rauf and Adekoya
*Journal of Electrical Systems and Inf Technol* (2023) 10:43
https://doi.org/10.1186/s43067-023-00104-2

Journal of Electrical Systems and Information Technology

RESEARCH Open Access

# Forecasting household energy consumption based on lifestyle data using hybrid machine learning

seidu agbor abdul rauf[1*] and Adebayo F. Adekoya[1]

*Correspondence:
rauf.seidu.stu@uenr.edu.gh

[1] Department of Computer Science and Informatics, University of Energy and Natural Resource, Sunyani, Ghana

## Abstract

Household lifestyle play a significant role in appliance consumption. The overall effects are that, it can be a determining factor in the healthy functioning of the household appliance or its abnormal functioning. The rapid growth in residential consumption has raised serious concerns toward limited energy resource and high electricity pricing. The propose 134% electricity tariffs adjustment by Electricity Company of Ghana (ECG) at the heat of economic hardships caused by Covid-19 has raised serious public agitation in Ghana (west Africa) . The unpredictable lifestyle of residential consumers in an attempt to attain a comfortable lifestyle and the rippling effects of population growth burdens energy demand at the residential sector. This study attempts to identify the lifestyle factors that have great influence on household energy consumption and predict future consumption of the household with mitigating factors to cushion the effects on high consumption. The study is based on lifestyle data using hybrid machine learning. The hybrid model achieved high accuracy (96%) as compared to previous models. The hybrid model performance was evaluated using mean absolute percentage error (MAPE), root mean square error (RMSE) and correlation coefficient (R) metrics.

**Keywords:** Hybrid-machine learning, Lifestyle data, Energy load forecasting, Artificial intelligence

## Introduction

Electricity is the backbone of every economic growth in a country [6] undoubtedly, everything of mankind depends on it either in direct or indirect form. The energy industry players and policy makers have serious concerns in the amount of electricity generated and consumed at all sectors since it has an adverse effect on development [34].

Globally, as a matter of urgency, policy makers consider efficient use of electrical energy by consumers as priority in order to mitigate high electricity tariff reducing the world poverty rate (millennium development goal)(Selamawit Mussie (AUC) et al., 2015). Energy efficiency is a challenge for a sustainable society [42]. The progressive

growth in population coupled with rural-urban migrations and industrialization creates high demand for energy, putting pressure on generation, [6]6. Study shows that 30% of electricity generated is consumed by residential users [10].

Locally, the residential sector energy consumption stood at 4,487 ktoe, representing 40.5% of the total final energy consumption for 2022 year ending. This figure is however projected to grow by 1.9 yearly according to the report by Energy Commission of Ghana. The commission's report further revealed the national electricity access rate value as 85.33% (Energy [12].

According to the U.S energy outlook report, as discussed by Jones et al and Martin et al in their studies, residential electricity consumption varies significantly depending on weather and lifestyle patterns of the consumer [18], for example, that an average home in the Pacific region (consisting of California, Oregon, and Washington) consumes 35% less energy than homes in the South-Central region. The differences are attributed to climate conditions. Further study by Martin et al details large family size, income and occupations as key factors in residential consumptions [26].

Studies by Hui et al [16], and Zangrando et al [40] have argued that family lifestyle of consumers affects residential consumption. The family lifestyles includes,age, family type, household appliance type possessed, occupation, house type, income, cultural belief, social life and appliance usage type [16, 16]. They concluded that these family lifestyles contributes to higher residential consumptions.

Since consumers' lifestyles are mostly activities-related that are unpredictable and in some cases uncontrollable, it makes it difficult to determine which particular lifestyle of the residential consumer significantly influences electrical consumptions at a given period and even more difficult for residential load forecasting.

Notwithstanding the difficulties stated, identifying the particular lifestyle factor(s) that significantly influences consumption helps in proper household energy management practices and accurate energy demands projections.

Residential load demands projections based on lifestyle data helps in the implementation of energy policies and programs and also helps consumers to know how their future energy consumption is affected by their lifestyle [33, 41].

In [9], authors have argued that the home is the largest single basic unit of electricity consumptions and when controlled, will reduce drastically the amount of electricity consume by the residential sector. Study by Ruan et al revealed that households in Canada consumes 1.4 million Tera-joules of energy and this is estimated to be up by 7.2%. The study stated the residential consumption figure in Canada as 44.6% [31].

There is a need for thorough investigations in to residential consumers' lifestyle to identify the various lifestyle factors that significantly influences residential consumption and to address these factors to curtail the rising consumption at the residential sector. Hermann et al have stated in his recent study that, domestic consumption in the household could be monitored and regulated if efficient energy management systems (EEMS) are installed in the home. The study further argued that residential users need to take an important decision based on their lifestyle preferences [15] to curtail domestic consumption.

Forecasting electrical consumption base on household lifestyle data using machine learning algorithms is a challenging task [29]. However, it lays a strong base for

effective demand response program and provides support in terms of maintenance, automation and timely generation meeting the energy demands of consumers [29].

This current study presents a hybrid machine-learning model to accurately predict household energy consumptions in the home using a lifestyle data gathered through a questionnaire.

The objectives of the study are tailored below:

1. To propose hybrid machine learning model to predict residential consumptions using the household lifestyle data.
2. To identify which particular lifestyle of consumer is highly predictor of consumptions
3. To evaluate the model using accuracy and error metrics. It achieves 4.20% for MAPE with 96.0% accuracy using lifestyle data in a timeline of 40 days, which validates the effectiveness of the proposed approach

Accurate prediction of the mode will enhance decision making by industry players in terms of generation, distributions and consumptions and more importantly it will be beneficial to the end-users in energy savings.

This study is guided by the following research questions:

1. How predictable is the lifestyle data in household electrical consumptions?
2. Which particular lifestyle of the household influences electrical load consumptions?
3. How reliable and accurate in prediction of electrical loads consumption of the model?

The purpose of this study is to investigate the household lifestyle aspects of the residential energy consumption. Lifestyle aspects include but not limited to family patterns, occupations, marital status, and age. The energy consumption is investigated based on the life schedules of each family member. To limit the scope of discussion, Tamale, the capital city of northern Ghana, was selected for the case study.

The remaining sections of the current paper are organized as follows: Section "Literature review" presents a review of the pertinent literature on electricity demand predictions. In section "Methodology and data", we discuss the materials and methods adopted for the current study. Section "Experimental setup" presents the outcome of the study, and discussions. Section "Conclusion" concludes the study and outlines the direction for future studies.

## Literature review

Several studies on residential energy use have been conducted by many researchers in response to increasing energy demands by residential consumers. Articles about trends of energy use and its relationship with household's lifestyle attributes are reviewed in this section.

Studies by Li et al indicated that residential energy consumption is expected to keep rising alongside an increased in household appliance ownership in Japan and across Asia countries [22]. Their survey in eighteen (18) western countries shows that

household appliance energy consumption is heading toward saturations. Household energy consumption trends based on family pattern, aging society and life schedules was carried-out by Luo et al [23]. Family member age and its influence on household electricity consumptions was carried out by Lazzari et al [21]. Their findings were supported by Gonzalez et al who analyzed household energy consumptions in terms of family patterns, employment status, employment sector, gender and age and concluded that lifestyle have a significant effect on household energy consumptions [14].

Chou et al study the changes in household occupant's behaviors in Hangzhou, China and predicted that residential energy consumption will continue to increase in the near future due to comfort living lifestyle and serious dependency on electrical appliances and concluded that there can be a great energy savings at the household if occupants are educated on energy savings measures [7].

Studies by Nti et al presented a monthly electricity demand prediction model using a soft-computing model in Bono region (Ghana) based on historical weather and demand data [28].

Using a multi-layer perceptron (MLP), decision tree (DT), and support vector machine (SVM), the researcher attained an accuracy of 95% for MLP, 67.2 for SVM, and 80.57% for DT. Zhao et al conducted a study in Tokyo using feature selection and multivariate linear regression (MLR) techniques to predict seasonal electricity demand for households based on end-user lifestyle data and concluded that lifestyle data is significant in energy demand forecasting using household factors such as family pattern, age, and building type as independent variables [42]. The study fails to predict the actual energy value.

Nti et al conducted a study in Sunyani (Ghana) from three hundred and fifty (350) household to forecast residential electricity consumption based on lifestyle data using artificial neural networks. The study achieved a good accuracy with RMSE (0.000726) and MAE (0.000976) of the proposed model as against (RMSE = 0.08816 and MAE = 0.06911) for support vector regression (SVR) and (RMSE = 0.0657 and MAE = 0.05714) for decision trees (DT). With household factors such as residential location, age of family head, employment sector of the family head, nature of employment, marital status and among others [29].

The key factors that have a great impact on total household electricity consumptions includes, income level and population size according to Kwac et al [20]. These are the significant factors that can change household energy consumptions. Increasing per capita income is highly correlated with household energy consumptions, however, household electricity consumption has a "U" nonlinear correlation with urbanizations and whiles electricity pricing have great negative impact correlation with influencing on household energy consumptions [35]. Temperature influence on household energy consumptions is dependent on regional locations. Hence, for household energy consumptions, per capita income, temperature and urbanizations all indicates nonlinear correlations on changes in household electricity consumptions [1].

On recent study, the household electricity uses and ownership of electrical gargets in Ireland were analyzed using logit regression (LR) on large micro-dataset by Almahamid et al in Northern Ireland. The study revealed that the usage of space and

water-heating by a household are more crucial than electrical machines in categorizing residential energy usage [2].

Asem et al used Markov model to build a prediction model based on user sensor deployed in four (4) residential homes to predict the presence of the user in the home(Asem [4].

Further studies have advanced in to temporal variables in household energy demand predictions. The variables include calendar readings (hours, days, weeks, months, years, holidays [37] others also used weather parameters (temperature, wind speed, irradiations and humidity, etc) which can be obtained from country meteorological departments for energy demand forecasting [25].

Different modeling techniques have been deployed by researchers to predict electricity load demands including support vector machines (SVM) (Edward et al, Li et al . Zhang et al,), artificial neural networks (ANN) (Nti et al, Liu et al, Edward et al), autoregressive integrated moving average (ARIMA) (Chou et al), clustering algorithms (Wang et al,) and regression models (Malatesta et al), etc.

For industrial and commercial electricity load forecasting, artificial neural networks (ANN) have been largely and extensively deployed, for nonlinear and time series problems, support vector machine is the most used algorithms in solving these problems and perform best in household energy experiments as argued by Meng et al [27]. Table 1 summarizes the articles reviewed

Figure 1 shows the electricity consumption per sectors of the economy for the year 2021 and 2022 sourced from the energy commission of Ghana (ECG). As indicated in Fig. 1, there is a progressive increment in energy consumption at the residential sector. The residential sector accounted for 65% of energy consumption for the year 2022 whiles industrial sectors in Ghana consumed 35% of final energy(Energy [12] (Energy [11].

Figure 2 shows the electricity consumption per region (household) in 2017, 2019 and 2021 using population. In 2017, the national population electricity access rate stood at 84.1% and 85% for 2019 while 85.3% for 2021 (Energy [13]. The regional population access rate can be determined by the formula:

**Table 1** Summary of selected articles reviewed

| Studies | Variables | Method(s) | Location (study) |
|---|---|---|---|
| Nti et al | Socio-economic factors (age, income, family size, etc) | Artificial neural network (ANN) | Ghana |
| zhang et al | Family pattern and aging society | Support vector machine | Japan |
| Edward et al | Life schedules | Artificial neural network | China |
| Grolinger et al | Electric gadgets | logit regression (LR) | Ireland |
| Malatesta et al | Socio-economic factors | ARIMA | China |
| Kwac et al | Occupant behavior | Linear regression | China |
| Chou et al | Occupant behavior | multivariate linear regression (MLR) | Tokyo |
| Alhussein et al | Income, temperature and population size | Regression model | China |
| Li et al | Weather and family size | Markov model | Japan |
| Almahamid et al | Calendar readings | Support vector machine (SVM) | Asia |

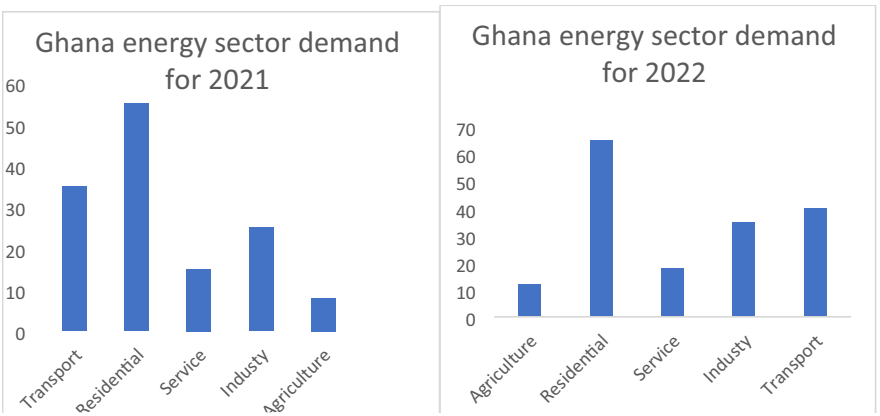


**Fig. 1** Ghana sector electricity demands; Source: energy commission report

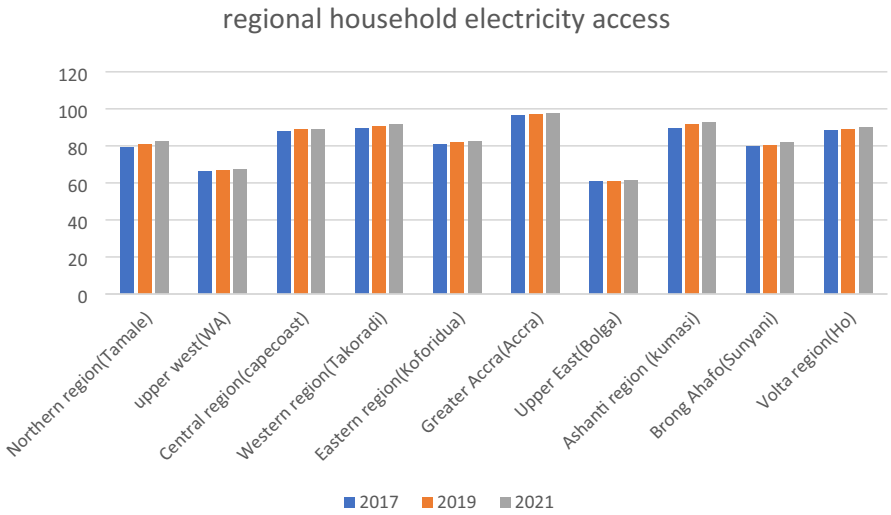


**Fig. 2** Regional household electricity access rate (Ghana)

$$\text{Regional population access rate: } = \frac{\text{total communities connetted to grid}}{\text{total regional population}} \times 100$$

## Methodology and data

This section presents the methods adopted for the implementation of the propose household energy consumption forecasting model based on lifestyle data using hybrid machine learning model.

### Lifestyle data acquisition

A total of 450 households were randomly selected from eight communities within the Tamale metropolitan assembly (Northern Ghana). Sixteen (16) set of questionnaires were developed to obtain the lifestyle data of each participant. It was initial administered

**Table 2** Questionnaire distribution

| Community | Community ID | No. of households | Percentage (%) |
|---|---|---|---|
| Sagnerigu central | SC1 | 55 | 12.2 |
| Sagnerigu west | SW 2 | 67 | 14.9 |
| Choggu east | CE1 | 71 | 15.8 |
| Choggu west | CW2 | 74 | 16.4 |
| Tamale central | TC1 | 43 | 9.6 |
| Tamale west | TW2 | 62 | 13.8 |
| Tamale south | TS3 | 33 | 7.3 |
| Tamale north | TN4 | 45 | 10 |
| Total | | 100 | 100 |

**Table 3** Variables and abbreviations

| Features | Abbreviation |
|---|---|
| Family size | FS |
| Income level | IL |
| employment | EMP |
| Marital status | MS |
| Educational level | EDUL |
| Age | AGE |
| Type of vehicle own | VT |
| Number of rooms | NR |
| Gender type | GT |
| Apartment location | AL |

to 10 household to fine-tune the questionnaire with their comments in Sagnerigu District of the metropolis. The final set of questionnaires were then administered to all selected eight communities with the help of research assistants. Table 2 shows the distribution of the questionnaires to the selected communities.

#### Data preprocessing

Basic steps were carried out on the collected data in order to prepare it for accurate and efficient forecasting. Detecting missed information, data cleaning, noise removal, data filtering and several other basic methods were applied as data pre-processing steps. The best features were selected. Important and common household lifestyles are taken into consideration in the selection of best features. Machine learning methods are then applied on the raw data.

#### Features selections

Redundant or non-informative features are removed using statistical approach from the model. The stepwise linear regression is used as a filter to evaluate the importance of each feature predictors outside the predictive model and subsequently models only the selected features that passes the criterions to increase the accuracy performance

model. All features served as an input of the hybrid model. Table 3 shows the variables and its abbreviation

#### Predictor's influence on consumption evaluations

Impact evaluation on each predictor of the lifestyle features was performed using predictive accuracy measures. Five of such measures used in this current study includes Cross-Validation (CV), Adjusted $R^2$, Akaike's Information Criterion (AIC), Corrected Akaike's Information Criterion ($AIC_c$) and Schwarz's Bayesian Information Criterion (BIC). These predictive accuracy measures can be calculated as follows

CV $= \frac{1}{T}\sum_{t=1}^{T}\left[{}^{e_t}/_{1-h_t}\right]^2$. where $e_t$. is the residual obtained from fitting the model to all T observations

Adjusted $R^2 = \left(1-R^2\right)\frac{T-1}{T-K-1}$. where T is the number of observation, K is the number of predictors. Maximizing Adjusted $R^2$. is good for selecting effective predictors, since it does not tend to err on selecting too many predictors.

A $=$ T $\log\frac{SSE}{T}$ $+$ $+2$(k $+$ 2) where SSE is sum of square error, T is the number of observations and K $+2$ is the number of parameters in the model. The idea is to penalize the fit of the model (SSE) with the number of parameters that need to be estimated. AIC $+\frac{2+(K+2)(K+3)}{T-K-3}$, the AIC tends to select too many predictors for smaller values of observations (T). The $AIC_c$. is a biased-corrected version of the AIC. Unlike the AIC, $AIC_c shouldbeminimized$.

BIC $=$ T $\log\frac{SSE}{T}$ $+$ (k $+$ 2) log T. unlike the AIC, minimizing the BIC is intended to get the best model. The model chosen by the BIC is either the same with the AIC or with fewer terms. Fewer terms means that the BIC have penalized the number of parameters more heavily than the AIC. For larger values of T, leave y-out cross validation is similar to minimizing BIC when y = T$\left\{\frac{1-1}{\log(T)-1}\right\}$

### Support vector machine (SVM) model

The support vector machine is supervised learning algorithm that mostly used in classification related task. Hyper plane is a best set parameter in SVM that classifies a dataset into a number of classes. Hyper plane is selected a straight line if there are only two classes which need classification. Distance from the hyper plane is the key working idea of SVM classifier. Greater the distance of a point from hyper plane, the SVM classifier can best classify a specific point into its corresponding class. In this work, SVM is use for forecasting and also correcting deviations in case of nonlinear. The SVM regression is formulated as follows:

y $=$ w$\varphi$(x) $+$ b. where $\varphi(x)$. is the feature which is nonlinear mapped r p '. 'w' and 'b' are coefficients estimated by minimizing;

$$\mathrm{R(C)} = \mathrm{C}\frac{1}{N}\sum_{i=1}^{N} L_{\varepsilon}\left(d_i,\ y_i\right) + \frac{1}{2}\|w\|^2$$

$$L_{\varepsilon}(\mathrm{d},\ \mathrm{y}) = \begin{cases} |d-y| - \varepsilon & |d-y| \geq \varepsilon \\ 0 & others \end{cases}$$

Where both C and $\varepsilon$. are prescribed parameters. The term $L_{\varepsilon}$. (d, y) is the $\varepsilon$. -intensive loss function, $d_i$. is the actual household consumption data in the ith period. This function indicates that errors below $\varepsilon$. are not penalized. Whiles $\mathrm{C}\frac{1}{N}\sum_{i=1}^{N} L_{\varepsilon}(d_i,\ y_i)$. is the empirical error and $\frac{1}{2}w^2$ measure the flatness of the function. 'C' evaluates the trade-off the empirical risk and the flatness of the model.

The Lagrange function and wolf duality theory with the kernel function k $(x_i, x_j)$ introduced, the function 'R (C)' can be transformed in to quadratic programming problem. The final obtained regression function can be expressed as follows;

$\mathrm{f}\left(x,\ \alpha_i\ \alpha_j^*\right) = \sum_{i=1}^{m}(\alpha_i - \alpha_j^*)\ k(x_i, x_j) + b$, where $\alpha_i$, $\alpha_j^*$ are Lagrange multipliers. The input vector $x_j$ corresponding to nonzero $\alpha_i$, $\alpha_j^*$ is the support vector. Thus, it is clear that the function 'f' totally depends on $'\alpha_i'$, $'\alpha_j^{*'}$. Finally with the given regression function, we can predict the deviation.

#### ARIMA model

The auto-regression integrated moving average (ARIMA) model is a time series analysis technique used to reflects trends [19]. The underlying purposes of ARIMA model are for searching and predictions. In this current study, it is mainly used for forecasting. Box and Jenkins introduced a general model that utilizes the auto-regression model together with the moving average [39].

The autoregression (AR) part is a time series model that assumes the data have an internal autocorrelation (internal structure) which is explored by forecasting methods. Future load in electricity consumption if the electricity load is a linear correlation of previous loads. Then, the AR is formulated as:

$$\mathrm{Y}\prime_{ts} = \sum\nolimits_{i=1}^{m} E_i x_{ts-i} + \beta_t = E_1 x_{ts-1} + E_2 x_{ts-2} + E_m x_{ts-m} + \beta_t$$

Where; $\beta_t$. is the distortion signal, $\mathrm{Y}\prime_{ts}$. is the values of the time series E= $E_1 + E_2 + E_m \ldots$.. Is the product of coefficient vector and m is the integer [39].

The moving average (MA) is a time series used to smoothen the previous history. The logic is that electricity demands observations are close and may be similar in values. $y_{ts-j} = \sum_{j=0}^{n} G_j y_{ts-j} = y_{ts} + G_1\ y_{ts-1} + G_2\ y_{ts-2} + \ G_n\ y_{ts-n}$ where $\mathrm{G} = G_1 + G_2 + G_n$ is the product coefficient vector.

T ARIMA model has three parameters and can be denoted by ARIMA (m, d, n); where the "**m**" notation is the number of auto-regressive arrangement (orders) in the model which determines the current predicted value from the previous values in the series. The **d** notation is the order of differences in the series before estimations of models and "**n**" notation analyzes the value of deviations in the series for previous values before predicting current values [20]. The ARIMA model, the important steps are the order determinations (the value of m and n) and parameter estimations. Currently, the method of order determination is the Akaike Information Criterion (AIC) and the methods of parameter

estimations are moment estimation, least-square estimations, maximum likelihood estimations [30], etc

The ARIMA configuration is given as:

$$Y\prime_{ts} = \sum_{i=1}^{m} E_i F^d x_{ts-i} + \sum_{j=0}^{n} G_j y_{ts-j}$$

Where: $F = 1 - Z^{-1}$. and $E_m$ (z) is stationary and $E_i,\ G_j,\ x_{ts},\ y_{ts}$ are the operators.

### ARIMA-SVM hybrid development

ARIMA and SVM models have their individual capability to capture data characteristics either in linear or nonlinear domains. This current study utilizes the individual strengths of both the ARIMA and SVM models to form the hybrid model (Fig. 3). Thus, both linear and nonlinear patterns are considered in the hybrid model to give an improved overall prediction performance. The hybrid model ($Z_t$) can be represented as;

$Z_t = Y_t + N_t$ where $Y_t$ is the linear part of the hybrid model and $N_t$ is the nonlinear part. Both $N_t$ and $Y_t$ are estimated from the dataset.

The residual ($\in_t$) at time 't' obtained from the ARIMA can be represented by;

$\in_t = Z_t - Y_t^{'}$ where $Y_t^{'}$ is the forecast value of the ARIMA model at time 't'. The residual modeled by the SVM can be represented by;

$\in_t = Z_t - Y_t^{'} + \Delta t$ where f is the nonlinear function modeled by the SVM and $\Delta t$ is the random error. The combined forecast is given by;

$$\check{Z}_t = \tilde{N}_t + \acute{Y}_t$$

### Error estimation

The current study employs Percentage Error estimation method for detecting outliers as shown in Fig. 3. With a tolerance level of 10% as a condition, any percentage error greater than the tolerance value is considered as an outlier. However, tolerance is subjected to change depending on set criteria and the type of data.

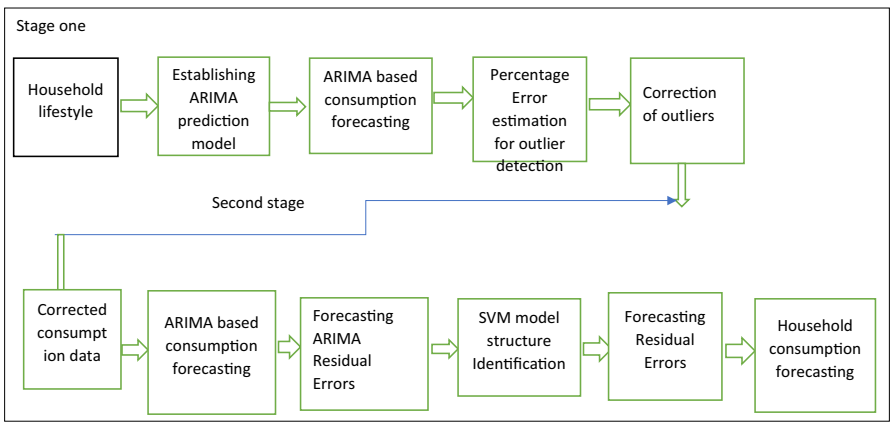


**Fig. 3** Propose hybrid model flowchart

### Evaluation metrics

The error metrics discussed in [8] are used to evaluate our model,

1. Root Mean Square Error (RMSE): It estimates the residual between the actual and the predicted values. A smaller RMSE value indicates a better performance model. While an RMSE value equal to zero indicates a perfect fit of the model. This is determined using the formula below

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\left(\sum_{v=1}^{m} t_v - y_v\right)}$$

2. The Mean Absolute Percentage Error (MAPE) discussed in [8] is also used to measure the performance of the proposed model. It is an aggregative indicator commonly used in power systems. It is mostly used to evaluate the forecasting performance of the whole predicting process comprehensively. It indicates an average of the absolute percentage error, however the lower the value of the MAPE, the better the performance of the model. This is determined using the formula below

$$\mathrm{M} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{t_v - y_v}{t_v}\right|$$

Where $t_v$ is the actual value, $y_v$ is the forecast value, M is the mean absolute percentage error and the n is the number of times the summation iteration happens.

3. The correlation coefficient (R): This criterion reveals the strength of relationship between actual values and predicted values. The correlation coefficient has a range from 0 to 1, where higher R means it has an excellent performance measure.

$\mathrm{R} = \frac{\sum_{v=1}^{m}(t_v - t^-)(y_v - y^-)}{\sqrt{\sum_{v=1}^{m}(t_v - t^-)^2 \sum_{v=1}^{m}(y_v - y^-)^2}}$ where $t^- = \frac{1}{m}\sum_{v=1}^{m} t_v$ and $y^- = \sum_{v=1}^{m} y_v$ are the average values of $t_v$ and $y_v$ respectively and $t_v$ is the actual value, $y_v$ is the predicted value produced by the model, and m is the total number of observations.

## Experimental setup

The obtained features out of the administered questionnaires were abbreviated as indicated in Table 3. The household lifestyle feature serves as an input parameter whiles the actual monthly consumption of participants from the Northern Electricity Company of Ghana (local supply authority) serves as output target. The qualitative response from the respondents are coded using the dummy variables. After which all the 450 received responses were queued with Microsoft excel in to comma-separates values (CSV) file format. The implementation of the proposed model was carried in Shinny App. Shiny app is a package in R programming language that makes it easy to build an interactive web application straight from R programming framework. The ARIMA model is implemented to forecast the household consumption as follows. Firstly, the household lifestyle is transferred in to a stationary time series by periodic difference transformation and a first-order difference transformation. Secondly, it is confirmed as ARIMA (2, 6, 3) through order determination and the values of parameters are obtained through parameter estimation. Finally, we use the confirmed ARIMA model to forecast the household consumption. The methods are repeated to obtain the deviation and forecasting data.

To improve the prediction accuracy, the SVM model is use to extract the sensitive component of the deviation. The deviation sample is use to train the SVM model with exponential kernel function in order to correct the deviation. Far-smallness and near bigness theory and similarity theory are used to construct the input sample, since SVM parameters have serious effect on prediction accuracy, the appropriate values determined by test are C= 22, $\epsilon = 0.18$ and $\sigma = 1$. This give the SVM high generalization. Hence we are able to predict the household consumption and the deviations. The overall forecast includes both the forecast for household consumption and the deviations forecast. Figure 3 shows the flowchart of the implemented hybrid model. The outcome of the proposed hybrid model was benchmarked with previous study.

## Results and analysis

This section presents the results and analysis of the propose ARIMA-SVM prediction hybrid framework for forecasting electricity consumption based on household users lifestyle characteristics.

### *Features selection analysis*

Features with significant weight of more than 0.5 are selected whiles features with weight less than 0.5 are considered unimportant and rejected. The features ranking index in Fig. 4 reveals that the family size of the household highly determine its electrical energy consumption. This affirm the study in [7] that reveals a direct relationship between family size and electrical energy consumption. The family size includes the nuclear family (children and parents), the extended family and domestic staffs in the households. Followed by income level, vehicle type, educational level, age, employment and marital status. Features such as apartment location, number of rooms and gender type have low correlation with energy consumption. They are treated as unimportant features and subsequently rejected for further analysis.

The results affirms study by [17, 39] that age of a household is a significant factor of the household electricity consumption. The results reveals a link between the type of vehicle owned and income level. This can be argued as the monthly income of the household determines the type of vehicle owned and also the number of electrical appliances the household uses. In simple terms, the socio-economic status of a household and

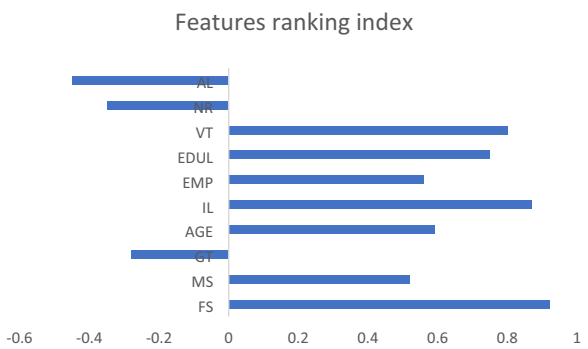


**Fig. 4** Features ranking index

the country at large determines the amount of electrical energy consumed. The results affirm the findings reported in [21, 21], it however disagrees with the findings in [8] that reported that there is no correlation between electrical energy consumption and household income level.

This current study reveals the family size as the top-highest significant feature. This can be interpreted that government agencies responsible for energy sector can project electrical energy demands based on residential population and again adequate measures, policies and programs can be designed to mitigate the growing demands of residential energy consumption.

***Predictor's performance evaluation on household consumption***

From the features ranking index in Fig. 4, the seven (7) selected features are further analyzed using multiple predictive accuracy measures to assess the individual performance on household electrical energy consumption. With the seven selected predictors, there are 128 possible models ( $2^7 = 128$). All the 128 models are fitted and the results are summarize in Table 4. A "1" indicates that the predictor was included in the model and a "0" means that the predictor was not included in the model.

The first row (Table 4) shows the measures of predictive accuracy for a model including all seven predictors. Applying the actual household consumption, the CV, BIC, AIC, AICc, and adjusted $R^2$ are calculated using the CV function.

CV (fit.comsMR)

#> CV AIC AICc BIC Adj $R^2$

#> 0.1163 −409.2980 −408.8314 −389.9114 0.7486

**Table 4** Summaries of predictors' performances on household consumption.

| EDU | Age | MS | FS | VT | IL | EMP | CV | AIC | AICc | BIC | ADJR2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.116 | -409.3 | -408.8 | -389.9 | 0.749 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.116 | -408.1 | -407.8 | -391.9 | 0.746 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0.118 | -407.5 | -407.1 | -391.3 | 0.745 |
| 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.129 | -388.7 | -388.5 | -375.8 | 0.716 |
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0.278 | -243.2 | -242.8 | -227 | 0.386 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.283 | -237.9 | -237.7 | -225 | 0.365 |
| 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.289 | -236.1 | -235.9 | -223.2 | 0.359 |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.293 | -234.4 | -234 | -218.2 | 0.356 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0.300 | -228.9 | -228.7 | -216 | 0.334 |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0.303 | -226.3 | -226.1 | -213.4 | 0.324 |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.306 | -224.6 | -224.4 | -211.7 | 0.318 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.314 | -219.6 | -219.5 | -209.9 | 0.296 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.314 | -217.7 | -217.5 | -208 | 0.288 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.372 | -185.4 | -185.3 | -175.7 | 0.154 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.414 | -164.1 | -164 | -154.4 | 0.092 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.432 | -155.1 | -155 | -148.6 | 0.062 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0.447 | -147.3 | -147 | -139.2 | 0.054 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0.455 | -139.1 | -139 | -127.1 | 0.049 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.485 | -125.2 | -124.9 | -110.6 | 0 |

These values are compared against the corresponding values from other models. For the CV, BIC, AIC and AICc measures, model with lowest value is selected whiles adjusted $R^2$ the model with highest value is selected. The results have been sorted according to the AICc values. The best model contains all the predictors. However, a closer look at the results reveals the individual strengths of the predictors. The family size (FS) had the highest significant impact on household consumption followed by marital status (MS). The results further reveals that the Employment and Education predictors are highly (negatively) correlated. This can mean that most of the predictive information in Employment is also contained in the Education variable. Also the first two rows have almost identical values of CV, BIC, AIC and AICc. So the Employment variable could possibly be dropped and get similar forecasts.

#### *Result and comparison*

The outcome of the hybrid model is compared with the results of Vinagre et al, Mahia et al, Yu et al, Zogaan et al and Atalay et al. for Vinagre et al. [36], the study employed support vector machine (SVM) to forecast residential consumption. The SVM was implemented on different framework (R and MATLAB). The best result achieved with the SVM presents an average error of 6.6% when implemented in R, it however raised to 7.0% in MATLAB software. They concluded that both frameworks can accurately predict residential consumption but the MATLAB is less consistent as compared to R. This current study used the R programming framework to implement the hybrid model. Mahia et al. [24] study employed ARIMA model with three set of parameters ( ARIMA (1,1,1), ARIMA(1,1,2) and ARIMA (1,1,7) ) to forecast residential consumption on two different dataset. Their experimental results shows that ARIMA (1, 1, 1) had high precision and stable predictions on both datasets. Yu et al. (Yu et al., n.d.) study implemented a Decision Tree (DT) to predict household consumption. A total of 100 trees was generated with a data sample of 247. This was implemented in MATLAB framework achieving an accuracy of 91%. Zogaan et al. [43] used random forest with one iteration with backward elimination in R package statistical model to forecast residential consumption achieving good accuracy of prediction. Atalay et al. [5] implemented ARIMA–RF model. Comparing the results between our proposed method and previously reported methods shows that our proposed model have outperformed the previous models using the $R^2$, RMSE and MAPE as summarized in Table 5.

**Table 5** Comparison of the propose hybrid model result with previous models results

| Model | RMSE | MAPE (%) | $R^2$ |
|---|---|---|---|
| DT | 40.23 | 6.12 | 0.5766 |
| RF | 38.94 | 5.40 | 0.5841 |
| SVM | 38.77 | 4.87 | 0.5991 |
| ARIMA | 43.49 | 5.16 | 0.5988 |
| ARIMA-RF | 37.63 | 4.75 | 0.6967 |
| ARIMA-SVM (propose work) | 32.75 | 4.20 | 0.7385 |

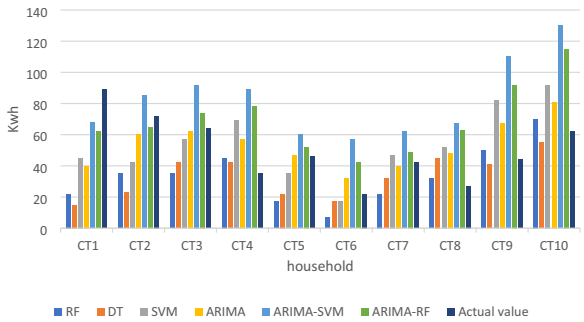


**Fig. 5** Comparison of models performance and actual household consumption

**Table 6** Comparison of ARIMA-SVM model on outlier detection

| s/n | Method | MAPE (%) |
|---|---|---|
| Phase-2 | ARIMA-SVM (with outlier detection) | 4.20 |
| Phase-1 | ARIMA-SVM (without outlier detection) | 6.15 |

It is clear from Table 5, there are improvement in the evaluation metrics using the hybrid model. The MAPE has reduced 4.20% with an improvement in correlation coefficient 0.7385 and RMSE value 32.75.

#### *Models performance and actual household consumption value comparison*

Again, comparing the predicted values of each model against the actual consumption of the household (Fig. 5) shows that our proposed ARIMA-SVM hybrid model outperformed the ARIMA-RF, DT, RF, SVM and ARIMA models. The result can be interpreted that, the residential users can effectively manage their electrical consumption based on the lifestyle they choose to live and also offer them a better opportunity to properly plan their budgets to meet their energy demands to prevent any unforeseen shortage of electricity.

#### *Impact of outlier detection on ARIMA-SVM mode*

The performance of ARIMA-SVM hybrid model was evaluated on two-phase. Phase-1; ARIMA-SVM without outlier detection and Phase-2; ARIMA-SVM with outlier detection. In phase-2, the percentage error outlier detection method was employed with tolerance value of 10%. The outlier detected using the percentage error method will give the least MAPE value of 5.4%. The SVM is trained with the sample with exponential kernel function to correct the deviations. The performance of phase-1 is compared with phase-2 to evaluate the impact of outlier detection in using the percentage error method on ARIMA-SVM model. Table 6 summarizes the result of the ARIMA-SVM performance with or without the outlier detection method.

It is clear from Table 6 that the overall forecasting ability has improved. This can be interpreted that the ARIMA and SVM has mutually supplemented each other with their individual advantages in the hybrid model.

## Conclusion

In this study, we conclude that the improvement of electricity management system from demand side can be achieved efficiently and effectively by using a combination of a novel data mining techniques such as ARIMA and SVM model if household lifestyle data is relied -on despite having a conventional ordinary electricity grid using electricity prepaid meters. The experimental setup with 450 household data randomly selected from eight communities within the Tamale metropolitan assembly (Northern Ghana) revealed the family size as the most influencing predictor of household electrical consumption. The model's accuracy (96.0%) as compared to previous models indicates that the proposed hybrid model can effectively predict household electrical consumption. The use of actual household consumption (Kwh) as target variable in the current study makes it independent on the influence of socio-economic factors since the actual unit of household electrical consumption depend on household energy management lifestyles.

One of the most popular method used for forecasting is the ARIMA technique. ARIMA method will provide a better forecasting accuracy as it requires historical load of the household and fewer assumptions but however, the household consumption is influence by other factors such as the lifestyles of the household. The artificial intelligence techniques will incorporate these factors which can improve the accuracy further. In this current study, hybrid methodology is employed using the ARIMA-SVM. The ARIMA is used to predict the household consumption based on the household lifestyle characteristics and then the SVM is used to improve the accuracy by correcting outliers. In this study, the percentage error method is used to detect the outliers in the household lifestyle data. Through the experimental setup, the result shows that the hybrid model is far much better than the individual ARIMA and SVM models in standalone implementation. Even though, the hybrid model had achieve good accuracy measure, we believe it can be improved by adding other parameters such as household appliance and technology parameters. This is necessary since appliance and technology features have an impact in household consumption.

**Acknowledgement**
We acknowledge all article being reference to during our research work and to God (Allah) almighty.

**Author contributions**
Questionnaire development AFA, Data gathering and survey was carried-out by SAAR, Data preprocessing and analysis was performed by AFA, Quantitative and qualitative analyzes was performed by AFA, Manuscript development was performed by SAAR, AFA revise the manuscript writing, Results analysis and interpretation was supervised by AFA, Algorithm implementation and analysis was performed by SAAR, Performance evaluation by AFA, All authors read and accepted the final manuscripts and choice of journal
This manuscript is coming to you without an associated data and material

**Funding**
The article has received no funding from any individual or organizations or institutions.

## Declarations

**Competiting interests**
This article has no competing interest and has no association with any institutions or organization it is pure academic research knowledge based

Received: 6 April 2023 Accepted: 29 June 2023
Published online: 19 September 2023

**References**

1. Alhussein M, Aurangzeb K, Member S (2020) Hybrid CNN-LSTM Model for short-term individual household load forecasting. 8. https://doi.org/10.1109/ACCESS.2020.3028281
2. Almahamid F, Grolinger K (2022) Agglomerative Hierarchical Clustering with Dynamic Time Warping for Household Load Curve Clustering.
3. Alqasim AR (2022) Using regression analysis for predicting energy consumption in dubai police by a capstone submitted in partial fulfilment of the requirements for.
4. Alzoubi A (2022) Machine learning for intelligent energy consumption in smart homes. Int J Comput Inform Manufact IJCIM 2(1):62–75. https://doi.org/10.54489/ijcim.v2i1.75
5. Atalay V (2023). POWER CONSUMPTION FORECASTING BY HYBRID Serkan Ozen. 42, 126–156. https://doi.org/10.31577/cai
6. Branco MP, Geukes SH, Baidillah MR, Takei M, Aron M, Lilienkamp T (2020). Prediction model of household appliance energy consumption based on machine learning Prediction model of household appliance energy consumption based on machine learning. https://doi.org/10.1088/1742-6596/1453/1/012064
7. Chou J, Tran D (2018) Forecasting energy consumption time series using machine learning. Energy. https://doi.org/10.1016/j.energy.2018.09.144
8. Dong B, Dong B, Li Z, Rahman SMM, Vega R (2015) A hybrid model approach for forecasting future residential electricity consumption a hybrid model approach for forecasting future residential electricity consumption. Energy & Build 117(September):341–351. https://doi.org/10.1016/j.enbuild.2015.09.033
9. Edwards RE, New J, Parker LE (2012) case study. Energy & Build. https://doi.org/10.1016/j.enbuild.2012.03.010
10. EIA (2020) Annual Energy Outlook 2021. 1–81. www.eia.gov/aeo
11. Energy Commission-Ghana. (2021). 2021 ENERGY OUTLOOK FOR GHANA, Demand and Supply Outlook (Issue April).
12. Energy Commission (2022a). *2022a* ENERGY OUTLOOK FOR GHANA ADDRESS Ghana Airways Avenue Airport Residential Area (behind Alliance Francaise) Private Mail Bag Ministries Post Office Demand and Supply Outlook (Issue April). www.energycom.gov.gh
13. Energy Commission (2022b) *2022b* National Energy Statistics (Issue April).
14. Gonzalez D, Patricio MA, Berlanga A, Molina JM (2022) Variational autoencoders for anomaly detection in the behaviour of the elderly using electricity consumption data. Exp Syst 39(4):1–12. https://doi.org/10.1111/exsy.12744
15. Herrmann MR, Costanza E, Brumby DP, Harries T, Brightwell G, Ramchurn S, Jennings NR (2021) Exploring domestic energy consumption feedback through interactive annotation. Energy Efficiency. https://doi.org/10.1007/s12053-021-09999-0
16. Hui M, Lee L, Ser YC, Selvachandran G, Thong PH, Cuong L, Son LH, Tuan NT, Gerogiannis VC (2022) A Comparative Study of Forecasting Electricity Consumption Using Machine Learning Models.
17. Hussein R (2022) Household energy consumption prediction using the stationary wavelet transform and transformers. IEEE Access 10:5171–5183. https://doi.org/10.1109/ACCESS.2022.3140818
18. Jones RV, Fuertes A, Lomas KJ (2015) The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. Renew Sustain Energy Rev 43:901–917. https://doi.org/10.1016/j.rser.2014.11.084
19. Kaytez F (2020) A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption. Energy 197:117200. https://doi.org/10.1016/j.energy.2020.117200
20. Kwac J, Member S, Flora J, Rajagopal R (2016) Lifestyle segmentation based on energy consumption data. 3053, 1–9. https://doi.org/10.1109/TSG.2016.2611600
21. Lazzari F, Mor G, Cipriano J, Gabaldon E, Grillone B, Chemisana D, Solsona F (2022) User behaviour models to forecast electricity consumption of residential customers based on smart metering data. Energy Rep 8:3680–3691. https://doi.org/10.1016/j.egyr.2022.02.260
22. Li Y, Pizer WA, Wu L (2019) Climate change and residential electricity consumption in the Yangtze River Delta, China. Proceed Acad Sci United States of Am 116(2):472–477. https://doi.org/10.1073/pnas.1804667115
23. Luo Q, Wen G, Zhang L, Zhan M (2020) An efficient algorithm combining spectral clustering with feature selection. Neural Process Lett. https://doi.org/10.1007/s11063-020-10297-6
24. Mahia F, Dey AR, Masud A, Mahmud MS (2019) Forecasting Electricity Consumption using ARIMA Model. 0, 24–25.
25. Malatesta T, Breadsell JK (2022) Identifying Home System of Practices for Energy Use with K-Means Clustering Techniques.
26. Martin L (2022) Annual energy outlook 2022 presentation to electricity advisory committee.
27. Meng Z, Sun H, Wang X (2022) Forecasting energy consumption based on SVR and markov model: a case study of China. Front Environ Sci 10:1–15.
28. Nti IK, Resources N, Adekoya AF, Resources N, Nyarko-boateng O (2020a) FORECASTING ELECTRICITY CONSUMPTION OF RESIDENTIAL USERS BASED FORECASTING ELECTRICITY CONSUMPTION OF RESIDENTIAL USERS BASED ON LIFESTYLE DATA USING ARTIFICIAL NEURAL NETWORKS. January. https://doi.org/10.21917/ijsc.2020.0300
29. Nti IK, Teimeh M, Adekoya AF, Nyarko-boateng O (2020) Forecasting electricity consumption of residential users based on lifestyle data using artificial neural networks. ICTACT J Soft Comput 10:2107–2116
30. Rashid M, Hamid A, Parah SA (2019) Analysis of streaming data using big data and hybrid machine learning. https://doi.org/10.1007/978-3-030-15887-3
31. Ruan Y, Wang G, Meng H, Qian F (2022) A hybrid model for power consumption forecasting using VMD-based the long short-term memory neural network. 9, 1–16. https://doi.org/10.3389/fenrg.2021.772508

32. Selamawit Mussie (AUC), Habaasa Gilbert (ECA/AUC), J. B., (AUC), Nougbodohoue Samson Bel-Aube (AUC), Mama Keita (ECA), Aissatou Gueye (ECA), D., Kellecioglu (ECA), Seung Jin Baek (ECA), J., Ameso (ECA), Maimouna Hama Natama (ECA), Stanley Kamara (UNDP), El Hadji Fall (UNDP), S., Berhane (UNDP) and James Neuhaus (UNDP), with technical inputs from Yemesrach Workie, (UNDP), Glenda Gallardo Zelaya (UNDP), F., Leigh (UNDP), Frederick Mugisha (UNDP), W., Reeves (UNDP), Fitsum G. Abraha (UNDP), J., Wakiaga (UNDP), Rogers Dhliwayo (UNDP), A., Bandara (UNDP), Becaye Diarra (UNDP), Celestin Tsassa (UNDP), G. M., Camara (UNDP), A. Mb. (UNDP) and, & Khady Ba Faye (UNDP). (2015). Assessing Progress in Africa Toward the Millennium Development Goals. In *Economic Commission for Africa*. 26 July 2015
33. Shaikh AK, Nazir A, Khan I, Shah AS (2022) Short term energy consumption forecasting using neural basis expansion analysis for interpretable time series. Scientific Reports, 1–18. https://doi.org/10.1038/s41598-022-26499-y
34. Sravani S, Naidu DS, Rohith V, Vardhan V (2021) *PREDICTION OF ELECTRICITY POWER* CONSUMPTION *USING MACHINE LEARNING APPROACH*. *03*, 1656–1662.
35. Thorve S, Baek YY, Swarup S, Mortveit H (2023) High resolution synthetic residential energy use profiles for the United States. 1–23. https://doi.org/10.1038/s41597-022-01914-1
36. Vinagre E, Pinto T, Ramos S, Vale Z, Corchado JM (2016) Electrical energy consumption forecast using support vector machines. 171–175. https://doi.org/10.1109/DEXA.2016.34
37. Wei Z, Wang H (2021). Characterizing residential load patterns by household demographic and socioeconomic factors. In: e-Energy 2021 - Proceedings of the 2021 12th ACM International Conference on Future Energy Systems (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3447555.3464867
38. Yu Z, Haghighat F, Fung BCM, Yoshino H (n.d.). *A decision tree method for building energy demand modeling*.
39. Yuan C, Liu S, Fang Z (2016) Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model. Energy 100:384–390. https://doi.org/10.1016/j.energy.2016.02.001
40. Zangrando N, Fraternali P, Petri M, Oreste N, Vago P, Luis S, González H (2022) Anomaly detection in quasi - periodic energy consumption data series: a comparison of algorithms. Energy Inform 5(4):1–25.
41. Zhang J, Zhang H, Ding S, Zhang X (2021) Power consumption predicting and anomaly detection based on transformer and K-means. 9, 1–8. https://doi.org/10.3389/fenrg.2021.779587
42. Zhao Q, Li H, Wang X, Pu T, Wang J (2019) Analysis of users' electricity consumption behavior based on ensemble clustering. Glob Energy Interconnect 2(6):479–488. https://doi.org/10.1016/j.gloei.2020.01.001
43. Zogaan WA (2022) Power consumption prediction using random. Forest Model 7(5):329–341


## Publisher's Note