1. Title of the thesis

Efficient and Scalable Smart Meter Analytics of Electricity Consumption Patterns and Profiles for Load Forecasting

2. Abstract

In the eyes of researchers and policymakers, the implementation of electricity consumption and conservation practices within residential households is of the utmost importance due to the considerable amount of electricity consumed by homes. It is vital to address this issue as electricity consumption and conservation efforts can lead to a significant reduction in electricity consumption, thereby reducing greenhouse gas emissions and promoting sustainability. Thus, initiatives aimed at encouraging homeowners to adopt energy-efficient technologies and practices. In the proposed research, machine learning algorithms for pattern recognition and a dimension reduction methodology using Self-Organizing Maps followed by KMeans clustering algorithm on Daily Electricity Consumption (D-EC) data have been applied. Finding similarities using SOMKMeans Clustering Algorithm between (D-EC) data points and grouping them is the aim of Clustering Algorithms. Consumers were divided into four groups and assigned labels depending on their consumption patterns. They were also classed further as either consistent or inconsistent consumers based on how much power they consumed with respect to consistency. A proposed algorithm utilizing the SOMKMeans Clustering Algorithm has been proposed to promote consumer awareness and provide alerts regarding their electricity consumption patterns. The algorithm also offers timely recommendations based on their daily minimum and maximum electricity consumption, aiding in efficient electricity consumption. It would be helpful to discover the effect of different variables to produce consumer electricity consumption patterns and identification of Consumption Profiles by determining the link between daily electricity consumption with housing and the demographic characteristics of households. We have analysed the relationship and interdependence of input and output features through various methods such as SOMKMeans Clustering Algorithm, Pearson's, Spearman's Rank, and Kendall's tau correlation techniques. Additionally, a statistical analysis employing correlation coefficients, correlation matrices, and internal assessment metrics has been done to ascertain the effect of housing and demographic characteristics on daily electricity usage. The dataset for the year

2013 consisted of 4,942 households used in the experiment, Studies have revealed that various factors such as household routines, demographic information, and the structural attributes of homes have an indirect effect on an individual's everyday electricity usage. Results of the evaluation metrics' scores of SOMKMeans Clustering Analysis with and without regard to household and demographic characteristics exposed that Housing and demographic factors did not have a significant impact on the training process of the classification model. Reason is that every variable, including family structure and age group, was present in each cluster. Daily Electricity Consumption has a 98%, 95%, and 89% success rate in predicting Consumption Profile according to the ANOVA Test, Chi2 Test, and Mutual Information Feature dependence approaches, respectively.

To simplify the model and improve prediction accuracy for the Consumption Profile of customers, daily electricity usage is an essential feature in calculating the electricity consumption profile of consumers. Recognizing the profile of consumers based on their electricity usage helps electricity suppliers make decisions about their policies and informs consumers about any problems with their regular usage. We compared six classification methods in this study project: KNN, SVM, DT, RF, MLFFNN, and GNB. The performance of classification algorithms was compared using a variety of assessment measures, i.e., Accuracy, Precision, Recall, F1score, AUC, Cohen's Kappa, Hamming Loss, and Mathews Correlation Coefficient to determine which performed the best in categorizing the consumers based on the clustered dataset of daily electricity usage. MLFFNN accurately categorized the electricity consumption of numerous families, with an overall accuracy of 97%. The findings also showed that MLFFNN performed well across all classes, with a F1score of more than 80%. Additionally, the effectiveness of the MLFFNN Classification Model to forecast the Consumption Profile of Consumers has been improved with the implementation of Bayesian Optimization as Hyper Parameter Tuning. After tuning hyperparameters, the MLFFNN-HPT model improved prediction outcomes by 98% overall with class-wise performance exceeding 90% for each metric including Accuracy, Precision, Recall, F1score, AUC, Cohen's Kappa, Hamming Loss, and Mathews Correlation Coefficient.

3. Brief description on the state of the art of the research topic

Researchers have thoroughly investigated the historical context and a concise summary of current research on electricity consumption patterns and profiles, including how much energy is consumed daily and how demographic and housing characteristics impact consumer consumption profiles, as well as many machine learning techniques that researchers have used to train the classification model.

3.1. Clustering of Consumers Based on Electricity Consumption

Customer segmentation based on time series data, such as Daily Electricity Consumption [D-EC], is valuable for creating consumption profiles that illustrate customer usage patterns. Energy Consumption Profiles [ECPs] are employed in various analytical applications, including consumption estimation and control, future energy demand, tariff planning, spotting abnormal electricity usage, and establishing power market strategies [19]. Clustering, which categorizes various consumers or load patterns into discrete classes, can help in the establishment of residential DR programs, and by the use of adequate DR programs, this approach may be utilized to target the right groups will support the utilities in developing customized time-of-use pricing structures or DR programs depending on demand patterns [17] as well as Clustering analysis of the power consumption data gathered by smart meters and other data collection terminals in a smart grid may be used to detect and extract distinct electricity consumption patterns of homeowners [32]. A clustering study of residential power consumption using the K-means [5, 4, 11, 19, 30] approach and agglomerative clustering [19, 30], Decision Tree [4], inverse function approach [8], shape-based clustering algorithm [26] have been presented by authors to determine an effective and comprehensive level of consumer strongly coupled electrical load profile, and few clusters are sufficient to group consumers [5].

KMeans clustering strategy considerably improves cluster quality over raw profile data by employing consumer characteristics that describe the structure of home power consumption profiles [28]. The authors analyzed the electricity usage pattern for three to six clusters, and KMeans clustering's initial centroids are picked at random, resulting in consistently changing clustering results [14]. Elbow uses the K-means approach to find the optimal number of clusters [23]. The K-means method was selected to run many times with varying cluster sizes,

allowing up to 10 groups of consumers with similar behavior to be identified using various consumption representations to know an ideal number of clusters [25].

Using hourly electricity and weather data, authors have proposed a Hidden Markov framework that describes domestic consumers' thermally-sensitive consumption, which helps create profiles for individuals and groups that may be used to guide DR programs [2]. According to experimental data analysis, hierarchical clustering was the best method to maintain consistent processing time regardless of the cluster count [12]. The daily power usage habits of low-voltage residential consumers in China are examined, and analysis uses a fuzzy cluster validity index (PBMF) and fuzzy c-means (FCM) clustering approach to identify the correct number of clusters [100] and an enhanced fuzzy clustering technique for mining home monthly power use patterns. Demand response (DR) programmers, or demand side management (DSM) in a broader sense, are efficient approaches to encourage family energy consumption behavioral adjustments using price- or incentive-based tactics, according to the economic paradigm as an application of Consumption Profile of Consumer [31].

3.2. Relationship among features of Residential Electricity Data

Consumer electricity usage and patterns depend on various factors when creating consumption profiles, including income, house structure, the number of rooms, the number of occupants, age group, day of the week, weather, outside and inside temperature, and others. The measure of correlation coefficient (r or R) provides statistics on the closeness of two variables and discusses how determining the degree of multicollinearity and mediating or regulating the status of independent factors in a model may be accomplished using the coefficient of linear correlation between two variables [21]. While creating Electricity Consumption Profile, electricity usage by consumer and pattern depends on Demographic-Household Features such as the yearly income of a household, information on the house's layout, the number of rooms, the number of occupants, the age range of the family members, and the daily temperature range, etc. may or may not show a significant role while doing household consumption profile examination. According to the behaviour-oriented paradigm, the intricate interactions between intrapersonal, interpersonal, and environmental factors frequently impact how much energy consumers consume have been discussed in [31]. Daily power usage annually is commonly

influenced by outside factors, such as the mean ambient temperature and the number of daylight hours each day, which generally exhibit similar patterns over time [15].

The authors have used Multiple Linear Regression on 620 urban families in Seremban, Malaysia, to examine how socioeconomic parameters like income and age relate to how much electricity is consumed [3, 20]. Authors found that yearly income and household size have a weakly positive connection with annual electricity consumption when considering the demographic characteristics component, and there was a statistically significant beneficial association between the yearly power use and the age of the household representative and the makeup of the family [16]. Based on the previous research, the primary objective is to find the relation between electricity consumption and demographic-household features and how this relationship will help design the classification model, which can predict the consumption profile of consumers unknown to the trained model.

3.3. Classification of Consumers

Logistic Regression was found to be the most accurate in classifying users based on the kind of intensity of electricity usage, with an average accuracy value of 99%, in the authors' comparison of K-Nearest Neighbor, Support Vector Machine, Decision Tree, and Logistic Regression methods of classification to determine the type of electricity consumption habits in daily household life with an accuracy of more than 95% and they were suitable algorithms for classifying consumers [6]. On electricity data, the authors utilized hierarchical clustering with Ward linkages as customer segmentation and Classification-Regression Tree as consumer classification, and they reached 95.8% prediction accuracy [9]. A classification-prediction model was developed using cross-training ensemble equations that employed the Artificial Neural Network [7, 22] and Decision Tree [7] as the classification method, and the model's effectiveness was evaluated by calculating the weighted average of the error. To determine the household consumption profile from the power data from smart meters, authors have investigated Neural Networks and Elastic Net Logistic regression as classification algorithms on the balanced dataset and achieved 63% and 60% accuracy [10].

For estimating hourly or sub-hourly energy demand in four distinct buildings, models of regression and a neural network-based model with data categorization is provided, and in

comparison, to traditional regression models, the recommended regression approaches and an artificial neural network (ANN) model with the suggested categorization produce highly accurate results for estimating energy consumption [18]. Using data on power usage that is already available, the network has been trained to categorize customers according to their type of electric meter (single-rate or dual-rate) and their residential area (city or village) and as a result, it can accurately estimate a customer's electric meter type with 77% accuracy and their residential zone with 82% accuracy[13]. The authors evaluated the prediction of the power consumption of fifteen anonymous individual households by London Hydro, a local utility company, from 2014 to 2016, using the support vector regression (SVR) modeling technique, applied to both daily and hourly data granularity and provided a relatively efficient and accurate approach [29]. Suggested a random search to adjust the numerous hyper-parameters involved in the method's performance, generating fewer models with competitive accuracy, and the smoothing process lowers the forecasting error [24, 27]. Numerous authors have addressed different algorithms for creating power consumption patterns and profiles. Efficient and scalable generation of ECPs will help train the classification model to predict the consumption profile of consumers who have yet to be known to the trained model.

4. Definition of the Problem

Due to changes in the lifestyle of consumers, climate conditions, and usage of a variety of utilities, day-by-day electricity usage is also increasing.

- Based on the electricity consumption, electricity consumption patterns of consumers can vary significantly within user groups. Discovering particular users' behavioral characteristics is difficult without data analytics and machine learning algorithms.
- It becomes essential to use effective and scalable data analytics and machine learning algorithms for accurate consumption patterns, which will help to predict consumer profiles and load forecasting in smart metering applications based on the responsible factors.
    - Specifically, factors include energy consumption, day-night temperature, number of persons in the family, age of family members, yearly income, number of appliances, house structure, number of bedrooms, and outside weather conditions.

5. Objective and Scope of Work

- Apply machine learning algorithms and statistical methods to find demographic and household factors influencing consumption habits while training the classification model.
- Provide macro-level judgments to electricity providers using machine learning algorithms by analysing consumer profiling based on usage similarity and divergence in daily electricity consumption patterns. This approach enables us to identify consumption patterns and offers valuable insights to electricity providers.
- Using statistical data analytics, machine learning algorithms, and visualization tools to provide a model-based demand-supply decision-making system with prompt recommendations for power savings by offering consumers an awareness and alert system. By indirectly participating in demand management, consumers play an active role in promoting energy efficiency.
- Using machine learning algorithms, developing a step-by-step consumer segmentation and classification model to visualize future energy demands for customized or integrated energy management solutions.

6. Original contribution by the thesis.

Using machine learning algorithms and statistical data analytics, the proposed research offers a chance to examine the understudied area of electricity consumption trends and profiles.

- The research will provide an effective clustering algorithm and classification model for implementing demand and response programs by visualization of incorporated consumption patterns, profiling, and future electricity needs.
- Using an effective SOMKMeans Two-Level clustering algorithm, consumers are segmented according to their daily consumption and consistency. Consumers have been divided into eight categories according to their everyday consumption scenarios, i.e., Low, Moderate, Elevated, and Extravagant, as well as Consistent and Non-Consistent Consumers.
- Consumption profile-based analysis helps to determine the required resources, such as fuels needed to operate the generating plants and other resources necessary to guarantee the efficient and continuous generation of electricity and its delivery to customers. Such Consumption Profiles help to avoid forecasting under-generation or over-generation and

can help energy companies to identify areas where customers could be using electricity more efficiently.

- The proposed MLFFNN using HPT Classification Model has been beneficial in predicting consumption profiles and has helped the energy generator and provider to make good planning decisions since they can estimate future consumption or load demand based on the current consumption habit of consumers.
- Effectiveness of using MLFFNN-based Consumption profiles prediction for load forecasting requires appropriate segmentation of the consumers generated using SOMKMeans Clustering Algorithms,
    - Such load profiles represent each consumer class and the proper selection of the inputs for the neural network.
    - For example, if the consumption profile shows that energy demand typically peaks in the afternoon during hot weather, energy providers can use this information to forecast that energy demand will be high during similar conditions in the future. They can then plan for the required energy generation and distribution capacity to meet this demand.
- To optimize customer satisfaction, utility companies can utilize consumption profiles to strategically plan maintenance services. For example, residential areas can be prioritized for residential areas as per the daily electricity consumption category where demand is minimal. This method enables utilities to gauge consumer demand and efficiently schedule maintenance tasks accurately.
- In essence, load forecasting aids in planning for the size, location, and type of the future generating plant.
    - The utilities most likely generate the electricity close to the load by identifying places or regions with high or rising demand. As a result, the expense of building out transmission and distribution infrastructures and the resulting losses are reduced.
    - Distributed generation systems can aid in reducing the need for long-distance transmission lines and lowering energy losses during transmission by producing electricity closer to where it is required.

- Overall, it offers an effective strategy for deciding area and region-wise future electricity essentials as well as household-wise consistent and non-consistent consumer information. Hence, the Electricity Consumption Profiles can result from applying Load Forecasting tools.

7. Methodology of Research and Results

Household electricity use includes power used by electric appliances. Electricity consumption is the actual demand for the resource applied to the supply. Since residential energy consumption accounts for a sizeable share of global energy consumption, many researchers and decision-makers are driven to encourage energy conservation in residential structures and homes. Residential Energy Consumption (REC) is the total amount of energy used by families for various tasks, including lighting, cooling rooms, running washing machines, cooking, dishwashers, and other everyday tasks. Different electrical end applications affect the REC as a whole, and this must be understood. The methodology is divided into three parts, as shown in Figure 1.

1. Segmentation of Consumers using SOMKMeans Clustering Algorithm
2. Influence of Demographic and Housing Characteristics to predict Consumer Profile
3. Prediction accuracy of MLFFNN using HPT Classification Model

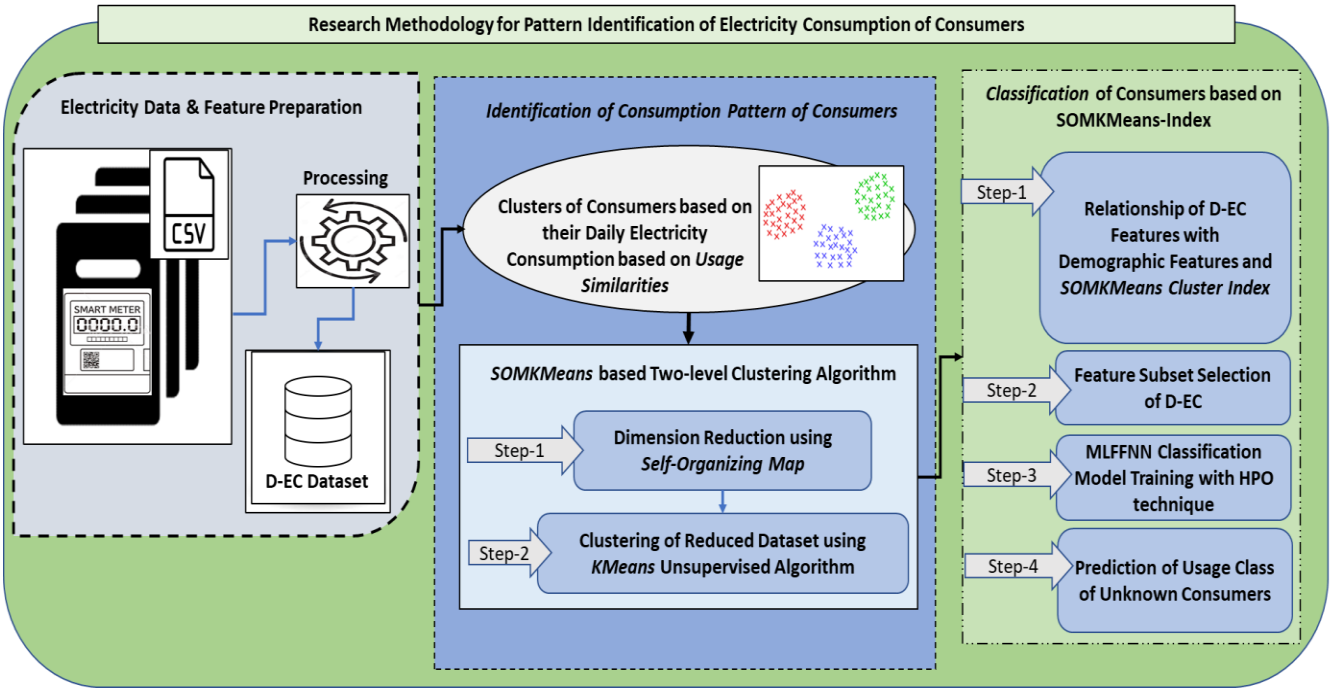

Figure 1 Research Methodology to generate Electricity Consumption Pattern and Profile

7.1 SOMKMeans Clustering Algorithm

Unsupervised Clustering algorithms help recognize patterns among the data points based on the similarity distance between them. Many researchers have applied various clustering algorithms to the Electricity Consumption dataset. Several Unsupervised Clustering Algorithms, including Fuzzy C-Means Clustering, KMeans Clustering, Hierarchical Clustering to construct clusters, and Self-organizing Map as a reduction of data dimensionality, have been used by several authors to analyze electricity consumption patterns and create consumer profiles.

Steps involved in Self Organizing Map based Two-level Clustering Methodology are described below. The formation and explanation of the Proposed SOMKMeans based Two-level Clustering Methodology consist of three stages: Cluster Tendency, Assign Cluster no. to Data Points, and Evaluation of clustering algorithms.

1) Hopkins Statistic's (H) as Cluster Tendency (CT) of D-EC data
2) Generate ECPs of D-EC using Unsupervised Clustering Algorithms
    a) First Method
        i) Find the optimum Number of Clusters (k) using Internal Validity Indices (IVI) Distortion (D) Elbow Method
        ii) Apply Unsupervised Clustering algorithms to assign Cluster Index to each Data Point based on the similarity distance.
    b) Second Method – Self-Organizing Map based Two-level Clustering Methodology to find clusters of D-EC dataset
        i) Apply Self-Organizing Map on D-EC for first-level clustering
        ii) Find the optimum Number of Clusters (k) using Internal Validity Indices (IVI) Distortion (D) Elbow Method of D-EC data, including the SOM Index.
        iii) Apply Unsupervised Clustering Algorithms on D-EC data with SOM-Index to assign SOM-based Cluster Index to each Data Point
3) Assessment of the Clustering Model
    a) Analysis of consumption ranges produced by various clustering algorithms and Methods 1 and 2 results.

b) Examine how Clustering Methods have performed using Silhouette Coefficient (S), Davies-Bouldin Score (DB), and Calinski-Harabasz Score (CH) as Evaluation Metrics.

4) Pick a Clustering approach with outstanding efficiency
5) Observations and Outcomes of Proposed SOMKMeans Clustering Algorithm.

Five clustering algorithms - KMeans Clustering, MiniBatchKMeans Clustering, Gaussian Mixture Clustering, Hierarchical Agglomerative Clustering, and Spectral Clustering were applied to D-EC data points without using Self-Organized Map. Results were compared based on the consumption ranges formed by each cluster. Given the consumption range, there is a problem with D-EC data points overlapping over the identified groups, indicating that data points needed better segmentation. So, the consumption profile label cannot be assigned to such consumers due to the mis-clustering of data. The same problem is found in the results generated by each clustering algorithm.

The experimental findings of D-EC data with SOM-Index and average daily load are presented in Figure 2. At cluster number 4, the SSE starts to flatten out, with a high value for Silhouette Score and a low value for Davies-Bouldin. Thus, based on the findings, four is thought to be the ideal number of clusters (k) to create clusters for the D-EC dataset. Then, Unsupervised Clustering Algorithms, i.e., KMeans Clustering, MiniBatchKMeans Clustering, Gaussian Mixture Clustering, Hierarchical Agglomerative Clustering, and Spectral Clustering, have been executed where average daily consumption with SOM Cluster Index used as input features to produce clusters based on likenesses of electricity usage of households.

Consumption ranges obtained by each clustering algorithm have been depicted in Table 1. All the clustering algorithms have generated distinct clusters regarding electricity consumption except for Gaussian Mixture Clustering Algorithm. Now it is to be checked which algorithm has performed well. Detailed Internal Evaluation Measures have been committed to identify which algorithm has outperformed presented in Figure 3.

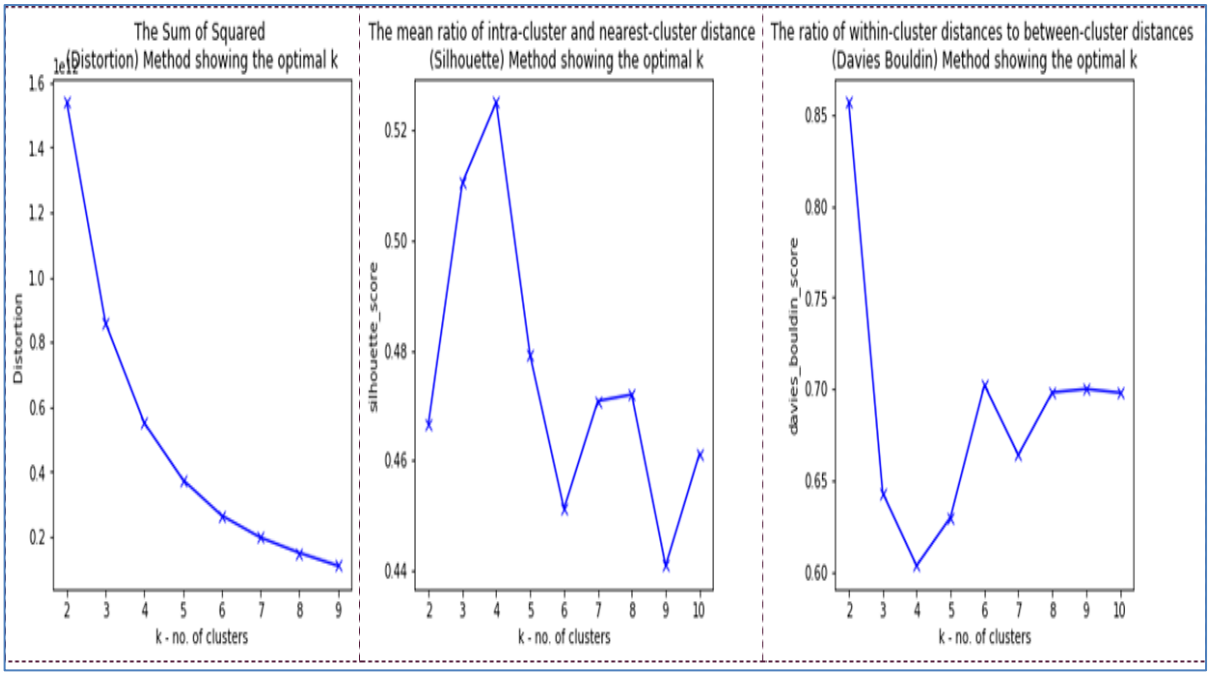

Figure 2 Optimum Cluster no. for D-EC dataset

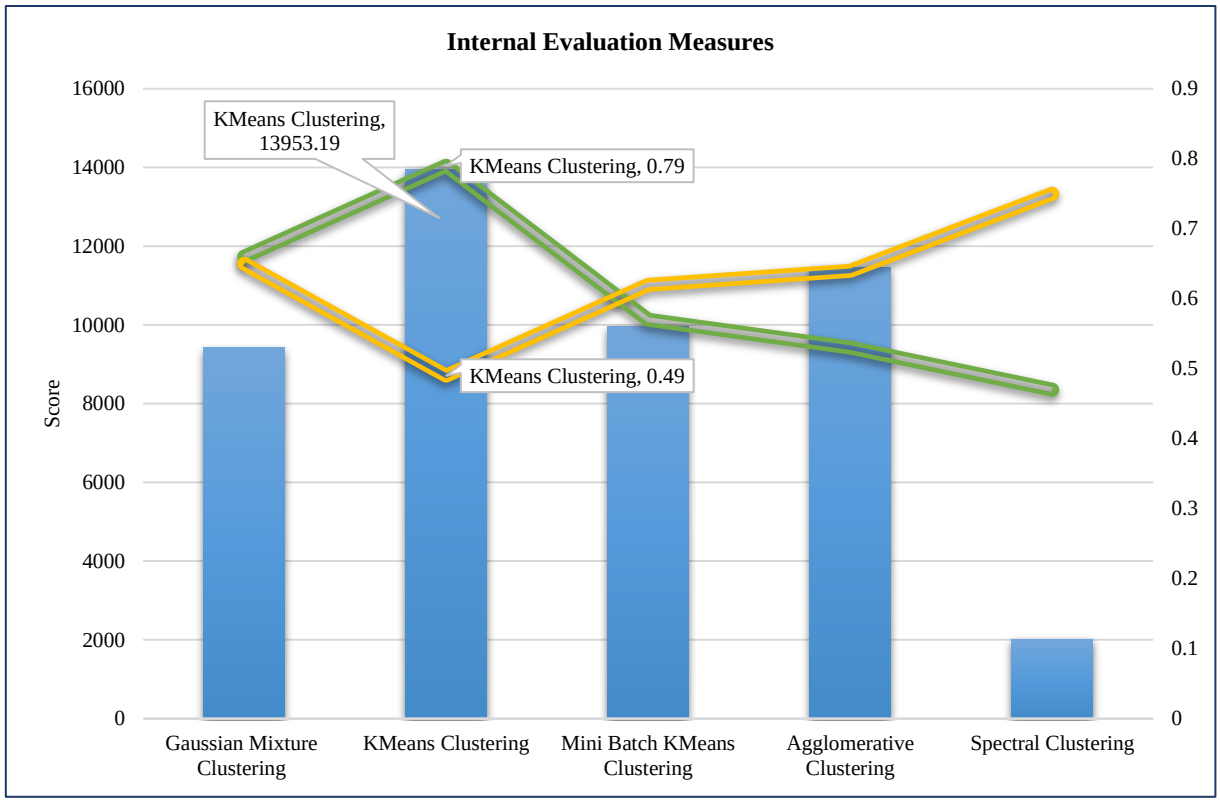

Figure 3 Internal Evaluation Measures

Table 1 Consumption Range

Name of Algorithm	Cluster Number	Total Consumers (%)	Mini-Max Usage (kWh)
Gaussian Mixture Clustering	0	88.46	0 - 57.9
	1	8.58	43.9 - 133.1
	2	2.06	108.8 - 367.1
	3	0.9	224.1 - 953.8
KMeans Clustering	0	93.53	0 - 62.5
	2	4.76	62.9 - 182
	1	1.44	182.3 - 428.1
	3	0.26	452.8 - 953.8
MiniBatchKMeans Clustering	2	64.73	0 - 28.8
	0	30.71	28.8 - 84.9
	3	3.51	85.4 - 238.8
	1	1.06	243.9 - 953.8
Agglomerative Clustering	3	56.12	0 - 25.5
	1	40.85	25.23 - 117.5
	2	2.37	120.9 - 292.2
	0	0.65	301.7 - 953.8
Spectral Clustering	2	24.58	0 - 14.9
	1	42.37	15 - 29.8
	0	25.01	29.8 - 53.8
	3	8.04	54 - 953.8

Internal evaluation metrics, such as the Silhouette Score, Davies Bouldin Score, and Calinski Harabasz Score, are used to measure the efficiency of applied clustering algorithms on the D-EC dataset.

The performance of the KMeans clustering algorithm is superior to other applicable clustering algorithms, according to experimental results of internal evaluation measures for clustering algorithms. It is based on greater Silhouette (S_Score) and Calinski-Harabasz (CH_Score) scores and a lower Davies Bouldin (DB_Score) score, as shown in Figure. 3, KMeans excels. Therefore, a distinctive clustering approach is presented for D-EC data to

produce Electricity Consumption Patterns and Profiles. It is based on SOM-based Two-level Clustering followed by the KMeans Clustering approach.

As a result of the SOMKMeans Clustering Algorithm, Four Clusters with SOMKMeans Cluster Index have been generated and shown in Table 2. Some consumers were found as non-consistent users within that cluster based on their monthly consumption, shown in Figure 5, as well as consumers shown in Figure 4 were using consistent electricity to carry on their day-to-day household activities. Sub-clustering was applied on four clusters using SOM-Kmeans Index and Average Daily Electricity (ADE) consumption to a group of such consumers in different clusters. Outcomes of Sub-clustering are presented in Table 3, and accomplish that in each cluster, a few percentages of consumers were identified as irregular consumers concerning electricity usage.

Table 2 SOMKMeans Cluster - Four Consumer Groups

Sr. No.	Cluster No.	Total Consumers	Percentage (%)	Averaged Consumption (in kWh) range		Consumption Label
				Daily	Monthly	
1	0	6028	93.5	0 - 62.5	0 - 1862.1	Low User
2	2	307	4.8	62.9 – 182	1873.7 – 5419.8	Moderate User
3	1	93	1.4	182.3 – 428.1	5429.5 – 12747.1	Elevated User
4	3	17	0.3	452.8 – 953.8	13484.4 – 28401	Extravagant User

Table 3 Sub-clustering of Consumers as Consistent and Non-Consistent Consumers

Cluster No.	Sub-Cluster No.	Total Consumers	Percentage (%)	Daily Averaged Consumption (in kWh) range	1st Level Label	2nd Level Label
0	0	5591	92.75	0.89 - 62.53	Low User	Consistent User
	1	437	7.25	0 - 62.38		Non-Consistent User
2	0	280	91.21	62.93 - 182.01	Moderate User	Consistent User
	1	27	8.79	62.92 - 163.47		Non-Consistent User
1	0	88	94.62	182.33 - 428.07	Elevated User	Consistent User
	1	5	5.38	198.24 - 375.4		Non-Consistent User
3	0	17	100	452.84 - 953.76	Extravagant User	Consistent User

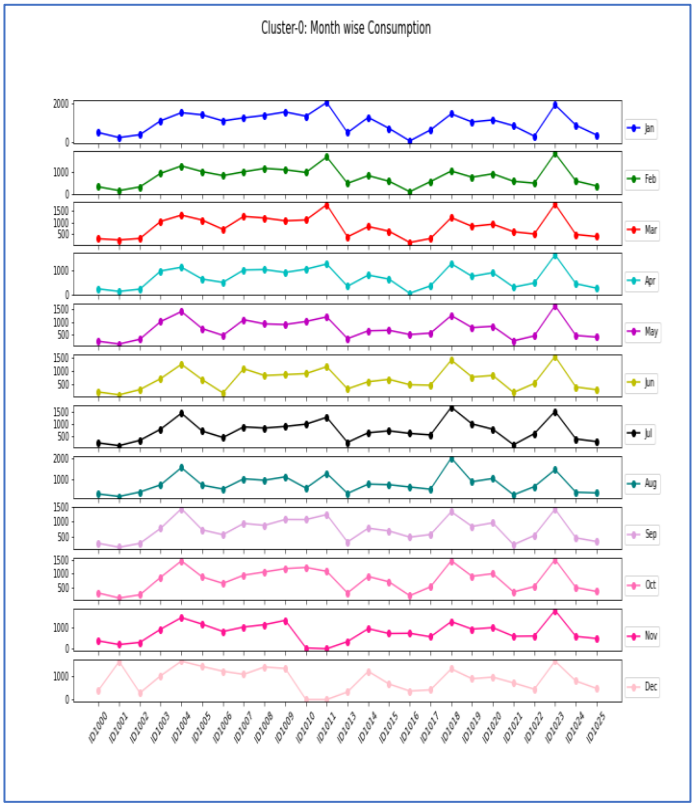

Figure 4 Cluster - 0 Consistent User

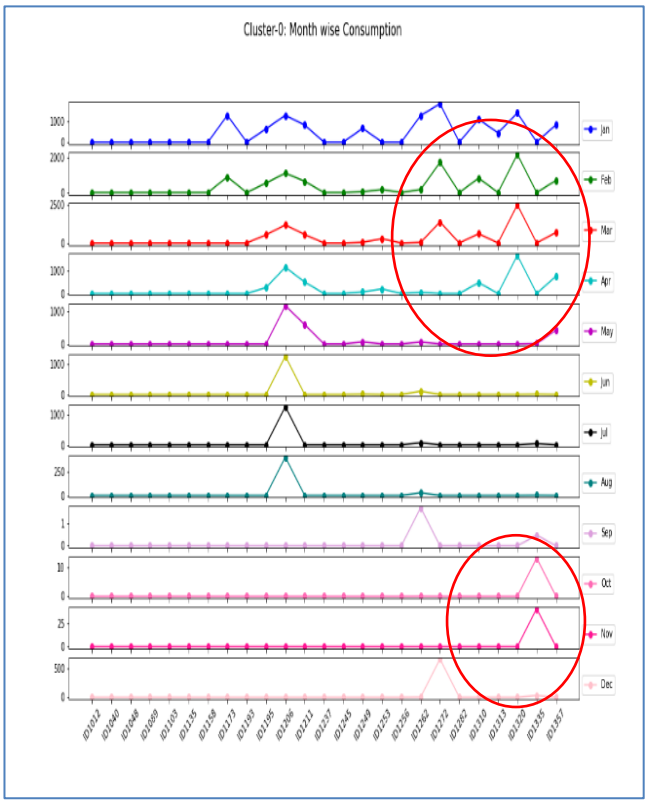

Figure 5 Cluster - 0 Non-Consistent

7.2 Influence of Demographic and Housing Characteristics to predict Consumer Profile

SOMKMeans Two-Level Clustering Algorithm, Correlation Coefficient-Matrix, and Features Dependency methods on Daily Electricity Consumption and Demographic-Household Features have been tested to find relevant features to train the classification model. The systematic objective of this proposed work is to find the Association of Household Features with Electricity Consumption and its Cluster Index, which will help decide whether to consider Household Features to generate Electricity Consumption Patterns and Profiles as well as Classification Models. Informative attributes of any dataset are referred to as Input Features and Output Features.

SOMKMeans Clustering Algorithm: Self-Organizing Map followed by Kmeans Clustering algorithm is used to generate clusters of DEC dataset based on similarities among the data values. Self-Organizing Map on DEC_F data with and without HF followed by KMeans Clustering Algorithms (SOMKMeans) have been implemented to assign SOMKMeans Cluster Index. Experimental results of Internal Evaluation Metrics, i.e., Davies-Bouldin Score,

Calinski-Harabasz Score, and Silhouette Score based on generated as Cluster Index of each consumer of DEC_F data with and without HF (Figure. 6) and a score of mentioned evaluation metrics are similar which have been plotted in Figure. 7.

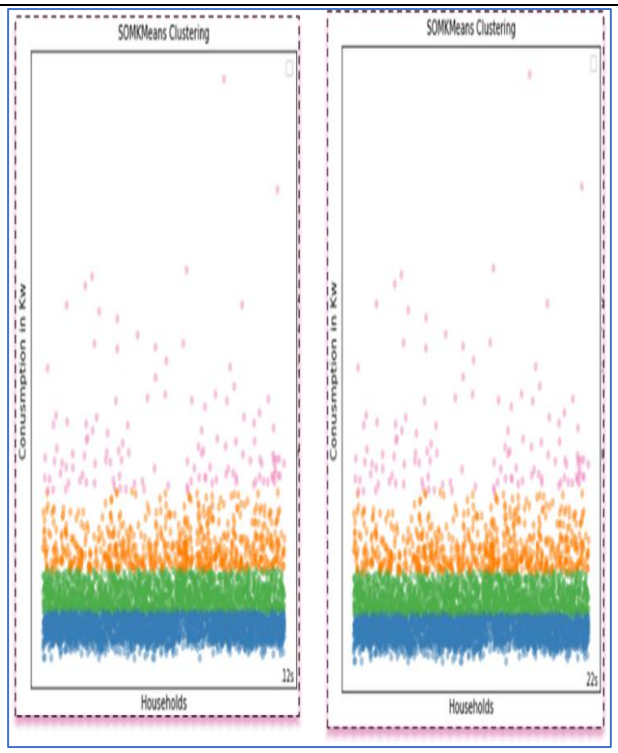

Figure 4 Results of SOMKMeans Clustering Algorithm performed on DEC_F with and without HF

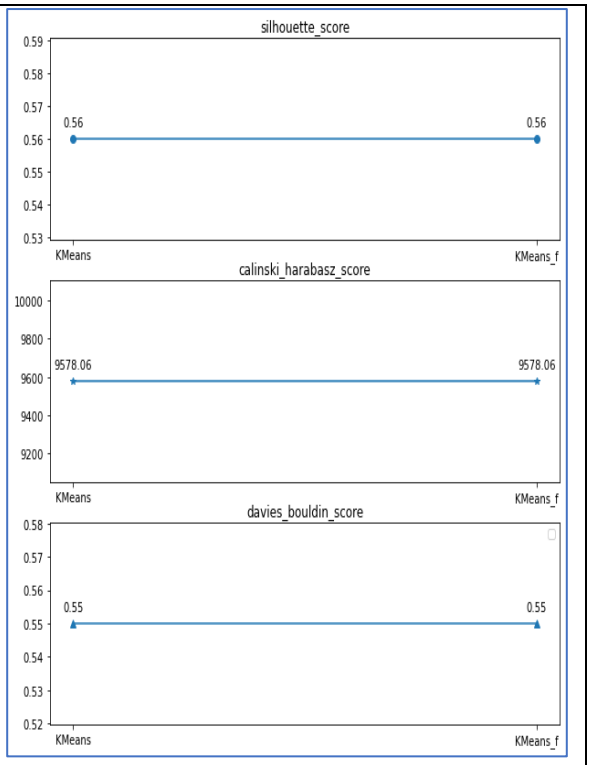

Figure 5 Results of Internal Evaluation Metrics

Correlation Coefficient: The statistical association between two variables is called their correlation. The correlation between average daily electricity consumption and monthly electricity consumption with household features was analyzed using Pearson's correlation coefficient, Spearman's Rank Correlation, and Kendall's tau Correlation methods. The coefficient score can be found in Figures 8(a) and 8(b). Figure. 8 (a) represents the score of each input feature by considering the average daily electricity consumption, and 8 (b) shows a score of the input feature in association with monthly electricity consumption.

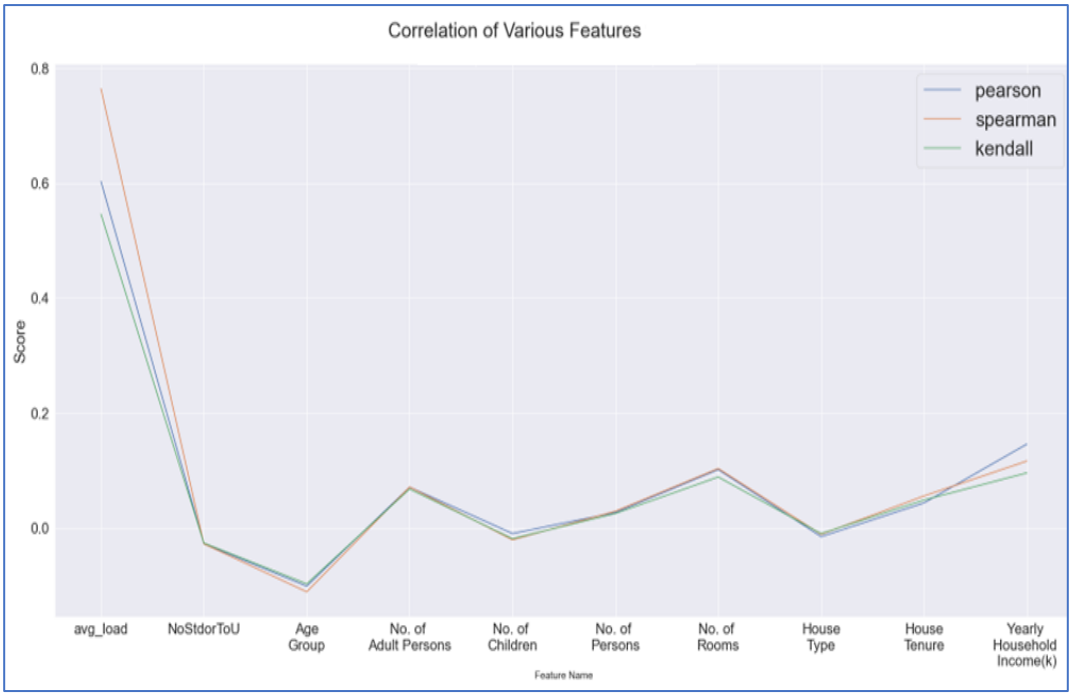

Figure 8 (a) Average Daily Electricity Consumption - Correlation Coefficient of Input Features

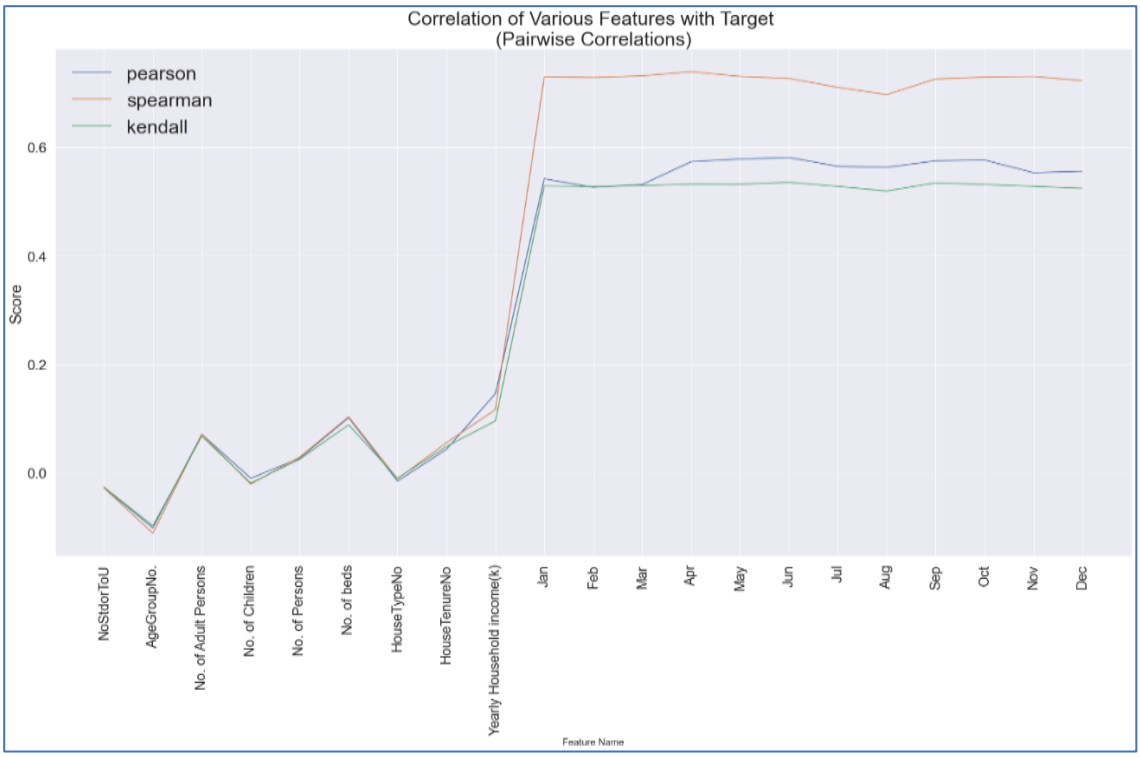

Figure 8 (b) Monthly Electricity Consumption - Correlation Coefficient of Input Features

Correlation Matrix: A correlation matrix can be incredibly useful in displaying the correlation between various input features. This 2D matrix represents the correlation coefficient of pairwise combinations of household features and average daily and monthly consumption. Pearson's correlation coefficient, Spearman's Rank Correlation, and Kendall's tau have been applied to determine the pairwise correlation among the input features, taking into account the daily electricity consumption. These findings have been effectively showcased in Figure 9.

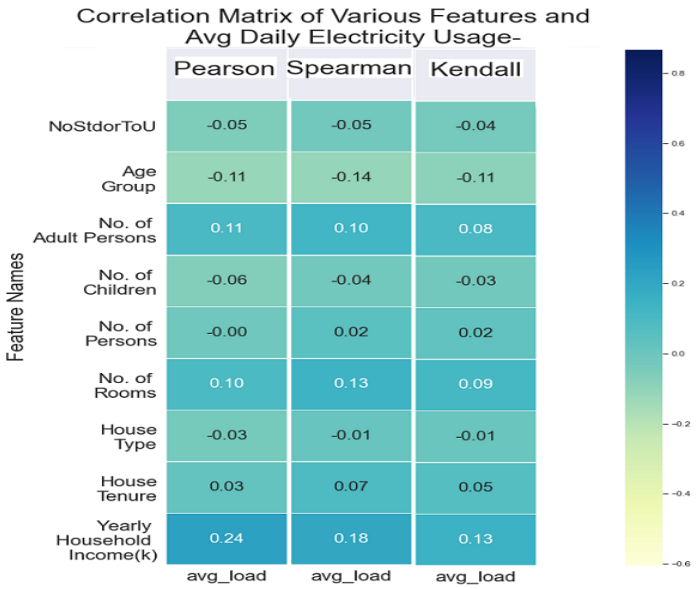

Figure 9 Pair-wise Correlation Coefficient of Input Features

Feature Dependency: Whether a feature is dependent on another feature to work is known as Feature Dependency. Three methods described below were instigated for numerical input features and a categorical target feature, i.e., Cluster Index of DEC_F dataset, to find the higher dependency of the Target Feature on Values of Input Features. Their results are displayed in Figure. 10.

- Feature Correlation with the Target Feature (Cluster Index) using ANOVA Test
- Feature Correlation with the Target Feature (Cluster Index) using Chi-square Test
- Estimate Mutual Information for a discrete Target/Output Feature.

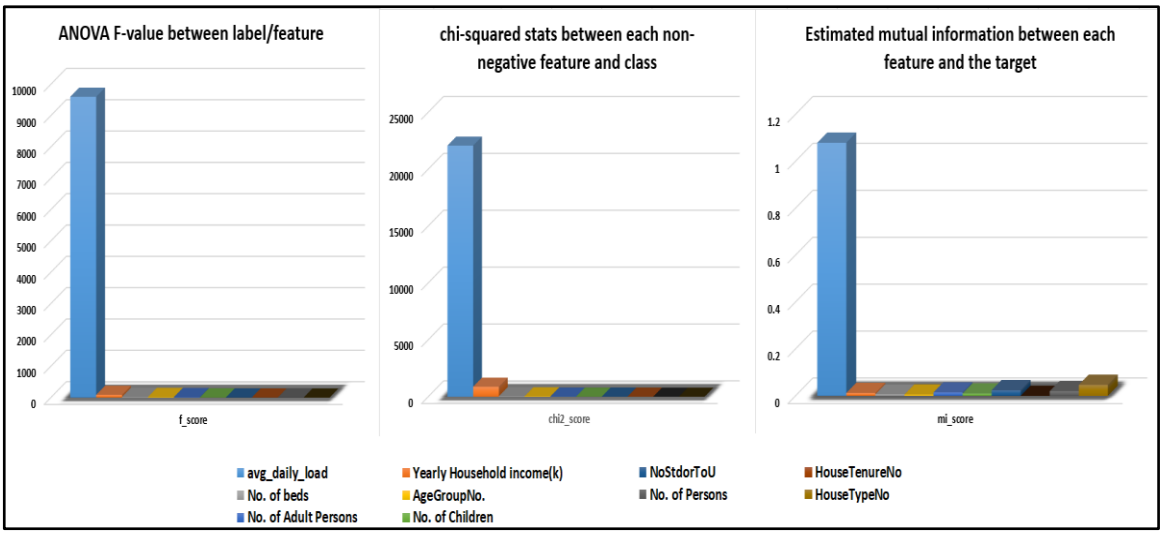

Figure 10 f_score, chi_score and mi_Score of Input Household Features

It has been observed that including the columns of Demographic-Household features does not significantly impact the identification of consumer clusters. Daily consumption alone is sufficient to identify these clusters. This was demonstrated by the SOMKMeans two-level clustering on the D-EC and D-EC with demographic-household features datasets. A significant observation is the existence of Demographic-Household features reflected in Daily Electricity Consumption Data. Observation carried out in this research will be used as an input to the classification model for its training process, and a decrease in the input features to train the classification model will increase the accuracy of predicting the consumption profile, which is the practical outcome concerning earlier study performed by other authors represented in previous research work of this paper. Based on the results presented in Figure 8, it can be seen that the target class has a higher reliance on electricity consumption compared to other input parameters. This is revealed through ANOVA F-test, Chi-squared Test, and Mutual Information, which scored 98%, 95%, and 89%, respectively.

7.3 MLFFNN using HPT Classification Model

A classification model is a machine learning algorithm used to predict the class or category of a new observation based on a set of input variables. Supervised classification algorithms are machine learning algorithms trained using labelled datasets to predict the class or category of new, unseen observations. To train the classification model for D-EC Dataset, six classification algorithms - KNN, SVM, DT, RF, GNB, and MLFFNN have been applied

SOMKmeans_DEC_Clustered$_n$ generated using SOMKMeans Clustering Algorithm shown in below Table 4.

Table 4 Clustered D-EC Dataset

meterID	Cluster Index	2010-1-1	2010-1-2	2010-1-3	2010-1-4
ID5829	0	64.443	66.34	39.081	33.363
ID3575	1	22.017	25.397	23.856	27.611
ID6733	2	340.141	360.5	289.322	376.726
ID9414	3	0	352.001	780.857	778.717
ID4073	4	26.338	26.356	26.542	109.371
ID6979	5	60.214	61.691	63.734	243.245
ID6329	6	728.393	693.443	690.153	452.632
ID5255	7	230.672	313.448	237.42	838.613

To balance the dataset, we applied the Borderline Synthetic Minority Oversampling Technique (SMOTE). This involved oversampling the minority classes by synthesizing new examples based on the Euclidean distance between the random data and its k nearest neighbors along the decision boundary between the two categories.

To evaluate the SOMKmeans_DEC_Clustered dataset, 653 records were set aside as test data while the remaining data was used to construct the model. The training dataset consists of 5800 observations of 365 variables (about 90% of the dataset), while the test data comprises 653 observations of 365 variables (about 10% of the dataset). To develop 6 classification models, it is necessary to create a model for each algorithm and call the fit() and predict() methods using their respective parameters. Once models are developed and evaluated, compare them to see which algorithm works best for the SOMKmeans_DEC_Clustered dataset based on the actual and predicted class of the test dataset.

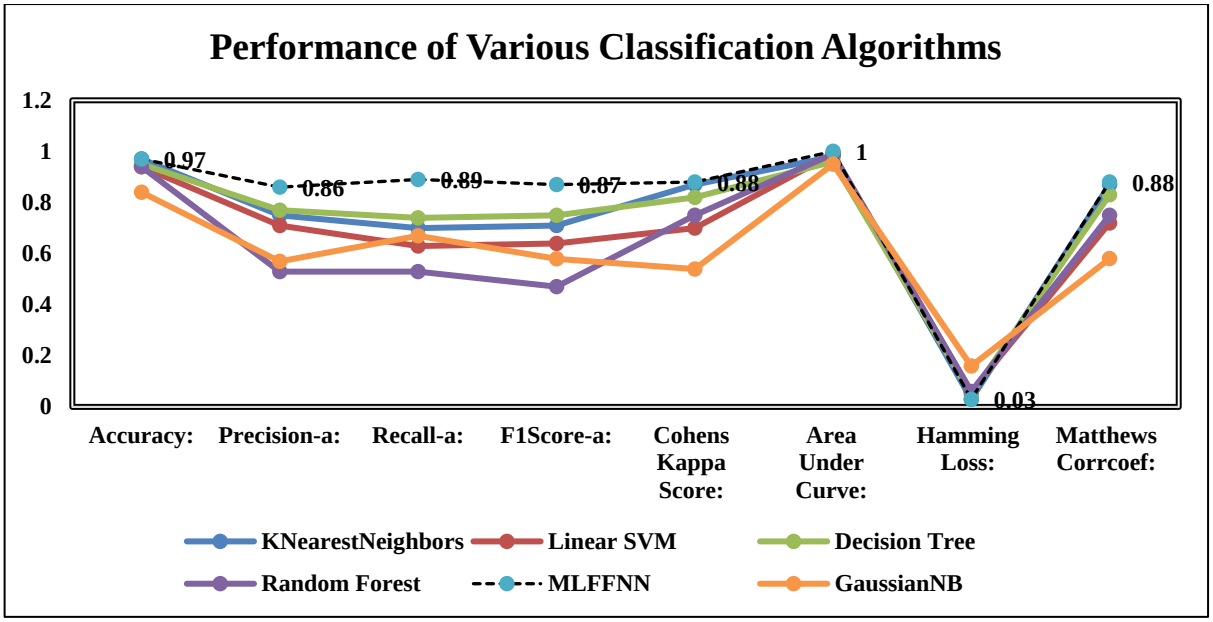

Figure 11 Performance of Classification Algorithms

By analysing the score of each evaluation metric, Multi-Layer Feed-Forward Neural Network (MLFFNN) has been significantly recognized as the best-performed algorithm compared to other classification algorithms. Based on the data from the SOMKmeans_DEC_Clustered$_{te}$ dataset, it has been found that the MLFFNN algorithm in Figure 11 accurately predicts the usage class of consumers with a success rate of around 97%, outperforming other classification algorithms. The MLFFNN also performed well, with an accuracy rate of over 80% for each input test dataset class. The results are represented in Figure 11 by a blue bar with a black dashed line surrounding it. Hyperparameter optimization will help achieve an improved performance of the MLFFNN algorithm.

In order to increase the accuracy of the MLFFNN algorithm, we have utilized the Baseline model and HPT model of MLFFNN on the SOMKmeans_DEC_Clustered dataset. We have plotted the loss and accuracy of both the training and validation data for both models using a line graph shown in Figure 12,13,14 and 15. This allows us to analyse how much the HPT model has exceeded the performance of the Baseline model.

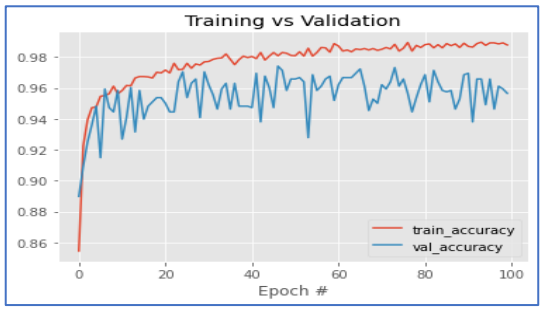

Figure 12 Training and Validation Accuracy of Baseline MLFFNN Model

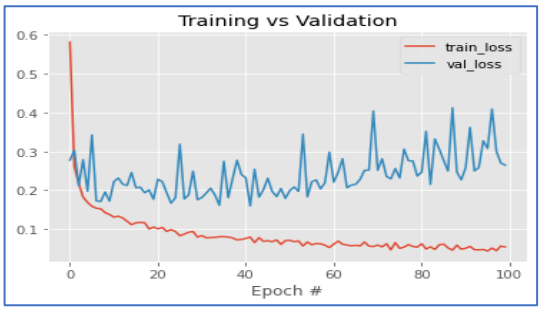

Figure 13 Training and Validation Loss of Baseline MLFFNN Model

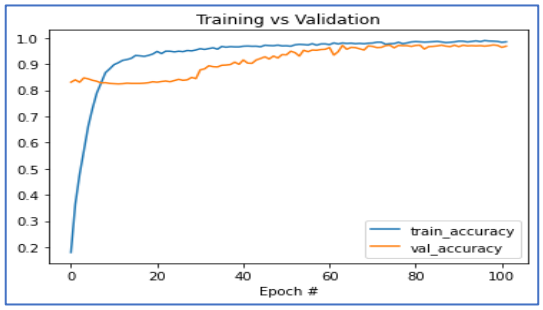

Figure 14 Training and Validation Accuracy of MLFFNN Model using HPT

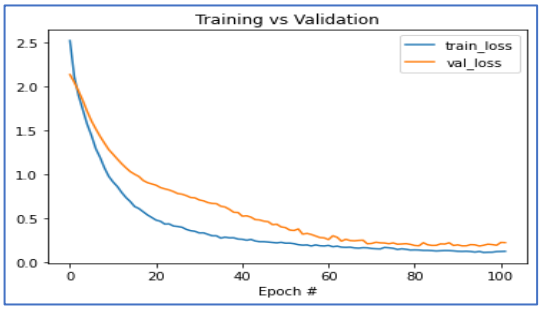

Figure 15 Training and Validation Loss of MLFFNN Model using HPT

The D-EC Dataset's performance has been enhanced through the application of Hyper Parameter Tuning. The results are depicted in Figures 14 and 15, indicating high accuracy in both the training and validation datasets for HPT based MLFFNN Model. At Epoch 100, the accuracy of training and validation data is increasing and stable. Now looking into the training and validation loss results, it is clear that from Epoch 65 to 100, the training and validation loss continuously decrease, and their differences are minimal presented in Figure 15.

While comparing the results of the Baseline MLFFNN Model with the HPT Model of MLFFNN, it has been observed that the accuracy of the MLFFNN Model using HPT has been increased up to 98% and observing the results of other evaluation metrics, i.e., Precision, Recall, F1-Score, CohensKappa Score, hamming Loss, and Mathews Correlation Coefficient, it clearly indicates that Hyper Parameter Tunning for MLFFNN has improved performance of Classification Model for predicting the electricity consumption class also termed as Consumption profile of consumer which is shown in Figure 16.

By observing the class-wise results produced through Evaluation Metrics Precision, Recall, and F1-Score, Hyper Parameter Tuning of the MLFFNN model have improved prediction results for each class which is more than 90%.

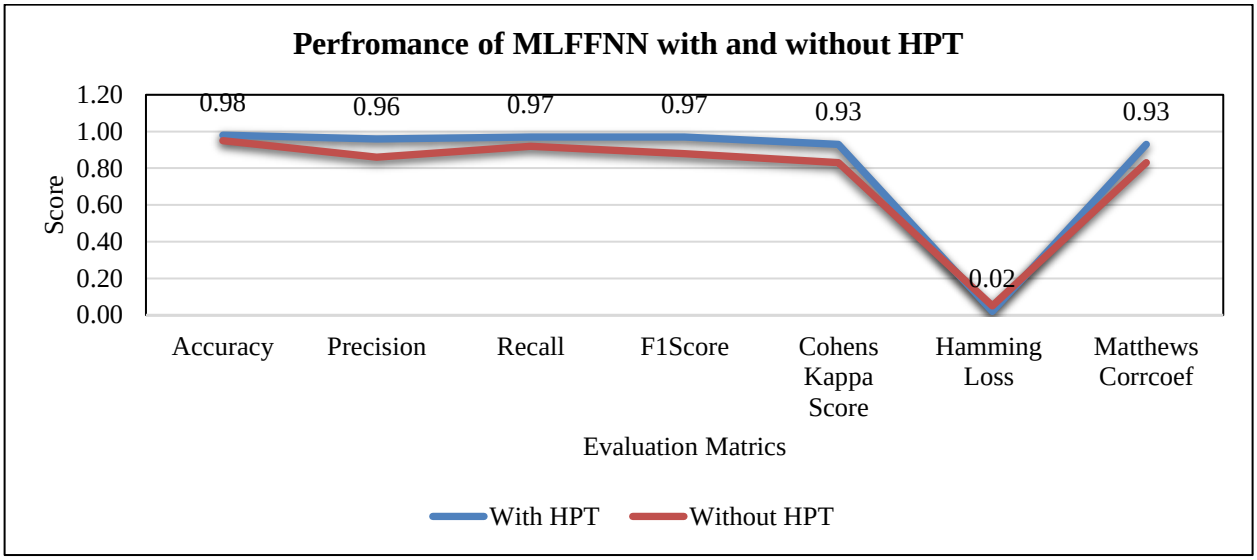

Figure 16 Performance of MLFFNN with and without HPT

8. Achievements concerning objectives

- Effective segmentation of consumers based on daily electricity consumption
- Daily Electricity Consumption has a significant impact on deciding Consumers Profile Class
- Hyperparameter Tuning of MLFFNN Classification Model is efficient up to 98% to predict the Consumer Profile who unknown to the MLFFNN morel
- Borderline Synthetic Minority Oversampling Technique has helped to synthesize observation for imbalanced classes of datasets in the training process of the MLFFNN Model

9. Conclusion

Efficient SOM-based Two-level Clustering is a cutting-edge methodology that utilizes smart meter data from residential consumers to generate Electricity Consumption Patterns (ECPs). Its main objective is to accurately identify the Consumption Profile of individuals by analyzing their Daily Electricity Consumption. The Consumer Pattern analysis using SOMKMeans Clustering Algorithm generated four Daily Electricity Consumer Profiles. Consumers are classified as Low, Moderate, Elevated, and Extravagant Usage based on their daily routines.

According to their Average Daily Electricity (ADE) usage, consumers were classified as consistent or non-consistent in each developed profile. The proposed methodology provides a chance to examine the previously unexplored extent of electricity consumption patterns and profiles using statistical analytics and machine learning techniques, highlighting research directions to advance applications in society and the economy to benefit the community. By adopting smarter energy consumption habits and optimizing appliance usage, people have the ability to substantially lower their energy expenses. In addition, energy providers can establish adaptable tariff structures that are tailored to specific regions, in order to promote consistent patterns of consumption, determine usage trends, and anticipate future energy requirements. Such generated Electricity Consumption Profiles (ECPs) can also help to prepare a model to classify the consumers based on their daily consumption. With the analysis of demographic features, house structure characteristics, and the min-max temperature of a specific day, it is possible to confidently generate a cluster of consumers and their respective classifications solely based on their electricity consumption patterns. Through the utilization of SOMKMeans Clustering Algorithm, Correlation Coefficient, and Feature Dependency, it has been observed that Daily Electricity Consumption holds significant influence in determining consumer profiles and generating Classification-Prediction Models. This is due to various factors such as the household's daily habits, number of rooms, house type and structure, and usage of appliances, all of which are indirectly reflected in the Daily recorded Electricity Consumption. By comparing the results of the different classification algorithms, Multi-Layer Feed Forward Neural Network has significantly outperformed on Daily Electricity Consumption Data as well as Bayesian Optimization as Hyper Parameter Tuning helped to improve the performance of the MLFFNN Classification Model up to 98% to predict the Consumption Profile of Consumers.

10. Copies of papers published and a list of all publications arising from the thesis

- Rinku Chavda, Dr. Sohil Pandya. 2018. “Analyzing Electricity Consumptions Pattern for Profiling and Forecasting-a Review.” International Journal of Innovative Research & Studies 8(11): 236–42.

- Rinku Chavda, Sohil D. Pandya, and Chetan D. Kotwal. 2022. “Electricity Consumption Patterns Using Som-Based Two-Level Clustering of Residential Households.” Indian Journal of Computer Science and Engineering 13(1): 93–107 (Scopus Indexed Journal).
- Rinku Chavda, Sohil D. Pandya. 2023. “Experimenting Sensor-based Effective Energy Saving Module for Household Electricity Consumption”, GU - JET (Journal of Engineering and Technology), ISSN 2249 – 6157 (Peer Reviewed Journal).
- Chavda R, Pandya S, Kotwal C 2023. Influence of Demographic-Household Features on Electricity Consumption to generate Consumption Patterns and Profiles. Indian Journal of Science and Technology 16(35): 2879-2888. https://doi.org/ 10.17485/IJST/v16i35.685 (Web of Science).

11. References

[1]. Abeykoon, Vibhatha, Kankanam Durage Nishadi, Pasika S Ranaweera, and Rajitha Udawapola. 2016. “Electricity Consumption Pattern Detection.” Annual Research Symposium (ARS), Faculty of Engineering, University of Ruhuna. (January).

[2]. Albert, Adrian, and Ram Rajagopal. 2013. “Building Dynamic Thermal Profiles of Energy Consumption for Individuals and Neighborhoods.” Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013: 723–28.

[3]. Ali, Sharif Shofirun Sharif et al. 2021. “Critical Determinants of Household Electricity Consumption in a Rapidly Growing City.” Sustainability (Switzerland) 13(8): 1–20.

[4]. Ali, U, C Buccella, and C Cecati. 2016. “Households Electricity Consumption Analysis with Data Mining Techniques.” In IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, 3966–71.

[5]. Al-Wakeel, Ali, and Jianzhong Wu. 2016. “K-Means Based Cluster Analysis of Residential Smart Meter Measurements.” Energy Procedia 88(June): 754–60. http://dx.doi.org/10.1016/j.egypro.2016.06.066.

[6]. Amiruddin, Brilian Putra, Evanbill Antonio Kore, and Dhiya Aldifa Ulhaq. 2020. “Comparison of Classification Algorithms on Household Electricity Consumption Data.”: 4–7.

[7]. Banihashemi, Saeed, Grace Ding, and Jack Wang. 2017. “Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption.” Energy

Procedia 110: 371–76. https://www.sciencedirect.com/science/article/pii/S1876610217301856.

[8]. Buttitta, Giuseppina, Olivier Neu, Will Turner, and Donal Finn. 2017. "Modelling Household Occupancy Profiles Using Data Mining Clustering Techniques on Time Use Data School of Electric and Electronic Engineering, University College Dublin, Dublin, Ireland School of Mechanical and Materials Engineering, University College." IBPSA Building Simulation 2017.

[9]. Capozzoli, Alfonso, Marco Savino Piscitelli, and Silvio Brandi. 2017. "Mining Typical Load Profiles in Buildings to Support Energy Management in the Smart City Context." Energy Procedia 134: 865–74. https://doi.org/10.1016/j.egypro.2017.09.545.

[10]. Carroll, Paula et al. 2018. "Household Classification Using Smart Meter Data."

[11]. Choi, Hyun Wong, Nawab Muhammad Faseeh Qureshi, and Dong Ryeol Shin. 2019. "Analysis of Electricity Consumption at Home Using K-Means Clustering Algorithm." International Conference on Advanced Communication Technology, ICACT 2019-February (September): 639–43.

[12]. Kim, Young Il, Jong-Min Ko, and Seung-Hwan Choi. 2011. "Methods for Generating TLPs (Typical Load Profiles) for Smart Grid-Based Energy Programs."

[13]. Knezevic, Dragana, and Marija Blagojević. 2019. "Classification of Electricity Consumers Using Artificial Neural Networks." Facta universitatis - series: Electronics and Energetics 32: 529–38.

[14]. Nepal, Bishnu, Motoi Yamaha, Sahashi, and Aya Yokoe. 2019. "Analysis of Building Electricity Use Pattern Using K-Means Clustering Algorithm by Determination of Better Initial Centroids and Number of Clusters." Energies 12: 2451.

[15]. Paatero, Jukka V., and Peter D. Lund. 2006. "A Model for Generating Household Electricity Load Profiles." International Journal of Energy Research 30(5): 273–90.

[16]. Papageorgiou, George, Andreas Efstathiades, Maria Poullou, and Alexander N. Ness. 2020. "Managing Household Electricity Consumption: A Correlational, Regression Analysis." International Journal of Sustainable Energy 0(0): 1–11. https://doi.org/10.1080/14786451.2020.1718675.

[17]. Rajabi, Amin et al. 2019. “A Pattern Recognition Methodology for Analyzing Residential Customers Load Data and Targeting Demand Response Applications.” Energy and Buildings 203: 109455.

[18]. Ridwana, Iffat, Nabil Nassif, and Wonchang Choi. 2020. “Modeling of Building Energy Consumption by Integrating Regression Analysis and Artificial Neural Network with Data Classification.” Buildings 10(11). https://www.mdpi.com/2075-5309/10/11/198.

[19]. Satre-Meloy, Aven, Marina Diakonova, and Philipp Grünewald. 2020. “Cluster Analysis and Prediction of Residential Peak Demand Profiles Using Occupant Activity Data.” Applied Energy 260: 114246. https://www.sciencedirect.com/science/article/pii/S0306261919319336.

[20]. Sena, Boni et al. 2021. “Determinant Factors of Electricity Consumption for a Malaysian Household Based on a Field Survey.” Sustainability (Switzerland) 13(2): 1–31.

[21]. Senthilnathan, Samithambe. 2019. “Usefulness of Correlation Analysis.” SSRN Electronic Journal (July).

[22]. Sulaiman, S. M., P. Aruna Jeyanthy, and D. Devaraj. 2016. “Artificial Neural Network Based Day Ahead Load Forecasting Using Smart Meter Data.” 2016 - Biennial International Conference on Power and Energy Systems: Towards Sustainable Energy, PESTSE 2016: 1–6.

[23]. Syakur, M. A., B. K. Khotimah, E. M.S. Rochman, and B. D. Satoto. 2018. “Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster.” IOP Conference Series: Materials Science and Engineering 336(1).

[24]. Torres, J. F., D. Gutiérrez-Avilés, A. Troncoso, and F. Martínez-Álvarez. 2019. “Random Hyper-Parameter Search-Based Deep Neural Network for Power Consumption Forecasting.” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11506 LNCS: 259–69.

[25]. Viloria, Amelec et al. 2020. “Electrical Consumption Patterns through Machine Learning.” Journal of Physics: Conference Series 1432: 12093.

[26]. Wen, Lulu, Kaile Zhou, and Shanlin Yang. 2019. “A Shape-Based Clustering Method for Pattern Recognition of Residential Electricity Consumption.” Journal of Cleaner Production 212: 475–88.

[27]. Yang, Li, and Abdallah Shami. 2020. “On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice.” Neurocomputing 415: 295–316. https://www.sciencedirect.com/science/article/pii/S0925231220311693.

[28]. Yilmaz, Selin, Jonathan Chambers, and Martin Patel. 2019. “Comparison of Clustering Approaches for Domestic Electricity Load Profile Characterisation - Implications for Demand Side Management.” Energy 180.

[29]. Zhang, Xiaoou Monica, Katarina Grolinger, Miriam A.M. Capretz, and Luke Seewald. 2019. “Forecasting Residential Energy Consumption: Single Household Perspective.” Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018 (December): 110–17.

[30]. Zhang, Zhenhua, and Timnah Zimet. “K-Means Based Clustering Analysis of Household Energy Consumption.”

[31]. Zhou, Kaile, and Shanlin Yang. 2016. “Understanding Household Energy Consumption Behavior: The Contribution of Energy Big Data Analytics.” Renewable and Sustainable Energy Reviews 56: 810–19. http://www.sciencedirect.com/science/article/pii/S1364032115013817.

[32]. Zhou, Kaile, Changhui Yang, and Jianxin Shen. 2017. “Discovering Residential Electricity Consumption Patterns through Smart-Meter Data Mining: A Case Study from China.” Utilities Policy 44: 73–84. http://www.sciencedirect.com/science/article/pii/S0957178717300176.