CAC-WOA: Context Aware Clustering with Whale Optimization Algorithm for Knowledge Discovery from Multidimensional Space in Electricity Application

Prashant Ahire (prasshantahire@gmail.com)
Dr. D.Y. Patil Institute Of Technology Pune
Pramod Patil
Dr. D.Y. Patil Institute Of Technology Pune

Research Article

Keywords: Artificial neural network, bio-inspired optimization, context aware clustering, knowledge discovery, electricity monitoring, whale optimization

Posted Date: August 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1907759/v1

CAC-WOA: Context Aware Clustering with Whale Optimization Algorithm for Knowledge Discovery from Multidimensional Space in Electricity Application

Prashant G Ahire[1]
Research Scholar
Dr. D. Y. Patil Institute of Technology
Pune, India
prasshantahire@gmail.com

Dr Pramod D Pati[2]
Professor
Dr. D. Y. Patil Institute of Technology
Pune, India
pdpatiljune@gmail.com

Abstract— **Due to the widespread use of smart metering infrastructure, multidimensional data on home electric consumption is easily available for studying its dynamics at finely resolved geographical and temporal scales. Effective forecasting and analysis of electric consumption are crucial for customer participation in time-of-use tariffs, critical peak pricing, and user-specific demand response programs derived from multidimensional data streams. Along with the enormous economic and sustainability ramifications, such as energy waste and the decarbonisation of the energy industry, precise consumption forecasts enable power system planning and reliable grid operations. Energy consumption forecasting is a hot field of research; despite the number of developed models, projecting electric consumption in residential buildings remains problematic owing to the significant unpredictability of occupant energy use behaviour. Discovering the electricity consumption knowledge from the Multi-Dimensional Data Streams (MDDS) of electricity logs is a challenging research problem. To end this, a novel electricity knowledge discovery model proposed from the MDDS using clustering and machine learning. Context-Aware Clustering with Whale Optimization Algorithm (CAC-WOA) is designed and explained in this research article. The CAC-WOA consists of two phases context-aware groups formation and WOA-based machine learning predictive model. In the CAC algorithm, group's formation using electricity contextual information to estimate the robust predictive features are proposed. Using such predictive features, the predictive model using the WOA-based Artificial Neural Network (ANN) is built. The modified ANN technique using the WOA algorithm is used to reduce the error rates and improve the prediction accuracy. The experimental outcomes using publically available electricity consumption datasets prove the efficiency of the CAC-WOA model.**

Keywords— Artificial neural network, bio-inspired optimization, context aware clustering, knowledge discovery, electricity monitoring, whale optimization.

I. INTRODUCTION

Data mining is a technique for extracting relevant data or patterns from massive data repositories such as relational databases, data warehouses, and XML repositories [1]. Pre-processing, data mining, and post-processing are the three steps of data mining, which is also known as the primary progression of Knowledge Discovery in Databases (KDD). Because the quantity of data in many databases has become incredibly big in recent years, knowledge discovery in databases (KDD) has become a technique of great interest. In big databases, KDD reflects the function of a nontrivial procedure for finding an efficient, consistent, possibly functional, and previously unknown pattern. Before the data mining algorithms are coupled with reliable data, the pre-processing step is employed to carry out the task. Data cleansing, integration, selection, and transformation are all included. The data mining phase of KDD is when a variety of methods are utilized to uncover buried knowledge. The mining result, as well as the users' demands and domain knowledge, are computed in the third step. The data mining step is the most important of all the phases. KDD is a mechanism for linking many branches of computer science expertise. It is the process of gathering data from a variety of sources and synthesizing it into valuable knowledge. Data mining software is a type of systematic data analysis tool. It helps users to look at data from a variety of perspectives to categorize and analyze the relationships that have been discovered. The approach of identifying connections or patterns using hundreds of fields in huge relational datasets is known as data mining [2]. Data mining is a technique for analyzing enormous amounts of data. It's also a set of instruments used to carry out the procedure. Data mining uses data from a variety of sources, including marketing, health, and communication.

In decision-making applications like electricity consumption, data streams play a critical role. Data mining techniques are required for knowledge discovery from such data sources about the electricity consumption patterns. Data streams are the continuous flow of information. Sensor data, call center records, network traffic, and other data streams are examples of data streams which forms the MDDS [3]. Authors sheer volume and pace provide a significant challenge for the data mining industry to extract knowledge. Data streams exhibit a wide range of characteristics, including feature evolution, indefinite duration, restricted labeled data, idea evolution, concept drift, and so on [4-7]. When the fundamental notion of data changes over time, concept drift occurs in data streams [8] [9]. When

separate classes form in streams, concept evolution occurs. Feature evolution occurs when the feature set in data streams evolves. Data streams also suffer from a lack of labeled data since it is impossible to manually classify all of the data points in the stream. Each of these characteristics adds to the difficulty of data stream mining.

Considering the electricity applications, the knowledge discovery from the different electricity logs that formed MDDS has received significant attention from researchers. By examining electric consumption statistics, policymakers and building owners can gain a better understanding of their customers' demand-consumption behaviors [10-12]. Moreover, the analysis and exact forecasting of energy consumption utilizing multi-dimensional data streams are vital for consumer participation in time-of-use tariffs, critical peak pricing, and user-specific demand response operations. Achieving accurate consumption forecasts helps to enhance power system planning and ensure dependable grid operations, in addition to having significant economic and environmental repercussions, such as reducing energy waste and accelerating the decarbonisation of the energy sector Energy consumption forecasting is a prominent issue in research; despite the availability of proven models, estimating electric consumption in residential buildings remains difficult due to the unpredictability of tenant energy use behavior [13][14]. As a result, the search for an appropriate model for making exact predictions about energy usage continues.

To acquire knowledge from the MDDS, researchers must overcome several challenges such as erroneous predictions, scalability, and efficiency. Many different types of data mining methods have been presented for the effective finding of information from multi-stream electricity datasets. Optimization methods, clustering methods, machine learning methods, etc. have frequently been used for electricity consumption prediction [15]. For accurate electricity prediction, context-aware electricity features and error-free predictive models [16] are two key requirements.

To overcome the challenges of existing solutions for electricity load forecasting or energy consumption forecasting, a novel CAC-WOA model is presented in this research article. As the name indicates, the model consists of two algorithms like clustering and WOA-based machine learning. The core functionality of CAC-WOA is briefly described below.

The researchers have proposed the CAC algorithm to perform the context-aware grouping of input multidimensional electric consumption data using the statistical features estimation technique. In CAC, the notion of self-supervised clustering is called CAC where the contextual knowledge is estimated as observation to devise clusters.

Compared to traditional clustering methods, the context-aware clustering method considers dependencies in goods at a greater contextual level to enhance prediction accuracy. The outcome of CAC is used as the prediction feature in the machine learning techniques for accurate forecasting.

To overcome the challenges of machine learning techniques, in this research article, researchers have modified the existing ANN classifier using the WOA for reducing the training MSE and errors and improving the overall accuracy. To obtain the MSE error, the WOA is used to initialize and update the weight vector of the ANN. The design, methodology, and experimental analysis of the CAC-WOA model using the publically available electricity consumption dataset are presented to justify the efficiency of the proposed model.

The remainder of the research article consists of the below sections. Section 2 presents a brief study of various KDD techniques. Section 3 presents the CAC-WOA design and methodology. Section 4 presents the simulation results. Section 5 presents the conclusion and future recommendations.

II. Related Works

Since the last decade, several novel methods have been introduced for the knowledge discovery from MDDS under different categories such as machine learning, bio-inspired optimization (BO), swarm intelligence (SI), and clustering under various applications. As the objective of the proposed research is on efficient knowledge discovery from the electricity logs, this section reviewed the recent methods on energy forecasting using clustering, BO, SI, and machine learning. To understand the role of clustering in knowledge discovery from the MDDS, researchers have mainly reviewed the clustering techniques.

A. State-of-Arts

Several strategies have recently been developed for predicting power usage by mining streaming live electrical data using various methodologies. These approaches all use multi-dimensional electric data streams as input to accomplish knowledge discovery in terms of load forecasting or energy consumption forecasting.

A suggested a clustering-based electricity forecast model in [17]. Authors created the subspace clustering based on the user's electricity usage characteristics' assessment index and then got various patterns for the user's power consumption. In this case, the clustering had been done automatically. Another clustering-based technique for energy consumption forecasting had presented in [18]. Authors created fundamental energy profile modeling based on the clustering of users. Using the streaming data from 288 commercial buildings, authors used the test technique to measure the system's correctness. The technique for estimating the energy consumption of the ancient educational building using the Support Vector Machine (SVM) and SARIMA (seasonal autoregressive integrated moving average) had devised in [19]. For electricity users, the short-term energy load prediction model presented utilizing the Fuzzy C-Mean (FCM) and K-means clustering in [20]. According to their electricity consumption parameters, authors used K-means clustering to separate the customers into two groups. The FCM technique was then used to filter out similar data. The data analysis processes provided suggested extracting the state of the appliance's power from its streaming energy consumption data in [21]. Authors created the method using a unique data learning model for nonintrusive load monitoring, which is a multi-target classification algorithm. The author of [22] looked at current models and proposed machine learning. Authors combined the Artificial Neural Network (ANN) with the Genetic Algorithm SI method (GA). Authors used a real-world testbed to put their prediction model to the test. A unique incremental learning

approach for predicting building energy consumption was examined in [23]. Authors created the swarm decision table method and compared it to the traditional decision tree approach. The experiments were carried out on multi-dimensional data streams connected to the Internet of Things (IoT) in real-time. For short-term power consumption forecasting, [24] developed a spatial and temporal ensemble forecasting model. Authors involved using cluster analysis and the k-means algorithm to investigate power usage profiles at the apartment level. The ensemble forecasting model of the two-deep learning models, Long Short-Term Memory Unit (LSTM) and Gated Recurrent Unit (GRU) were used in their model (GRU). The energy consumption forecast model for the Microsoft Azure Cloud-based method was introduced in [25]. For the prediction model, authors employed classifiers like SVM, ANN, and k-Nearest Neighbor (KNN). Another recent effort, the occupant-behavior-sensitive prediction model, was built for the prediction of building energy usage in [26]. Authors created the ANN, DNN (Deep Neural Network), EBT (Ensemble Bagging Trees), and CART machine learning algorithms (Classification and Regression Trees). The power consumption estimation model [27] was recently presented utilizing machine learning techniques in Agartala (India). Authors created a prediction model to forecast the load for the following 24 hours, then one week to one month. Authors created random forest (RF) and XGBoost classifiers, which are machine learning techniques. The hybrid machine learning approach for predicting appliance energy usage and peak demand had presented in [28]. For the forecast of appliance energy usage and customer peak demand, authors recommended the quicker k-medoids clustering technique, SVM, and ANN. Another contemporary clustering approach for energy forecasting, the two-layer Distributed Clustering Algorithm (DCA), was developed in [29] and uses affinity propagation and k-means clustering algorithms. Researchers went on to discuss the incentive Demand Response (DR), as well as the user-side DR flexibility. The dragonfly algorithm (DA) was proposed as a meta-heuristic optimization approach in [30]. Authors devised an algorithm to address the real-world problem of power monitoring in single and multiple smart homes. Authors divided the appliances into two categories: non-shiftable and shiftable. In [31], authors have developed a new technique for predicting individual consumer appliance energy usage from a set of connected user's electronic loads. For online applications, authors first acquire and store the current data of each appliance with changing load. After that, authors calculated individual load currents. To anticipate the electrical demand of individual users, authors developed the Artificial Bee Colony (ABC) method, which is a search-based optimization technique. For home energy forecasting, [32] examined the use of ensemble approaches to enhance the execution of ANN models. The instance was a house in Portugal with solar panels and batteries and a Home Energy Management System (HEMS) in charge. The largest reciprocal knowledge coefficient for monthly electricity use had proposed in [33]. First, the highest reciprocal knowledge coefficient had established between monthly power usage and its affecting elements. The high-relevance components have screened out established on the highest reciprocal information coefficient. The data with strong relevance parameters were combined. Finally, the random forest performed the prediction of electricity consumption prediction. In [34], a well-known deep learning transformer had used with the clustering method K-means to figure out how much power was used over time. The Transformer model was used to predict how much power the next hour would use, and the K-means clustering technique had employed to improve the prediction results.

B. Motivation

In this research article, researchers have discovered that knowledge discovery from MDDS is a difficult research topic from the recent literature on MDDS mining approaches in the electrical consumption forecasting domain. Clustering, machine learning algorithms, and SI are among the most often employed strategies for knowledge discovery from the MDDS. Based on the findings of these recent investigations, researchers have identified the following research gaps that motivate the proposed CAC-WOA model.

- Existing clustering-based methods [17] [18] [20] [24] [28] [29] [34] utilized the conventional (k-means, k-medoids, and FCM) approach for the group formation where the context information about energy consumption was neglected and just relied on the consumed energy for the group formation. It results in an erroneous approach for electricity monitoring.
- The machine learning techniques such as ANN, SVM, KNN, random forest, etc. were utilized for the prediction either individually or combined with clustering [17-33]. These techniques already suffered from Mean Square Error (MSE), training error, and accuracy challenges.
- Various optimization techniques such as DA [30], ABC [31], etc. were used to enhance the performance of machine learning techniques, but the selection of robust and efficient bio-inspired optimization is still a research problem. The deep learning techniques [34] not exploited into this domain due to lack of predefined deep learning models and higher time complexity.

C. Contributions

To overcome the challenges of existing solutions for electricity load forecasting or energy consumption forecasting, researchers proposed a novel CAC-WOA model in this research article. As the name indicates, the model consists of two algorithms like clustering and WOA-based machine learning. The contributions of CAC-WOA are briefly described below.

- Researchers proposed the CAC algorithm to perform the context-aware grouping of input multidimensional electric consumption data using the statistical features estimation technique. In CAC, the notion of self-supervised clustering is called CAC where the contextual knowledge is estimated as observation to devise clusters.
- Compared to traditional clustering methods, the context-aware clustering method considers dependencies in goods at a greater contextual level to enhance prediction accuracy. The outcome of CAC is

used as the prediction feature in the machine learning techniques for accurate forecasting.

- To overcome the challenges of machine learning techniques, researchers modified the existing ANN classifier using the WOA for reducing the training MSE and errors and improving the overall accuracy. To obtain the MSE error, the WOA is used to initialize and update the weight vector of the ANN.
- The design, methodology, and experimental analysis of the CAC-WOA model using the publically available electricity consumption dataset are presented to justify the efficiency of the proposed model.

III. CAC-WOA METHODOLOGY

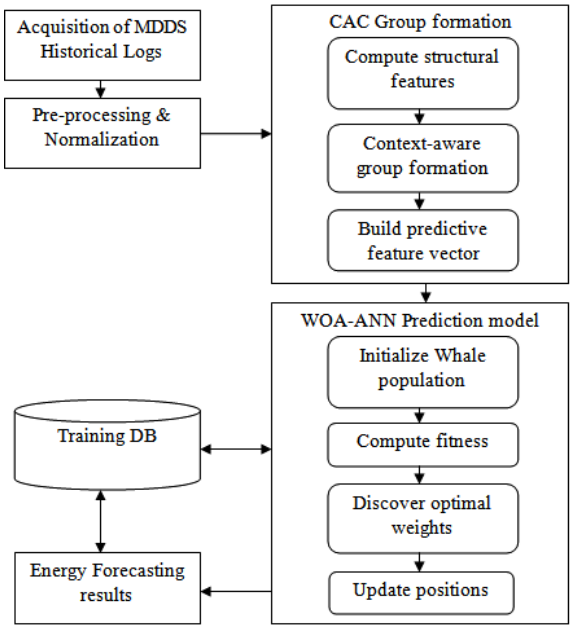

Figure 1. Proposed electricity consumption prediction model from MDDS

According to the contributions discussed above, the design and methodology of the proposed CAC-WOA model are presented in this section. Figure 1 shows the overall architecture of the proposed electricity knowledge discovery and prediction model from the MDDS logs. As shown in figure 1, the electric consumption data collected from the household residence is commonly called the MDDS historical logs. As such logs are collected for each home in the residential buildings, it may have raw or messy data which may lead to wrong knowledge discovery. Therefore, researchers first applied the lightweight pre-processing and data normalization algorithm to address the different types of data noises. After pre-processing of input electricity MDDS data, researchers applied the proposed CAC group formation where the context-aware structural features are estimated using the distance similarity measures. According to estimated features, researchers have sorted and formed the clusters. Each cluster is labeled with either of four electricity consumption patterns (e.g. seasonal, occasional, flat, and frequent). This process is called the build predictive feature vector for the further processing for electricity consumption prediction. The next step of the proposed architecture is to train and test the input predictive features using the proposed WOA-based ANN algorithm. As mentioned earlier, the process of ANN training is enhanced by applying the WOA technique to minimize the overall MSE rate and improve the prediction accuracy. The WOA technique performs the weight optimization using the current fitness evaluations. After training, researchers performed the testing for energy prediction by varying the test ratios 10 %, 20%, and 30%. As per the prediction outcomes, researchers estimated the prediction results in terms of accuracy, precision, recall, and MSE parameters.

A. *Data Pre-processing*

The data may have different kinds of noises such as wrong information, missing information, or incomplete information for some parameters. To overcome all such issues, researchers have proposed a lightweight mechanism using Natural Language Processing (NLP) and messy data handling techniques. The real-time applications connected with the IoT technology sometimes may produce the attributes in some tuples with messy data or incomplete data such as Inf, $Null$, or NaN values in place of the actual value. The occurrence of such values might lead to erroneous knowledge discovery and forecasting. As a result, such chaotic data from input raw logs must be suppressed or cleaned. The suggested pre-processing and data normalization technique is demonstrated in Algorithm 1. First, researchers have performed the detection and removed the special characters (SC) e.g., @, #, etc., complex characters (CC) e.g., a+, B-, etc., URLs, and stop words (SW) using NLP. Then, researchers have detected the location of messy data (Inf, $Null$, NaN) and then replaces that messy data with a average value corresponding to that parameter. The $mean(.)$ function discovers the average value from the entire column by suppressing the messy values. It ensures the clearance of such data from multi- dimensional sources.

Algorithm 1: Data pre-processing
Input D: Input electricity consumption data $M \leftarrow \{Inf, NaN, Null\}$: Messy data parameters Output N: Normalized dataset

1. Acquisition of MSSD data D
2. $[m,n] \leftarrow size(D)$
3. $N \leftarrow zeros\ (m,n)$
4. For each $i = 1{:}m$
5. For each $j = 1{:}n$
6. If $(D(i,j) \neq SC \ || \ D(i,j) \neq URL \ || \ D(i,j) \neq CC || D(i,j) \neq SW)$
7. $N(i,j) \leftarrow \mathrm{D}(i,j)$
8. Else
9. $N(i,j) \leftarrow \mathrm{mean}(:,j)$
10. End If
11. If $(D(i,j) == M)$
12. $N(i,j) \leftarrow \mathrm{mean}(:,j)$
13. Else
14. $N(i,j) \leftarrow \mathrm{D}(i,j)$
15. End If
16. End For
17. End For
18. Return (N)

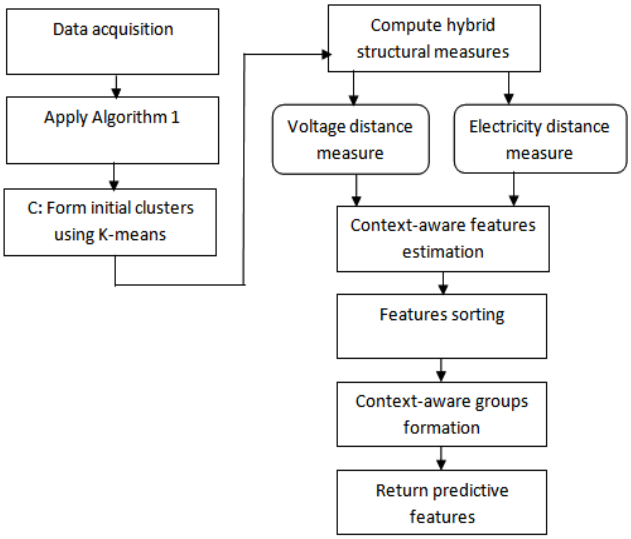

Figure 2. Architecture of CAC for group formation of electricity data

B. *CAC Group Formation*

The objective of context-aware group formation of the input pre-processed electricity data is to achieve reliable and efficient forecasting according to the consumption patterns compared to existing mechanisms. This section presents the design of the proposed CAC group formation approach. Figure 2 shows the overall functionality of the proposed clustering mechanism. As shown in figure 2, the input raw electricity data is first pre-processed using algorithm 1. After pre-processing, researchers have initially performed the k-means clustering to discover the centroids. After the formation of initial groups, researchers have performed the optimization of these groups to achieve the context-aware clustering of input data. In this article, researchers have extracted two parameters of each electricity log as velocity and consumed electricity to perform the context-aware clustering. The distance for velocity and electricity consumption of two rows is estimated using a well-known Euclidean distance. Once the distances are estimated for all the group members, researchers sorted them into ascending order. Finally, according to the sorted scores, the groups are re-formed. This process not only achieves the accuracy of prediction but also prevents data loss.

The normalized electricity data N is fed to k-means clustering cluster to compute the centroids and their group members (GMs) initially. The groups G with its centroid are computed using:

$$\mathrm{G} = \mathrm{kmeans}\ (\mathrm{N}, \mathrm{c}) \quad (1)$$

Where, G represents the groups formed using k-means technique of data N, c stands for the number of clusters (researchers set c=4 in this work). Each cluster $\mathrm{G}^{\mathrm{i}}, \mathrm{i} \in \mathrm{c}$ has at least r CMs. The value of r is not the same for each cluster, i.e., the number CMs in each cluster can be distinct.

As discussed earlier, the conventional group formation technique leads to several challenges in the domain of knowledge discovery and forecasting. To end this, researchers proposed a mechanism to optimize the initially formed clusters using a context-aware approach. Algorithm 2 shows the complete functionality of the proposed group formation technique. The optimization of the groups is possible by computing the similarity score of each electricity log in each cluster using two parameters such as velocity and electricity. As shown in algorithm 2, researchers initialized the output clustering vector H of size n and $\mathrm{m}+2$. The additional two columns are introduced to add the newly computed integrated predictive feature using the $\mathrm{hybridScore}\ (.)$ function and corresponding predictive label as a cluster number. The $\mathrm{hybridScore}\ (.)$ function estimates the similarity measure into variable h among the current GM log and centroid of the current cluster$\mathrm{C}^{\mathrm{i}}_{\mathrm{centroid}}$. Iteratively, for each cluster, the current row/log and its context-aware value h recorded into the vector H for all the n number of rows in the dataset N across the c number of clusters. All the entries in vector H are then sorted according to corresponding value h in ascending order into the vector S. After that researcher initialized the Predictive vector P similar to vector H to record the optimized groups with its labels. All the entries in S are grouped according to the context-aware score

with its group number into the output P. The number of logs (GMs) in each group should satisfy the constraint (n/c), where n is the total logs and c is the number of clusters.
In hybridScore $(.)$ function, researchers have measured the Euclidean distance among velocity parameters and electricity consumption parameters of current CM log and centroid of cluster to which the CM log belongs. The computations are elaborated below.

$$v_j^i = \text{dist}\left(\log^j.\text{Velocity}, C^i_{\text{centroid}}.\text{Velocity}, 'euclidean'\right) \quad (2)$$

Where, v_j^i represents the Euclidean distance for j^{th} log $\log^j$ that belongs to i^{th} cluster using its centroid C^i_{centroid}.

$$e_j^i = \text{dist}\left(\log^j.\text{Electricity}, C^i_{\text{centroid}}.\text{Electricity}, 'euclidean'\right) \quad (3)$$

Where, e_j^i represents the Euclidean distance for j^{th} log $\log^j$ that belongs to i^{th} cluster using its centroid C^i_{centroid}.

Algorithm 2: CAC

Inputs

$C: groups\ with\ centroids$

$c: total\ groups$

$N: pre-processed\ dataset$

$n: number\ of\ logs/rows\ in\ N$

$m: number\ of\ column\ in\ each\ row/log$

Output

$P: Context\ aware\ groups\ with\ labels\ as\ predictive\ feat$

1. Initialize**:** $H \leftarrow ones(n, m+2)$
2. $t = 1, t \in n$
3. for $i = 1: c$
4. for $j = 1: size\ (C^i)$
5. $log \leftarrow CM^i(j)$
6. h $\leftarrow hybridScore(\log,\ C^i_{centroid})$
7. $H(t,\ 1:m) \leftarrow log$
8. $H(t, m+1) \leftarrow$ h
9. $t \leftarrow t+1$
10. end For
11. end For
12. $S \leftarrow getSort(H(:, m+1), "ascending")$
13. Initialize**:** $P \leftarrow ones(n, m+2)$
14. for $i = 1: lenght\ (S)$
15. for $j = 1: c$
16. if $\left(lenght(C^j) \leq \left|\frac{n}{c}\right|\right)$
17. $C^j \leftarrow add\ \left(S(i, :)\right)$
18. $P \leftarrow S(i, :)$
19. $P(i, m+2) \leftarrow$ j, assign label
20. end if
21. end for
22. end for
23. Return (P)

Finally, the integrated hybrid context-aware or structural similarity score is computed using the weight-based approach. The hybrid score h_j for j^{th} log $\log^j$ is computed by:

$$h_j = \left(a^1 \times v_j^i\right) + \left(a^2 \times e_j^i\right) \quad (4)$$

Where a^1 and a^1 represent weights for each structural measure. The value of both weight parameters is $a^1 + a^2 = 1$. Both the velocity and electricity characteristics are given equal weight in this study., i.e., $a^1 = 0.5$ &$a^2 = 0.5$.

C. WOA-ANN Predictive Model

A WOA-based ANN is used to anticipate power usage. To attain the lowest MSE and greater training accuracy, the WOA is used to initialize and update the weight vector of the ANN. The suggested WOAANN model can address erroneous forecasting and predicting power usage. Because of its simplicity and efficacy, the ANN is frequently employed as a classifier in machine learning. It's also been used in power system intrusion detection models. The ANN's training is still a difficult challenge. Traditional training algorithms have a hard time dealing with slow convergence and local optima. To identify the appropriate weights and biases for issues like grey wolf optimization (GWO) and moth flame optimization (MFO), one recent trend is to train the ANN using bio-inspired meta-heuristic algorithms that imitate biological or physical events.

In this research article, researchers have suggested the WOAANN electricity forecasting prediction model, which is based on an ANN that is trained with the WOA. Feed-backward,

feed-forwards, and self-organizing maps are the three types of ANN architectures and associated neuron layouts. A multilayer perceptron (MLP) is a feed-forward neural network that uses the hidden layer to turn inputs into outputs. To train the network, the back-propagation algorithm was employed as the supervised learning pattern. WOA is a swarm-based intelligent search approach [35] that is used to discover assaults while overcoming the delayed convergence difficulty and the "local minima" trap associated with ANN. WOA is known as an effective and competent method for solving optimization issues since it can search and estimate the identified neighborhood space of the global optimum. Researchers have employed a WOA algorithm as a trainer for a feed-forward neural network to address the problems associated with the learning technique in neural networks. It has been demonstrated that this approach can tackle a broad range of optimization issues and outperforms other existing algorithms in training MLPs because it is a gradient-free and flexible machine capable of local-optima avoidance.

Figure 3 demonstrates the functionality of the proposed WOA-based ANN training approach. This takes the input as the vector P as the predictive feature vector. The WOA consists of steps such as population initialization (also called as whale search agents initialization), fitness computation of each whale agent, evaluation and updation of hunting whales position, and estimation of optimal solution on convergence. Each whale search agent is initialized to optimize a candidate neural network in WOA-based ANN training. An MLP network has weight and bias vectors that indicate the relationships among the input and hidden layers, as well as the hidden and output layers. Equation (5) illustrates the total number of weights and biases (WB) variable combinations in the MLP network that will be improved using WOA. In this equation, n represents the total number of input nodes (the total number of rows in the input vector P), and q represents the total number of neurons in the hidden layer.

$$WB = n.q + 2.n + 1 \quad (5)$$

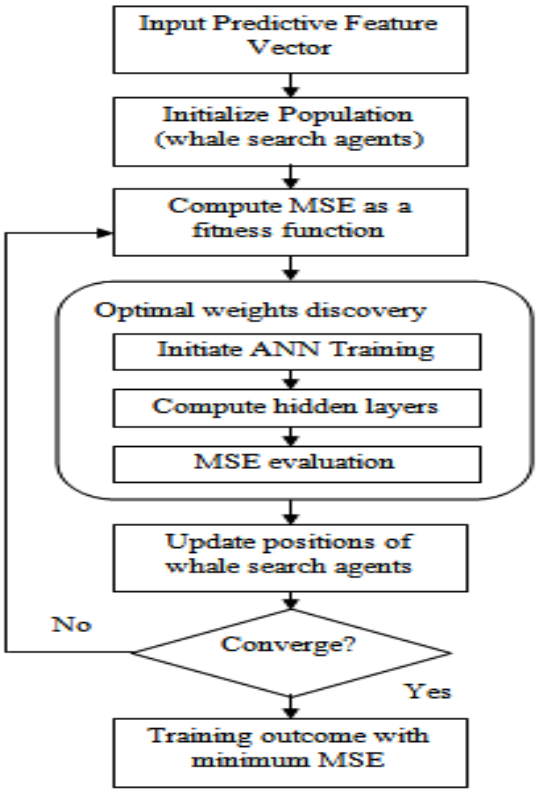

Figure 3. Architecture of WOA-based ANN training

Essentially, the MSE of the MLP network functions as a fitness function, and it is utilized by the whale search agents to assess how much difference there is between actual A and anticipated classes $\bar{A}$. The MSE for k number of samples is computed by:

$$mse = \frac{\sum_{x=1}^{k}(A_x - \bar{A}_x)^2}{k} \quad (6)$$

This fitness value is utilized in the WOA optimization algorithm as shown in figure 3. Before performing the WOAANN training and classification steps, the numbers of feature vectors are estimated from the vector P. These features vectors are randomly distributed into training and testing ratios. In our experiment, researchers have divided into three combinations such as 90 % (training)-10 % (testing), 80 % (training)-20 % (testing), and 70 % (training)-30 % (testing). Finally, the testing data is input into the ANN classification model using the optimal weights and biases established during the training phase to assess the model's accuracy. The WOAANN is very adept at avoiding local optima and reducing the MSE rate of ANN training. It increases the likelihood of identifying the best MLP weights and biases for the proposed model, which are linked with high accuracy and efficiency.

IV. Simulation Results

In this research article, researchers have implemented the proposed model using Python 3.10.2 under the Windows 10 OS with an Intel I5 processor with 8GB RAM and Intel graphics. Researchers have collected the MDDS electricity logs from the publically available research dataset [36]. The dataset contains power usage statistics from a high-rise residential building on the IIT Bombay campus from December 2016 to January 2018. Each of the 60 3BHK "(3 Bedrooms, Hall, and a Kitchen)" flats in the building is equipped with a smart meter that records data every 5-8 seconds. All timestamps in the dataset are in GMT+5.30 (Indian Standard Time). Each entry in the dataset originally consists of 6 fields as timestamp, phase 1 voltage, phase 2 voltage, phase 3 voltage, phase 1 electricity consumption, and phase 2 electricity consumption. After applying algorithms 1 and 2, researchers have a dataset with additional two entries such as integrated predictive feature and corresponding group number (1, 2, 3, or 4). As mentioned earlier, researchers have analyzed the performance of the proposed WOA-based ANN model with the other conventional classifiers such as SVM, ANN, and KNN. For all these classifiers, researchers fed the predictive feature vector with labelling P. The results are analyzed in terms of training MSE, training accuracy, testing accuracy, testing precision, and testing recall for energy consumption forecasting. Furthermore, the performance of the proposed model is compared with state-of-art methods in terms of accuracy, precision, recall, and forecasting duration in section B.

A. *Performance Investigation*

Figures 4-7 demonstrate the comparative investigation of the training MSE performance using the different predictive models such as KNN, SVM, ANN, and WOA-ANN for varying test ratios. Figure 4, figure 5, and figure 6 show the training MSE with a forecasting span of 200 minutes for 10 %, 20 %, and 30 % testing samples respectively. From these results, researchers have first noticed that as the forecasting span increases, the training error minimizes due to increasing historical data with increasing forecasting duration. It also indicates that training on initial samples will lead to higher errors. Therefore, historical MDDS plays a significant role in accurate forecasting. The second thing that researchers have observed from the results shown in figures 4-6 is the impact of test samples size. The lower test samples (means higher training samples) have delivered the minimum MSE due to higher training data for the classification. Among KNN, SVM, ANN, and WOA-ANN methods, the proposed forecasting model using WOA-ANN reduced the training MSE significantly compared to CAC with KNN, SVM, and ANN forecasting models. It is due to the optimized ANN approach using the WOA technique. The WOA can reduce the MSE while ANN training and classification phase compared to conventional classifiers. Therefore, it significantly affects the forecasting outcomes as well.

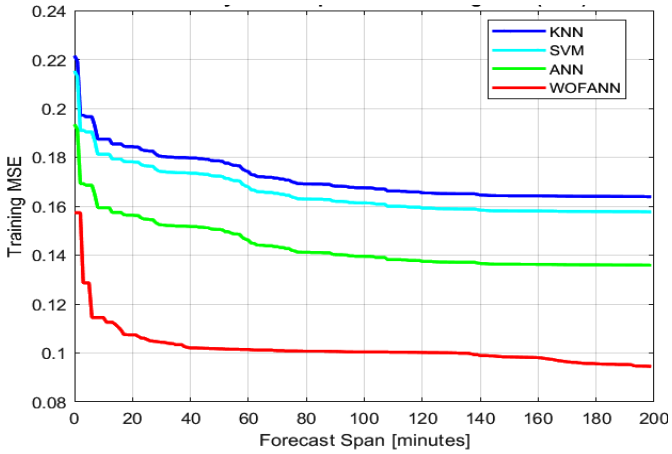

Figure 4. Training MSE analysis for 10 % test ratio

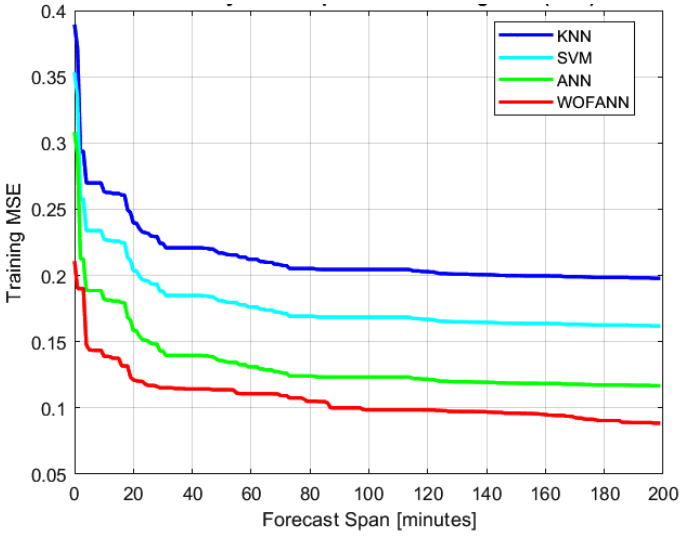

Figure 5. Training MSE analysis for 20 % test ratio

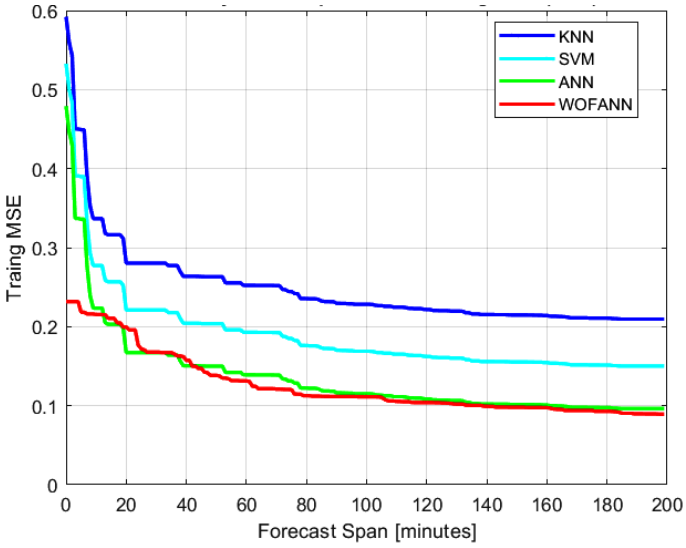

Figure 6. Training MSE analysis for 30 % test ratio

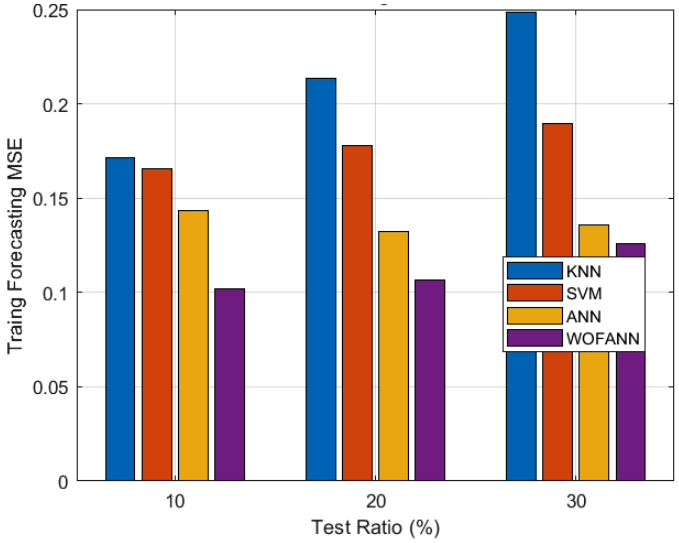

Figure 7. Average analysis of training MSE

Figure 7 demonstrates the average outcomes of training MSE for each test ratio scenario. The average training MSE is 0.1, 0.115, and 0.122 for 10 %, 20 %, and 30 % training samples. These outcomes show a significant reduction compared to conventional classifiers. The reduction in MSE has a direct impact on the training accuracy as shown in figure 8. The average training accuracy for each test sample ratio is computed for KNN, SVM, ANN, and WOA-ANN techniques. The accuracy of the proposed WOANN-based knowledge discovery or energy consumption forecasting model is improved compared to other prediction models. It is due to the proposed context-aware clustering with the WOA-ANN mechanism that reforms the groups with effective predictive features followed by an optimized ANN classification model.

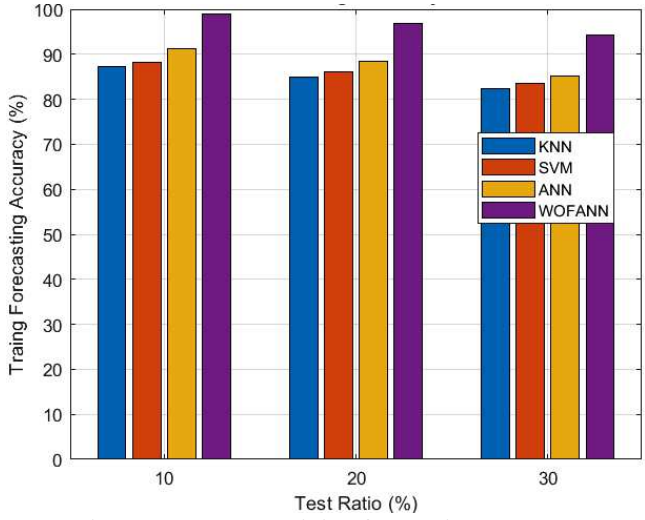

Figure 8. Average training forecasting accuracy

Furthermore, researchers have analyzed the performance of the proposed electricity consumption forecasting results in terms of 10%, 20%, and 30% test ratio classification. The testing accuracy, precision, and recall rates performances are measured for each test ratio scenario using KNN, SVM, ANN, and WOA-NN classifiers. Figure 9 demonstrates the outcome of overall forecasting accuracy for 10-30% test samples using different classifiers. Similarly, figure 10 and 11 shows the outcome of overall forecasting precision and recall for 10-30% test samples using different classifiers respectively. It shows the increasing number of test samples reduces the overall forecasting accuracy, precision, and recall rates due to the reduced number of training samples. The proposed CAC with the WOA-ANN forecasting model delivered higher prediction accuracy, precision, and recall rates compared to KNN, SVM, and ANN approaches. These results also claim that optimizing the group formations for knowledge discovery is not enough for improving the forecasting performances. The performance of the proposed CAC group formation algorithm has been enhanced by applying the optimized predictive model called WOA-ANN (i.e., CAC-WOA).

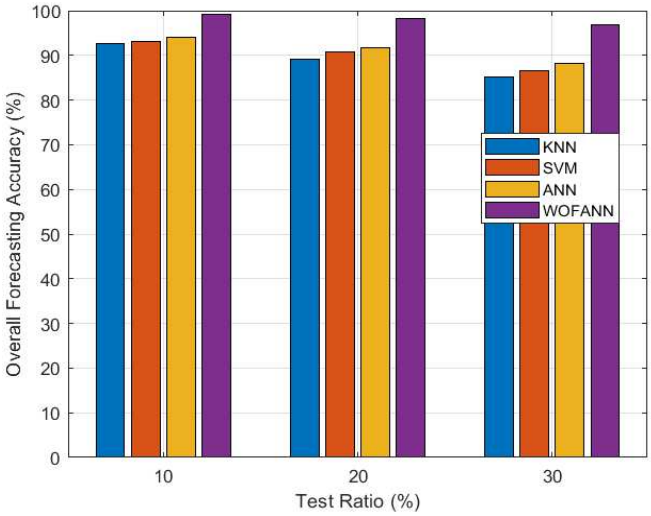

Figure 9. Performance analysis of forecasting accuracy

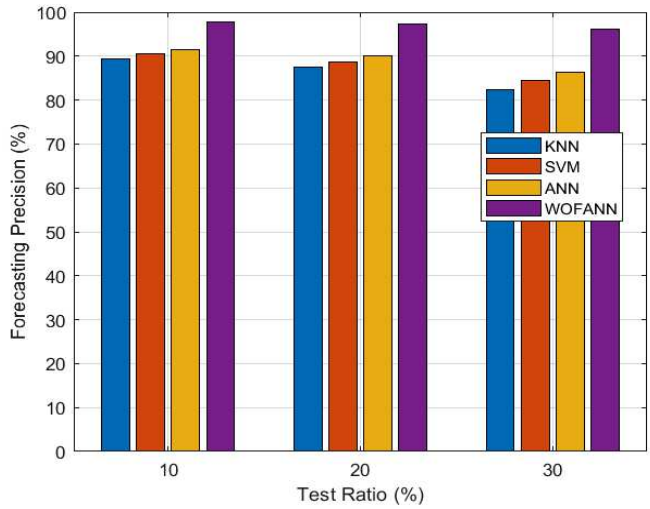

Figure 10. Performance analysis of forecasting precision

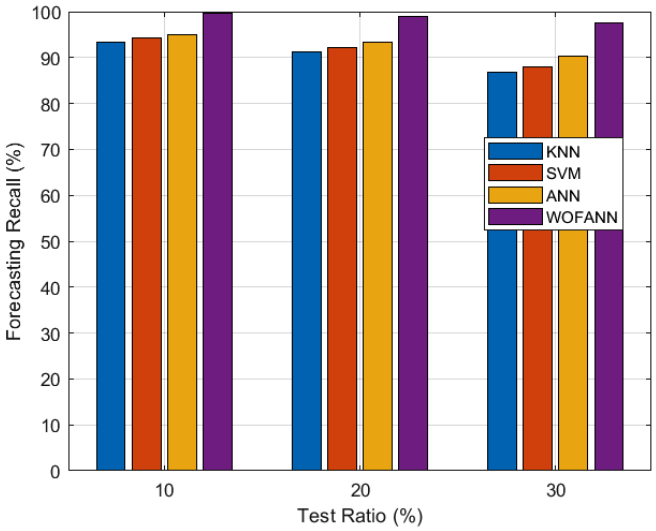

Figure 11. Performance analysis of forecasting recall

B. *State-of-Art Analysis*

This section presents the comparative analysis of the proposed model compared to recent existing energy consumption forecasting techniques using different approaches. Researchers have compared the performance of the proposed model CAC-WOA with four recent techniques such as Hussain et. al [30], Ghosh et. al [31], Bot et. al [32], and Pang et.al [33]. These methods are closely related to the proposed model where they have used Dragonfly and genetic algorithm in [30], Artificial bee optimization in [31], ensemble forecasting model in [32], and random forest model in [33] for energy consumption prediction. Researchers have applied these methods using the dataset and other experimental settings mentioned above for the comparative study in this research article. Table 1 demonstrates the average outcomes in terms of accuracy, precision, recall, and forecasting time. These results are estimated for 70 % (training) and 30 % (testing) via 10 executions of each method. From the outcomes presented in table 1, the proposed CAC-WOA method significantly improved the overall knowledge discovery from MDDS and forecasting results compared to recent solutions. The performances are improved due to optimized clustering using the context-aware group formation using structural similarity measures and modified ANN using the WOA approach.

Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Forecasting Time (Seconds)**
Hussain et. al [30]	95.72	93.74	96.32	1.99
Ghosh et. al [31]	94.32	93.23	95.37	1.64
Bot et. al [32]	94.12	93.02	95.06	1.57
Pang et.al [33]	93.21	92.43	94.73	1.53
CAC-WOA	97.43	96.57	98.67	1.49

V. 5. Conclusion and Future Works

This research article proposed the novel CAC-WOA framework for efficient knowledge discovery from the multi-dimensional electricity consumption data logs. These knowledge discoveries have been performed for the forecasting of electricity consumption for smart electricity monitoring applications. The proposed model consists of data pre-processing, CAC algorithm, and ANN-WOA. In data pre-processing, researchers have improved the quality of raw electricity logs to improve the forecasting reliability and efficiency. In the CAC approach, researchers have addressed the limitations of existing clustering mechanisms by forming the groups according to the context information of the dataset. The outcome of the CAC approach is the predictive feature vector with its predictive labels. Finally, researchers have designed the WOA-based ANN predictive model for accurate prediction accuracy and reduced MSE. The experimental results revealed that the proposed CAC-WOA model has reduced the MSE rate by 35 % compared to conventional techniques. Apart from this, the overall forecasting accuracy, precision, and recall rate performances are improved by 7.5 %, 8.9 %, and 8.69 % respectively compared to recent solutions. Investigating the CAC-WOA approach using the dataset of different dimensions and scales will be the first future work. Secondly, improving the CAC-WOA approach by using other optimization techniques for CAC is another interesting future direction.

Funding
Authors are responsible for funding.

Informed consent
Author and co-author are well aware about publication.

Conflict of interest
There is no conflict of Interest in the presented research work.

Data Availability
The datasets generated during and/or analyzed during the current study are publicly available.

References

[1] Khan L., Fan W. Tutorial: Data Stream Mining and Its Applications. In: Lee S., Peng Z., Zhou X., Moon YS., Unland R., Yoo J. (eds) Database Systems for Advanced Applications. DASFAA 2012. Lecture Notes in Computer Science, vol 7239. Springer, Berlin, Heidelberg, 2012.

[2] Mohamed, Hoda. Data Stream Mining. In Proc. of the 1st International Conference on Machine and Web Intelligence (ICMWI'2010), Algiers, Algeria, Oct. 2010.

[3] Pramod, S & Vyas, O. Data Stream Mining: A Review. 10.1007/978-1-4614-3363-7_75, 2013.

[4] Alothali, Eiman & Alashwal, Hany & Harous, S. Data stream mining techniques: a review. TELKOMNIKA (Telecommunication Computing Electronics and Control). 17, 2019.

[5] Agrawal, Lalit. Survey and Research Issues in Data Stream Mining. Bioscience Biotechnology Research Communications. 13. 146-149, 2020.

[6] R, Padma. Review in Data Stream Mining in Big Data. International Journal for Research in Applied Science and Engineering Technology. 8. 405-408, 2020.

[7] Rutkowski, Leszek & Jaworski, Maciej & Duda, Piotr. Decision Trees in Data Stream Mining. 10.1007/978-3-030-13962-9_3, 2020..

[8] Rutkowski, Leszek & Jaworski, Maciej & Duda, Piotr. Basic Concepts of Data Stream Mining. 10.1007/978-3-030-13962-9_2, 2020.

[9] A.Mehdi, Osama & Pardede, Eric & Ali, Nawfal. KAPPA as Drift Detector in Data Stream Mining. Procedia Computer Science. 184. 314-321, 2021.

[10] Bot K., Ruano A., da Graça Ruano M. Forecasting Electricity Consumption in Residential Buildings for Home Energy Management Systems. In: Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1237. Springer, Cham., 2020.

[11] Nti, I.K., Teimeh, M., Nyarko-Boateng, O. et al. Electricity load forecasting: a systematic review. Journal of Electrical Systems and Inf Technol 7, 13 ,2020.

[12] Gonzalez-Briones, A., Hernandez, G., Corchado, J. M., Omatu, S., & Mohamad, M. S. Machine Learning Models for Electricity Consumption Forecasting: A Review. 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019.

[13] Ferlito, S., Atrigna, M., Graditi, G., De Vito, S., Salvato, M., Buonanno, A., & Di Francia, G. Predictive models for building's energy consumption: An Artificial Neural Network (ANN) approach. 2015 XVIII AISEM Annual Conference, 2015.

[14] Bahij, Mouad & Labbadi, Moussa & Cherkaoui, Mohamed & Chatri, Chakib & Elkhatiri, Ali & Elouerghi, Achraf. A Review on the Prediction of Energy Consumption in the Industry Sector Based on Machine Learning Approaches. 01-05, 2021.

[15] R., Arumugam, P., & Jose, P. Revealing Household Electricity Power Consumption Using Data Mining Algorithms. *International Journal Of Statistics And Reliability Engineering, 7*(3), 350-354, 2021.

[16] González Briones, Alfonso & Hernández, Guillermo & Pinto, Tiago & Vale, Zita & Corchado Rodríguez, Juan. A Review of the Main Machine Learning Methods for Predicting Residential Energy Consumption.. 1-6, 2019.

[17] Zhou Xiangyu, Miao Qingqing, Lin Tao, Zhang Lei, and Zhou Jianquan. Linear Regression Electricity Prediction Method Based on Clustering of Electric Characteristics. In Proceedings of the 2nd International Conference on Big Data Technologies (ICBDT2019), Association for Computing Machinery, New York, NY, USA, 171–176, 2019.

[18] Shchetinin, Eugene. Cluster-based energy consumption forecasting in smart grids. Journal of Physics: Conference Series, 2019.

[19] Qiao, Q., Yunusa-Kaltungo, A., & Edwards, R. Hybrid method for building energy consumption prediction based on limited data, 2020.

[20] Haihong Bian, Yiqun Zhong, Jianshuo Sun, Fangchu Shi. Study on power consumption load forecast based on K-means clustering and FCM–BP model. Energy Reports, Volume 6, Supplement 9, Pages 693-700, 2020.

[21] Buddhahai, B., Wongseree, W., & Rakkwamsuk, P. An Energy Prediction Approach for a Nonintrusive Load Monitoring in Home Appliances. IEEE Transactions on Consumer Electronics, 1–1, 2019.

[22] Bourhnane, S., Abid, M.R., Lghoul, R. et al. Machine learning for energy consumption prediction and scheduling in smart buildings. SN Appl. Sci. 2, 297, 2020.

[23] Li, T., Fong, S., Li, X., Lu, Z., & Gandomi, A. H. Swarm Decision Table and Ensemble Search Methods in Fog Computing Environment: Case of Day-Ahead Prediction of Building Energy Demands Using IoT Sensors. IEEE Internet of Things Journal, 7(3), 2321–2342, 2020.

[24] Khan A-N, Iqbal N, Rizwan A, Ahmad R, Kim D-H. An Ensemble Energy Consumption Forecasting Model Based on Spatial-Temporal Clustering Analysis in Residential Buildings. Energies. 2021.

[25] Mel Keytingan M. Shapi, Nor Azuana Ramli, Lilik J. Awalin. Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. Developments in the Built Environment, Volume 5, 2021.

[26] Kadir Amasyali, Nora El-Gohary. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. Renewable and Sustainable Energy Reviews, Volume 142, 2021.

[27] Banik, R., Das, P., Ray, S. et al. Prediction of electrical energy consumption based on machine learning technique. Electr Eng 103, 909–920, 2021.

[28] Ejaz Ul Haq, Xue Lyu, Youwei Jia, Mengyuan Hua, Fiaz Ahmad. Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach. Energy Reports, Volume 6, Supplement 9, Pages 1099-1105, 2020.

[29] Ma, Hongyan. The role of clustering algorithm-based big data processing in information economy development. PLOS ONE. 16, 2021.

[30] Hussain, Ullah, M., Ullah, I., Bibi, A., Naeem, M., Singh, M., & Singh, D. Optimizing Energy Consumption in the Home Energy Management System via a Bio-Inspired Dragonfly Algorithm and the Genetic Algorithm. *Electronics*, *9*(3), 406, 2020.

[31] Ghosh, S., & Chatterjee, D. Artificial Bee Colony Optimization Based Non-Intrusive Appliances Load Monitoring Technique in a Smart Home. IEEE Transactions on Consumer Electronics, 67(1), 77–86, 2021.

[32] Bot, Santos, S., Laouali, I., Ruano, A., & Ruano, M. da G. Design of Ensemble Forecasting Models for Home Energy Management Systems. *Energies*, *14*(22), 7664, 2021.

[33] Pang, X., Luan, C., Liu, L. *et al.* Data-driven random forest forecasting method of monthly electricity consumption. *Electr Eng,* 2022.

[34] Zhang, J., Zhang, H., Ding, S., & Zhang, X. Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. Frontiers in Energy Research. Vol.9, 2021.

[35] Aljarah, I., Faris, H., & Mirjalili, S. Optimizing connection weights in neural networks using the whale optimization algorithm. Soft Computing, 22(1), 1–15, 2016.

[36] http://seil.cse.iitb.ac.in/residential-dataset/