances; G = geyser; WP = water purifier; TV = television; CF = ceiling fans.

Table 9. 2nd-stage SOM + KM with clustering analysis of C1.

Parameters	SS Mean	Avg_kWh	Avg_kWh (S)	Influencing QS Variables						
				TV*	E	G	TV	EG	ZB	CF
Min	0.44	182	148	0	0	0	1	0	0	2
Max	0.51	279	187	6	2	3	3	1	6	7
Mean	0.47	185	169	3.6	0.9	0.6	1.1	0.1	3.6	3.5
Median	0.46	184	168	4	1	0	1	0	4	6
Mn-Md	0	1	1	0	0	0.6	0	0	0	−2.5

Note: Note: QS = questionnaire survey; SS = silhouette score; Avg_kWh = average kilowatt-hours; Avg_kWh (s) = average kilowatt-hours seasonal; TV* = television star rating; E = earth leakage circuit breaker; G = geyser; EG = electronic games; CF = ceiling fans.

Table 10. 2nd-stage SOM + KM with clustering analysis of C2.

Parameters	SS Mean	Avg_kWh	Avg_kWh (S)	Influencing QS Variables						
				TV	R	CF	WP	EF	G	5O
Min	0.53	0	65	0	0	0	0	0	0	0
Max	0.63	181	97	2	4	6	3	3	2	13
Mean	0.60	88	76	1.0	1.3	2.5	0.4	0.1	0.3	3.4
Median	0.60	87	74	1	2	3	0	0	0	3
Mn-Md	0	1	2	0	−0.7	0	0	0	0	0

Note: QS = questionnaire survey; SS = silhouette score; Avg_kWh = average kilowatt-hours; Avg_kWh (s) = average kilowatt-hours seasonal; TV = television; R = refrigerator; CF = ceiling fans; WP = water purifier; EF = exhaust fan; G = geyser; 5O = five-year-old appliances.

Furthermore, C1 was influenced by two seasons, i.e., the rainy season from June to September and the winter season from October to January, whereas C2 was influenced by two seasons, i.e., the summer season from February to May, as well as the winter season. The case study is in the Indian context, so seasons are considered as per the Indian monsoon. Tables 9 and 10 show the performance of the 2nd-stage SOM with KM indirect clustering with the C1 and C2 clusters. In the C1 and C2 clusters, it was observed that energy consumption was mainly influenced by the winter season. Moreover, except for the KM clusters, the other algorithm showed outlier values in a separate cluster. Table 11 shows the status of the outliers in all of the models.

4.3. Result Comparision within Expert Classification and Clustering Methods

The 2nd-stage indirect SOM with KM clustering method was compared to the remaining seven models and the best-fitted model is proposed. This perfect-fit clustering model was considered for a result comparison with the household consumer classification by the expert classification method. The overlap-free sample parameter was used for the comparison of the classification and clustering methods. The proposed clustering method had 12% overlapping and 88% efficiency, as shown in Figure 9. The monthly consumption of 225 houses ranged from 0 kWh to 805 kWh. In the clustering process, clusters C1 and C2 were overlapped between the 103 kWh and 181 kWh consumption values. The cluster

C1 was overlapped with 26 samples that were removed due to having a lower SS mean than C2. Furthermore, the seven outliers were identified using the post-clustering method. Among the seven outliers, six outliers were also included in the EPEDH class, as shown in Figure 9. Figure 9 and Table 12 conclude that cluster C2 consisted of two classes: less energy demand house (LEDH) and moderate energy demand house (MEDH). The class of peak energy demand house (PEDH) belongs to the cluster C1. Moreover, the 12% overlapped data can be minimized by using data validation using significant features or questions from the QS.

Table 11. Comparative result analysis for direct and indirect clustering methods.

Clustering Methods	Sr. No.	Parameters	KM	Hierarchical	SOM
Direct clustering method	1	Total Samples	225	225	225
	2	No. of clusters	03	05	05
	3	No. of outlier clusters	00	01	01
	4	Total outlier samples	06	09	11
	5	Min (kWh) Mean	60	56	13
	6	Max (kWh) Mean	221	223	225
	7	Avg (kWh) Mean (Mn)	141	128	106
	8	Avg (kWh) Median (Md)	141	130	94
	9	Avg(kWh) Md-Mn Mean	0	+2	−12
	10	SS Mean	0.522	0.027	−0.054
	11	SS Median	0.527	0.069	0.011
	12	Overlap-free samples	0%	0%	0%
Indirect clustering method: 1st stage	1	Total Samples	225	225	225
	2	No. of clusters	02	03	05
	3	No. of outlier clusters	00	01	01
	4	Total outlier samples	30	08	44
	5	Min (kWh) Mean	51	42	32
	6	Max (kWh) Mean	230	239	169
	7	Avg (kWh) Mean (Mn)	137	135	96
	8	Avg (kWh) Median (Md)	136	136	92
	9	Avg(kWh) Md-Mn Mean	−1	+1	−4
	10	SS Mean	0.505	0.295	0.196
	11	SS Median	0.531	0.403	0.222
	12	Overlap-free samples	87%	82%	0%
Indirect clustering method: 2nd stage	1	Total Samples		225	225
	2	No. of Clusters		03	02
	3	No. of outlier clusters		01	00
	4	Total outlier samples		08	33
	5	Min (kWh) Mean		51	51
	6	Max (kWh) Mean	NA	222	230
	7	Avg (kWh) Mean (Mn)		131	136
	8	Avg (kWh) Median (Md)		129	135
	9	Avg(kWh) Md-Mn Mean		−2	−1
	10	SS Mean		0.531	0.532
	11	SS Median		0.530	0.531
	12	Overlap-free samples		83%	88%

Note: KM = K-means; SOM = self-organizing map; mean = Mn; median = Md.

Table 12. Consumer classification and its validation using the SOM with KM indirect clustering method.

Sr. No.	CN	CR (kWh)	MC (kWh)	CV (kWh)	PCA
1	LEDH	0 to 60	38	37	C2
2	MEDH	61 to 120	95	92	C2
3	PEDH	121 to 280	86	63	C1 + C2
4	EPEDH	Above 280	06	07	Expert

Table 12. *Cont.*

Sr. No.	CN	CR (kWh)	MC (kWh)	CV (kWh)	PCA
		Overlapping	00	26	
		Total	225	225	

Note: CN = consumer class; CR = class range; MC = manual classification; CV = clustering validation; PCA = post clustering analysis; LEDH = less energy demand house; MEDH = moderate energy demand house; PEDH = peak energy demand house; EPEDH = extra peak energy demand house; C1 = cluster 1; C2 = cluster 2.

In addition, the overlapped samples may be present due to energy theft or other issues that need to be identified. Thus, the open-ended research question is overlapped with data and its correction mechanism [18]. Table 12 shows the expert consumer classification result validation using the 2nd-stage SOM with KM indirect clustering techniques. Table 12 has notations such as class name (CN), class range (CR), manual-based classification (MC), validation using clustering (VC), and post-clustering analysis (PCA). C2 exhibited 97% clustering efficiency compared to the expert consumer classification approach, as shown in Table 12.

4.4. Case Study

This case study applied a random sampling technique to collect 225 samples from three districts of Maharashtra, India. There are four significant implications of this work: First, to classify domestic consumers to form groups with the same levels of consumption; Second, to optimize the memory size of the collected data sets and improve the classification and clustering operation speed, using different techniques such as correlations and feature engineering; Third, to help select the best clustering model by comparing other models based on different parameters; Fourth, this study collected primary and secondary data sets of individual domestic consumers.

This addresses ground-level consumption-related issues through experience and observations of study field visits that may help policy makers, and helps the company to attract consumers by providing consumption-based offers. The consumer classification was carried out by studying the tariff orders of India's top seven utilities, and in order to understand the data sets, four new classes of consumers were proposed. Under the indirect clustering method, the correlation methods such as Spearman and Pearson and different feature-engineering techniques were applied in order to improve the data quality and optimize the data set features [43]. The indirect clustering method was further implemented in two stages. The result shows the self-organizing map with K-means having 88% overlap-free samples and a silhouette score mean of 0.532. This case study will help to develop a novel method for household electricity consumer classification and clustering approaches. This research work will benefit researchers, policy makers, and electricity distribution companies. This work provides the model for efficient electricity distribution and offers better services to consumers.

4.5. Threats to Validity for Classification Results

The direct comparison between the result of the clustering methods results and the existing models or techniques for the HEC study is challenging. This challenge is due to variations in the implications, proposed methodology, data-collection methods, data characteristics, selected variables, and the used validation indicators that can be applied differently in different works, which can vary the results of clustering models. The performance of the clustering model can be compared or verified when the model works on the same data sets, methodology, and implications. In comparison, every clustering algorithm has its potential and limitations. Thus, one algorithm cannot perform the best on all data sets and applications under all conditions [10,25,44]. Due to these threats, this paper cannot directly compare the proposed results with the existing techniques or models for the HEC study. This paper used the Orange tool to develop clustering methods and fine-tune the

hyperparameters of the exact data, methodology, and implications using KM, H, and SOM algorithms. This method validated the expert consumer classification results.

5. Conclusions and Future Directions

This study addressed the challenges in OHCC using conventional-grid- and SG-system-based energy meter data sets. This work was based on conventional grid infrastructure, which includes manual meter reading and billing processes. For the case study, data sets from 225 houses in three districts of Maharashtra, India were used. The work noted that from March 2019 to August 2020, around 19% of electricity bills were average bills and 81% were normal bills. Thus, the MEC data set does not always reflect the actual energy consumption pattern and energy demand. Due to this, the conventional-grid-based MEC data set may not be adequately helpful to provide the OHCC. In addition, smart-meter-based energy data only has energy consumption data and neglects other required data sets for OHCC such as house characteristics, socio-demographic factors, appliance characteristics, feedback and awareness, and so on. However, according to the literature, this is the first survey-based study that will predict the individual household MEC for energy optimization using different locations and various data sets [22,53]. Obtaining the OHCC is the first stage for achieving a better accuracy of prediction. With this in mind, the work proposed a novel methodology to obtain OHCC through expert classification and clustering methods. This work used data sets, namely QS, MEC, and tariff orders of utility companies. The QS was structured and consisted of six parts, namely basic information, electricity bill, house characteristics, socio-demographic factors, appliance characteristics, feedback, and awareness. For OHCC, the expert classification method provided four new classes in kWh, namely 0 to 60 (LEDH), 61 to 120 (MEDH), 121 to 280 (PEDH), and above 280 (EPEDH). The clustering methods for OHCC used direct, indirect, and multi-stage clustering models and developed eight clustering models. After comparing the results of eight clustering models, the two-stage SOM model with a KM clustering was found to be the most accurate. Further, the classification results of expert classification and two-stage SOM with the KM clustering model were compared. The result shows that the SOM model with KM clustering provided 88% overlap-free samples and a silhouette score mean of 0.532. Moreover, this work can be used to improve the energy consumption prediction accuracy. The OHCC may be useful for utility companies to make an informed decision about DRM. In addition to this, utility companies may target specific household consumers to provide better offers and services to them.

This work can be explored by deploying the proposed methodology on different regions, smart-meter data sets and a more comprehensive history of consumption data, increased sample sizes, and to apply other consumer types such as commercial and industrial. Moreover, this work can also be extended by considering different variables such as behavioral consumption, the types of days (workday, weekend, etc.), types of seasons (summer, winter), weather parameters (temperature, wind speed, humidity, etc.), and appliance-consumption-based data. The restrictions of the developed method are that the proposed clustering model cannot be directly deployed to improve the smart-meter data set analysis.

Author Contributions: Conceptualization, G.S.R., H.R., S.M.M. and K.K.; methodology, G.S.R.; software, G.S.R.; validation, G.S.R., H.R., S.M.M. and K.K.; formal analysis, G.S.R., H.R., S.M.M. and K.K.; investigation, G.S.R., H.R., S.M.M. and K.K.; resources, G.S.R., H.R., S.M.M. and K.K.; data curation, G.S.R.; writing—G.S.R.; writing—review and editing, G.S.R., H.R., S.M.M. and K.K.; visualization, G.S.R.; supervision, H.R., S.M.M. and K.K.; project administration, G.S.R., H.R., S.M.M. and K.K. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the study is not completed and having the limitations to share the data.

References

1. Ullah, F.U.M.; Ullah, A.; Haq, I.U.; Rho, S.; Baik, S.W. Short-term prediction of residential power energy consumption via CNN and multi-layer bi-directional LSTM networks. *IEEE Access* **2020**, *8*, 123369–123380. [CrossRef]
2. Cai, H.; Shen, S.; Lin, Q.; Li, X.; Xiao, H. Predicting the energy consumption of residential buildings for regional electricity supply-side and demand-side management. *IEEE Access* **2019**, *7*, 30386–30397. [CrossRef]
3. Ramnath, G.S.; Harikrishnan, R. Households electricity consumption analysis: A bibliometric approach. *Libr. Philos. Pract.* **2021**, *5098*, 1–21. Available online: https://digitalcommons.unl.edu/libphilprac/5098 (accessed on 24 April 2021).
4. Pablo-Romero, M.D.P.; Pozo-Barajas, R.; Yñiguez, R. Global changes in residential energy consumption. *Energy Policy* **2017**, *101*, 342–352. [CrossRef]
5. Nejat, P.; Jomehzadeh, F.; Taheri, M.M.; Gohari, M.; Majid, M.Z.A. A global review of energy consumption, CO_2 emissions and policy in the residential sector (with an overview of the top ten CO_2 emitting countries). *Renew. Sustain. Energy Rev.* **2015**, *43*, 843–862. [CrossRef]
6. Ozawa, A.; Kudoh, Y.; Yoshida, Y. A new method for household energy use modeling: A questionnaire-based approach. *Energy Build.* **2018**, *162*, 32–41. [CrossRef]
7. Wang, Y.; Chen, Q.; Kang, C.; Zhang, M.; Wang, K.; Zhao, Y. Load profiling and its application to demand response: A review. *Tsinghua Sci. Technol.* **2015**, *20*, 2. [CrossRef]
8. Rajabi, A.; Li, L.; Zhang, J.; Zhu, J.; Ghavidel, S.; Ghadi, M.J. A review on clustering of residential electricity customers and its application. In Proceedings of the 20th International Conference on Electrical Machines and Systems, Sydney, NSW, Australia, 11–14 August 2017. [CrossRef]
9. Rhodes, J.D.; Cole, W.J.; Upshaw, C.R.; Edgar, T.F.; Webber, M.E. Clustering analysis of residential electricity demand profiles. *Appl. Energy* **2014**, *135*, 461–471. [CrossRef]
10. Wang, Y.; Chen, Q.; Hong, T.; Kang, C. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Trans. Smart Grid* **2019**, *10*, 3125–3148. [CrossRef]
11. Chicco, G.; Napoli, R.; Piglione, F. Comparisons Among Clustering Techniques for Electricity Customer Classification. *IEEE Trans. Power Syst.* **2006**, *21*, 933–940. [CrossRef]
12. Chicco, G.; Napoli, R.; Piglione, F.; Postolache, P.; Scutariu, M.; Toader, C. Load pattern-based classification of electricity customers. *IEEE Trans. Power Syst.* **2004**, *19*, 1232–1239. [CrossRef]
13. Chen, X.; Zanocco, C.; Flora, J.; Rajagopal, R. Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation. *Appl. Energy* **2022**, *318*, 119109. [CrossRef]
14. Ramnath, G.S.; Harikrishnan, R. Social Welfare Maximization in Smart Grid: Review. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1099*, 012023. [CrossRef]
15. Ramnath, G.S.; Harikrishnan, R. A Demand Response Program for Social Welfare Maximization in the Context of the Indian Smart Grid: A Review. In *Cyber-Physical, IoT, and Autonomous Systems in Industry 4.0*; CRC Press: Boca Raton, FL, USA, 2021.
16. Yilmaz, S.; Chambers, J.; Patel, M. Comparison of clustering approaches for domestic electricity load profile characterisation-Implications for demand side management. *Energy* **2019**, *180*, 665–677. [CrossRef]
17. Before the Maharashtra Electricity Regulatory Commission World Trade Center, 2020, 3, 1–752. Available online: https://www.mahadiscom.in/consumer/wp-content/uploads/2020/03/Order-322-of-2019.pdf (accessed on 26 June 2021).
18. Jang, M.; Jeong, H.C.; Kim, T.; Suh, D.H. Empirical analysis of the impact of COVID-19 social distancing on residential electricity consumption based on demographic characteristics and load shape. *Energies* **2021**, *14*, 7523. [CrossRef]
19. Ramnath, G.S.; Harikrishnan, R. A statistical and predictive modeling study to analyze impact of seasons and COVID-19 factors on household electricity consumption. *J. Energy Syst.* **2021**, *5*, 252–267. [CrossRef]
20. Krarti, M.; Aldubyan, M. Review analysis of COVID-19 impact on electricity demand for residential buildings. *Renew. Sustain. Energy Rev.* **2021**, *143*, 110888. [CrossRef]
21. Kumari, A.; Tanwar, S. A data analytics scheme for security-aware demand response management in smart grid system. In Proceedings of the IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Prayagraj, India, 27–29 November 2020. [CrossRef]
22. Ali, M.; Prakash, K.; Macana, C.; Bashir, A.K.; Jolfaei, A.; Bokhari, A.; Klemeš, J.J.; Pota, H. Modeling residential electricity consumption from public demographic data for sustainable cities. *Energies* **2022**, *15*, 2163. [CrossRef]
23. Yildiz, B.; Bilbao, J.I.; Dore, J.; Sproul, A. Household electricity load forecasting using historical smart meter data with clustering and classification techniques. In Proceedings of the IEEE Innovative Smart Grid Technologies—Asia (ISGT Asia), Singapore, 22–25 May 2018; pp. 873–879. [CrossRef]

24. Figueiredo, V.; Rodrigues, F.; Vale, Z.; Gouveia, J. An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. *IEEE Trans. Power Syst.* **2005**, *20*, 596–602. [CrossRef]
25. Ramnath, G.S.; Harikrishnan, R. Problem-based learning on household electricity consumption analysis using predictive models and tools. *Comput. Appl. Eng. Educ.* **2022**, 1–17. [CrossRef]
26. Prahastono, I.; King, D.; Özveren, C.S. A Review of Electricity Load Profile Classification Methods. In Proceedings of the 42nd International Universities Power Engineering Conference, Brighton, UK, 4–6 September 2007; pp. 1187–1191.
27. Zhou, K.-L.; Yang, S.-L.; Shen, C. A review of electric load classification in smart grid environment. *Renew. Sustain. Energy Rev.* **2013**, *24*, 103–110. [CrossRef]
28. Singh, N.; Singh, D. Performance evaluation of k-means and hierarchal clustering in terms of accuracy and running time. *Int. J. Comput. Sci. Inf. Technol.* **2012**, *3*, 4119–4121.
29. Chicco, G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* **2012**, *42*, 68–80. [CrossRef]
30. Zhang, Y.; Chen, W.; Xu, R.; Black, J. A Cluster-Based Method for Calculating Baselines for Residential Loads. *IEEE Trans. Smart Grid* **2016**, *7*, 2368–2377. [CrossRef]
31. Tanwar, S.; Kumari, A.; Vekaria, D.; Raboaca, M.S.; Alqahtani, F.; Tolba, A.; Neagu, B.-C.; Sharma, R. GrAb: A deep learning-based data-driven analytics scheme for energy theft detection. *Sensors* **2022**, *22*, 4048. [CrossRef]
32. Tsekouras, G.J.; Hatziargyriou, N.D.; Dialynas, E.N. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Trans. Power Syst.* **2007**, *22*, 1120–1128. [CrossRef]
33. Chicco, G.; Napoli, R.; Postolache, P.; Scutariu, M.; Toader, C. Customer characterization options for improving the tariff offer. *IEEE Trans. Power Syst.* **2002**, *22*, 11. [CrossRef]
34. Wijaya, T.K.; Vasirani, M.; Humeau, S.; Aberer, K. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In Proceedings of the IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 879–887. [CrossRef]
35. Zakaria, Z.; Lo, K.L. Two-stage fuzzy clustering approach for load profiling. In Proceedings of the 44th International Universities Power Engineering Conference (UPEC), Glasgow, UK, 1–4 September 2009.
36. Teeraratkul, T.; O'Neill, D.; Lall, S. Shape-based approach to household electric load curve clustering and prediction. *IEEE Trans. Smart Grid* **2017**, *9*, 5. [CrossRef]
37. Piao, M.; Shon, H.S.; Lee, J.Y.; Ryu, K.H. Subspace projection method based clustering analysis in load profiling. *IEEE Trans. Power Syst.* **2014**, *29*, 2628–2635. [CrossRef]
38. Yang, J.; Ning, C.; Deb, C.; Zhang, F.; Cheong, D.; Lee, S.E.; Sekhar, C.; Tham, K.W. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy Build.* **2017**, *146*, 27–37. [CrossRef]
39. Sun, M.; Konstantelos, I.; Strbac, G. C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data. *IEEE Trans. Power Syst.* **2017**, *32*, 2382–2393. [CrossRef]
40. Al-Jarrah, O.Y.; Al-Hammadi, Y.; Yoo, P.D.; Muhaidat, S. Multi-Layered Clustering for Power Consumption Profiling in Smart Grids. *IEEE Access* **2017**, *5*, 18459–18468. [CrossRef]
41. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the 33rd International Conference on Machine Learning (ICML-2016), New York, NY, USA, 5 February 2016; Volume 1, pp. 740–749.
42. Lin, X.; Yu, H.; Wang, M.; Li, C.; Wang, Z.; Tang, Y. Electricity consumption forecast of high-rise office buildings based on the long short-term memory method. *Energies* **2021**, *14*, 4785. [CrossRef]
43. Papageorgiou, G.; Efstathiades, A.; Poullou, M.; Ness, A.N. Managing household electricity consumption: A correlational, regression analysis. *Int. J. Sustain. Energy* **2020**, *39*, 1–11. [CrossRef]
44. Wang, K.; Wang, B.; Peng, L. CVAP: Validation for cluster analyses. *Data Sci. J.* **2009**, *8*, 88–93. [CrossRef]
45. Truing Up for FY 2019-20, Determination of ARR and Tariff for FY 2021–2022 for Uttar Gujarat Vij Company Limited (UGVCL) 2021. Available online: https://www.gercin.org (accessed on 10 September 2021).
46. Tariff Structure for FY 2021′22 as Per Tariff Order Issued by Hon'ble PSERC Vide Its Order 2021, pp. 1–6. Available online: http://pserc.gov.in/pages/tariff-orders.html (accessed on 18 November 2021).
47. Haryana Electricity Regulatory Commission, Distribution & Retail Supply Tariff Approved by the Commission for the, 2021 Notes: In Case of Arc Furnaces/Steel Rolling Mills for Supply at 33 kV and Above, the HT Industrial Tariff at the Corresponding 2021, p. 2. Available online: https://herc.gov.in/WriteReadData/Pdf/DR20210401.pdf (accessed on 26 November 2021).
48. Karnataka_Tariff_Order_MESCOM Electricity Tariff Order 2021, Volume 314. Available online: https://bescom.karnataka.gov.in/storage/pdf-files/RA%20section/Tariff%20rates%20%20FY-2021-22.pdf (accessed on 21 November 2021).
49. Electricity Regulatory Commission. Order on Tariff for Retail Sale of Electricity during 2021. Available online: https://www.apspdcl.in/pdf/Tariff%20Order%20for%20FY%202021-22.pdf (accessed on 16 December 2021).
50. Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data mining toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
51. Sanquist, T.F.; Orr, H.; Shui, B.; Bittner, A.C. Lifestyle factors in U.S. residential electricity consumption. *Energy Policy* **2012**, *42*, 354–364. [CrossRef]

52. True Up for FY 2019-20 2019, APR for FY 2020–2021 and Revised ARR and Tariff for FY for Assam Power Distribution Company Limited (APDCL) 2021. Available online: http://www.aerc.gov.in/APDCL_Tariff_Order_2021_22.pdf (accessed on 23 August 2021).

53. Christantonis, K.; Tjortjis, C. Data mining for smart cities: Predicting electricity consumption by classification. In Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 15–17 July 2019; pp. 1–7. [CrossRef]