Demand-Side Electricity Management through Machine Learning

Rishabh Joshi ,
Department of AIT-CSE,
Chandigarh University
joshirishabh333@gmail.com

Rahul Kumar,
Department of AIT-CSE,Chandigarh University
helloitsrahul25@gmail.com

Prabhjot Singh Bali,Department of AIT-CSE, Chandigarh University
prabjot.e16592@cuchd.in

***Abstract*—This project aims to predict the energy consumption in household. The primary objective is to learn the patterns in the electricity consumption and further test the accuracy of our result. The project makes use of signal processing methods and some statistical methods to better understand the electricity usage patterns and predict the usage of the energy in future. There are many techniques in signal processing and machine learning that, when combined together are a very good analyzer for the given system. Here, in this project we do the analysis using the Fourier transform and convolution to get to the results. We also make use of convolution and curve fitting to best analyze our result.**

Keywords—signal processing, statistical methods, electricity, patterns, prediction, energy consumption, curve fitting, Fourier transform, convolution

I. Introduction

Electricity has now become an integral part for the today world. Most of the appliances at our home, the devices we use all run-on electricity. With the change in seasons and time, the usage of electricity also varies. For example- A building containing an office will need more electricity in the working time and no electricity in the night and morning. We want to understand the usage pattern of an average household for the electricity consumption and get inference on how the house uses the electricity in a full year. We also wanted to use some signal processing methods like the Fast Fourier Transform and convolution on our data since there will be many fluctuations in our data. We make the use of Sci-kit Learn package present in the Python language to make some predictions on the data. NumPy and SciPy are used to process in input data and apply the algorithms. Vedo, a library that makes use of VTK is used for plotting. Vedo was the choice because it can easily scale to the number of datapoints present in the data. Jupyter Notebook is an interactive web notebook which is used to present the data.

II. Literature Review

It is certain that we are not the first researchers on this topic. We rely heavily on previous researchers and the high-quality paper they put on. They introduced many interesting concepts and applied techniques that helped gain a good understanding of the data. The authors in [1] used four techniques: Bagging, Stochastic Gradient Boosting, Model Averaged Neural Network and the K-Nearest Neighbors. They found out that GBM was able to predict with an outstanding accuracy of 96%. They also argue about the limitations like overfitting, high variance and bias in the data. They concluded that GBM has a great potential and can help accurately predict accuracy of consumption. They ask future researchers to integrate GBM with soft computing methods like Moth Flame Optimization, Harris Hawks Algorithms, etc. The authors in [2] start by telling the importance of predictions and then proceed to categorize the types of prediction in different categories. The authors conclude their research by emphasizing the need of more research in this area. Authors believe that the advancement can lead to savings for the corporations. The authors in [3] make use of Deep learning methods that predict the usage in short term, medium term and long-term. They called the model DLA-PM, Deep Learning Architecture for Power Management. It is designed for managing the energy fluctuations. The home appliances are connected to the cloud server using home automation techniques linked to the servers using IoT system. They conclude by saying that DLA-PM can be used in future years for smart cities to meet the power needs and preserve natural resources in future years. The authors in [4] use ensemble learning techniques like AdaBoost, Bagging, SVM and Decision Tree. The calculation of error made the use of MSE, MAE and Mean Absolute Percentage Error. To reduce the number of features in the dataset, they make the use of PCA and LDA to check the influence of features. They tried to find the optimal number of features for the best prediction. The authors in [5] argue about the increment of the use of ML in this field due to evolving infrastructure. They tried to make the paper more accessible to the people of engineering and CS background because of the lack of real-world models. They argued about the use of physics-based ML approaches that require less data but give good precision. Many papers made the use of the neural networks like, in [6], they compared the ANN and RF for classification purpose. RF means Random Forest. They also modeled the parameters like the number of guests to increase the realism of the prediction and increase the accuracy of their model. For future use, they directed to explore the other methods and big-data technologies for training and deployment. Again, the researchers in [7], did a data-driven predictive modelling of the appliances in our household. They evaluated four statistical methods: Multiple Linear Regression, Support Vector Machine with radial kernel, Random Forest, and Gradient Boosting Machines (GBM), with GBM performing best in all of the test cases taken. The authors in [7] knew about the limitation of the model. They knew that the data being only for one house cannot be the best in quality for the analysis. The authors in [8], [12], [13] made a model that would predict the next day usage of electricity given today's usage. Their system uses a sensor network tat monitors the consumption to predict next-day usage, this includes what devices will be used, when and how long it will be used. They tried to automate the task as much as possible. The ultimate goal for the authors was to minimize the energy bill and improve smart grid efficiency. The authors in [9], [10] did the work on load forecasting. They argued that the short-term load forecasting is necessary for corporations because it gives idea about the approximate usage of electricity. They also talk about the ANN techniques that are used for the purpose like backpropagation used for calculating the weights in the network. At last, they talk about the limitation of the networks and computing capability of the machines. The authors in [14], did a great job in finding patterns in electricity usage during in the CoVid-19 pandemic. They found that unlike other states like New York which saw a decrease in power usage, Florida and Texas witnessed a surprising increase. Authors in this study propose new machine learning models to improve short-term electricity consumption prediction in the context of COVID-19's impact. They explored Long-Short-Term-Memory, Convolutional Neural Networks, and Support Vector Machines, comparing performance with ensemble learning techniques. Their work highlights the potential of these during uncertain events. Finally in the paper [14], the authors analyzed the usage of IoT systems in different areas like healthcare and data collection. They also discuss about Wireless Sensor Network or WSN which is a collection of sensors connected to a central point. This paper focused more on the IoT side and the usage in medicine, but it provided a good start by telling the things like WSN, central server and usage in medicine for collecting data.

III. Methodology

A. *Downloading the dataset*

Before we can do any inference on average electricity consumption, we need some data which can help us do statistics. The dataset should be large enough that we can see the yearly variation in the electricity consumption of the household. Since this kind of dataset will take a large space the disk and will be computationally expensive, we used a powerful enough machine to process the data we were exploring. The dataset used in this paper is "Individual household electricity consumption". This dataset contains time series of 2,075,259 points. The data collected is between December 2006 and November 2010.

B. *Feature Selection*

This step is a crucial step in our method, or any other machine learning project. In this step, we identify the potential features we need or can use in our method. The machine learning methods rely heavily on the quality of features in our data. For a data processing task, feature is the column of the data which we can make use of, for our training.

This step requires us to find relevant features in the data which are relevant to our problem. This often requires expertise in the domain and good understanding of our data. The features used in our problems are the time and global intensity. These both are continuous time-series data and are a reliable source for predicting consumption at the moment.

Feature selection comes under the domain of feature engineering. Feature engineering plays and important role in the implementation of any machine learning project. By carefully selecting and transforming the data, we can increase the effectiveness of our model.

C. *Preprocessing the data*

Any machine learning method relies heavily on the quality of the data. It can be further simplified: Machine learning models require clean and well-structured data for good output. Only loading data and doing analytics just doesn't work. For example, NumPy requires the datatypes to be "int", or a "float" to calculate the mean and other quantitative overview of the data. By "int" and "float", we mean that the datatype can be any "int" or "float", it can also be "int8", "int16", "float16", "float32" or any other datatype other than "None" or "str". But, if the data contains even a single "None" type, it will change the datatype of the whole data to be "str". And, in "str" datatype, we cannot do general analytics like mean, deviation and variance.

The size of our dataset, when uncompressed, as a plain Comma Separated Values (CSV) file was around 130MB which is huge compared to most of the other models. For performance reasons, we made the use of SQLite database to decrease the time taken in calculation in the data. It also has an in-built support in the Python language so we used it for further calculations.

D. *Data Exploration*

Data Exploration is a method of finding patterns in the data by simply visualizing the data using a software. This step is sometimes also known as Visual Data Mining because it is easy to find the patterns in data by making a graph in the data.

There are a plenty of data visualization libraries present in the Python ecosystem but we chose the Vedo library to be the best candidate. Vedo is a wrapper library around another package called VTK. VTK stands for Visualization Toolkit and used for processing the data in bulk. It contains many types of plots and it can also represent 3D very efficiently.

While plotting the data, there was clearly a trend in the data. The consumption of electricity was very low in the morning time (as expected) but spiked around 8PM, which was also expected as most of the people come home at this time.

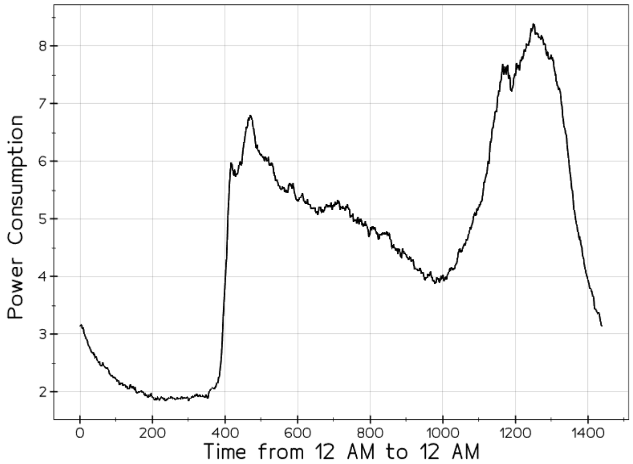

Fig.1 Average Energy Consumption in 24 hours

While visualizing the data between electricity consumption vs. any given time, it also showed an interesting pattern. It was clear that at the start of the year, the consumption was very low. It gradually increased till the mid-year and again started to drop around the month of August or so. It was producing a clear pattern of usage and looked like a sinusoid wave. The sinusoid behavior of the data lead to the use of some algorithms like Fast Fourier Transform and Convolution.

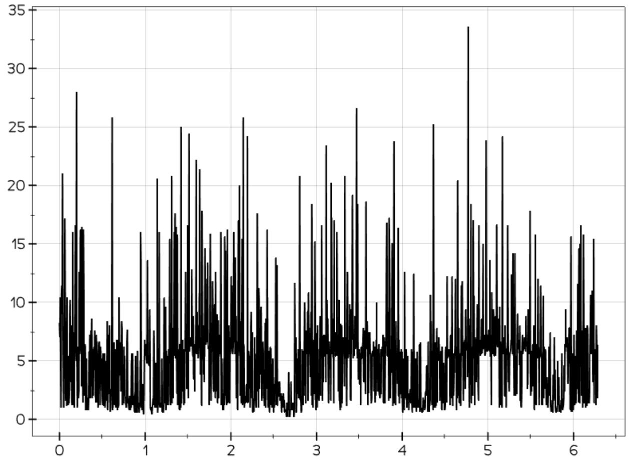

Fig. 2 A sample pattern around 12 noon.

IV. Fourier Analysis

In signal processing applications, it is important to understand the frequency response of a variable to another. Fourier analysis is a powerful tool to understand the signals by decomposing it into the constituent sinusoidal components. The Fourier series is given by:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx)$$

Fourier series finds it's applications in many of the signal processing domain. Some of the application of Fourier series are:

- **Signal Analysis**: It helps identify the different frequency and amplitude components that make a signal
- **Signal Reconstruction**: Since it can analyze a signal, it also helps in reconstructing a signal. An application of signal reconstruction is a Function generator. It is an electronic component which can create an electric wave of any frequency whichever is needed.
- **Communication Systems**: It also finds it's use in communication systems. It is used to filter out the unwanted frequency, consider a noise in the signal.

Fourier Transform is another part of Fourier analysis. Fourier Transform decomposes the given signal to the non-periodic signals into the frequency domain. The formula is represented by:

$$\hat{f}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx$$

The above formula gives the formula of Fourier transform, which is sometimes also represented by:

$$\mathcal{F}\{f(x)\} = \hat{f}(x)$$

Fourier transform finds its application like:

- **System Analysis:** It is used to analyze the frequency content of some non-periodic signals.
- **Image and Signal Processing:** It is also used in places like image filtering, compression and spectral analysis of the signals.

We ourselves used the Fourier transform to clean our input signals. It was clear from the initial visualizations that there are many spikes in the energy consumptions. To get a better understanding of the pattern, we had to remove some signals that were having negligible contributions. After doing many trail and errors in cleaning data, we settled to put the threshold to be half of the mean of the Fourier transform, that was produced. To recreate the signal from the data we just got, we used Inverse Fast Fourier Transform algorithm to further help us in getting to the result.

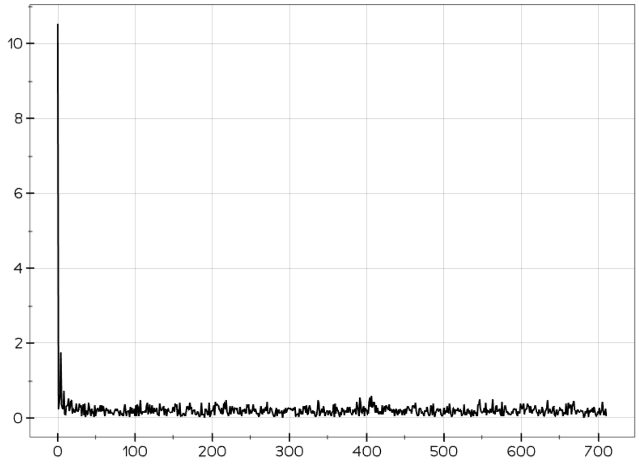

Fig. 3 Result of Fourier Transform

After applying FFT, and cleaning the signal now looks like the plot given below:

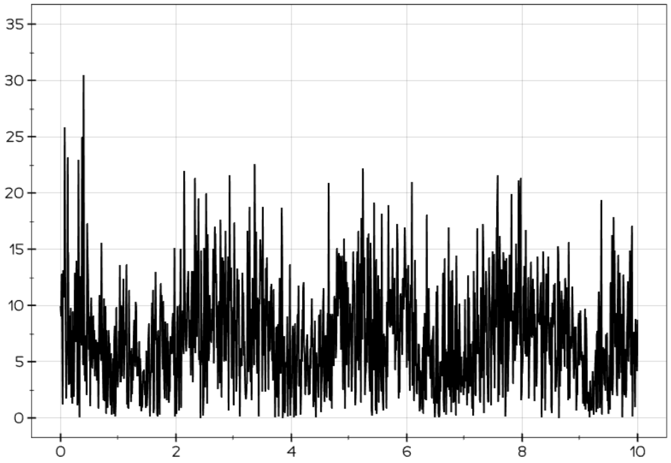

Fig. 4 Recreated signal after FFT is applied

V. Convolution

Convolution is a fundamental operation in signal processing. It is used to analyze the relationship between the input signal and the response output that we get from the system.

Mathematically speaking, convolution between two signals $x(t)$ and $h(t)$ is defined as following:

$$x * h = \int_{-\infty}^{\infty} x(t)h(t-\tau)d\tau$$

It is denoted by the operator "asterisk" and it is considered a basic operation like Addition and Subtraction. We can naturally describe convolution of a function $x(t)$ over another function by flipping the function, sliding the function and adding it over all the possible values.

The properties of convolution are as follows:

- **Commutative**: $x(t) * h(t) = h(t) * x(t)$
- **Associative**: $x(t) * \big(h(t) * g(t)\big) = \big(x(t) * h(t)\big) * g(t)$
- **Distributive**: $x(t) * \big(h(t) + g(t)\big) = x(t) * h(t) + x(t) * g(t)$

Convolution plays an important role in system analysis and filtering of a signal. Some important applications of convolution are given as follows:

- **Filtering**: Convolution is used in finding the linear filtering of a signal. Here the signal is passed through a filter, that has an impulse response $h(t)$. The output signal $y(t)$ gives us the filtered output of the signal. The filtration is done through convolution.
- **Signal Correlation**: Convolution can be used to compute the cross-correlation between two signals. It can measure similarity between them as a function of time lag.

We used correlation to filter our signal and find some patterns in the input data. The input data was the usage of electricity per minute, so there were fluctuations in the use of energy from time to time. By using a convolution function of a rectangular wave, we found out the rolling mean of the usage of energy was producing a clear sine wave. Most of the times, the sine wave was being produced but some other times, it was just garbage. We tried many convolution signals, from a fundamental frequency to a very low frequency sine wave. But rolling mean gave us the best result.

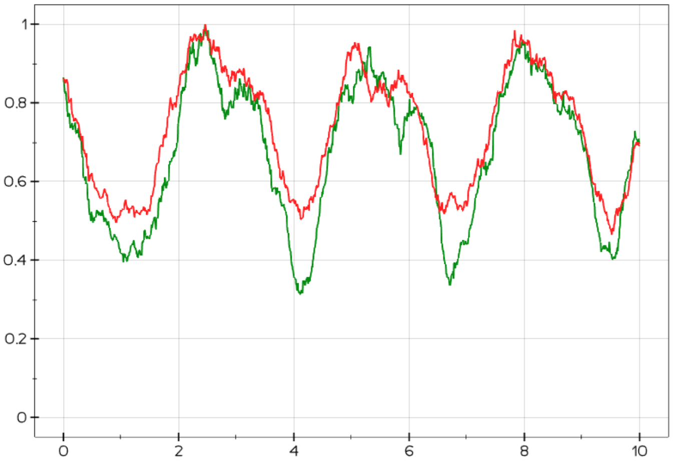

Fig. 5 The figure shows the rolling mean of the original signal and the cleaned signal. Original is the green one.

VI. Curve Fitting

Curve fitting plays an important role in optimizing and finding a polynomial curve that best fits our data. It is used to find a mathematical function that best captures the underlying relation in the data.

Given a set of points x_i and y_i , where the x_i is the independent variable and the y_i are the dependent variable. The use of curve fitting is to find a curve that minimizes the error between the generated polynomial and the values of y_i.

Given below are some common curve fitting methods commonly used in the field of optimizations.

- **Polynomial**: This is a simple and easy technique for modeling a smooth curve that fits our data. It finds a continuous curve that best fit the data. Overfitting can arise when a very high order polynomial approximation is done by the user, so sometimes it is necessary to use it with precautions.

- **Exponential**: It is used to find an exponential growth of a function or an exponential decay of a function. Bacterial growth and population growth is a good example of places where exponential approximations are used.
- **Trigonometric**: These functions are suitable where the data is periodic or we are trying to model a cyclic pattern.

Curve fitting finds it's role in Signal processing, machine learning and STEM. These places generally include an experimental data, so we try to fit the data to the theoretical models and parameter estimation.

We make use of Numpy library and the Scipy library to optimize the curve, scaling to the data and help us predict the usage of the next time. With Numpy, we tried to fit a 20-degree polynomial to the data and gives about 14% total error in calculation. Scipy was used to calculate a sinusoidal curve that best fits the data. Sinusoid curve didn't always work well but the polynomial approximation did a good job at finding a curve that fits our data.The figure below was generated with the help of Numpy polyfit function.

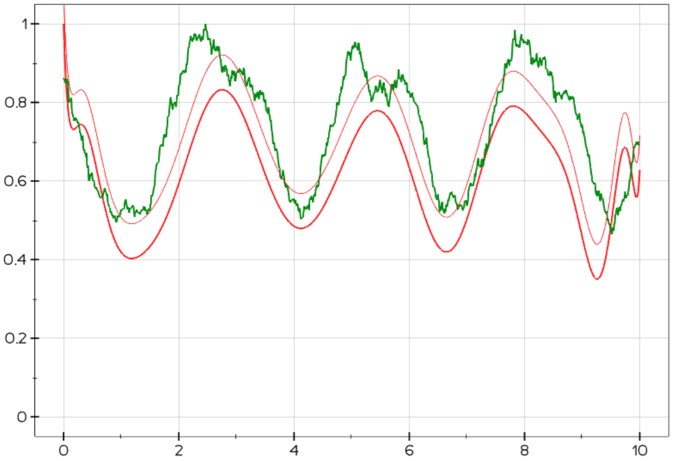

Fig. 6 The figure shows a 20-order polynomial that fits the data Small red line denotes the standard deviation.

Since our polyfit was showing a sinusoid curve for the data, we also tried to fit the data using the following trigonometric formula.

$$y = a * \sin(2 * b * x + c) + d$$

We had to use the numeric value 2 in the curve fitting since our prior value of 1 was not working correctly.The sinusoidal curve that fits the data is shown in the next figure. Everyone should be warned that sinusoid wave was not working all the time. There were many a time when the spikes in the data were not equidistant and therefore shooted very high. Our sinusoid having maxing amplitude of a and max y-intercept of d, could only go to the max value of $a + d$.

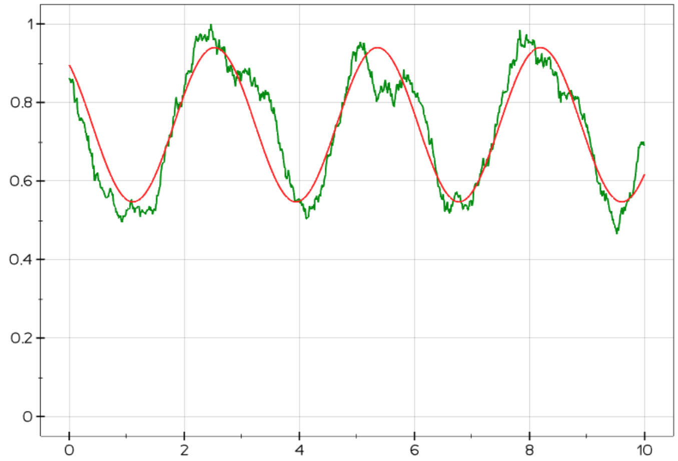

Fig. 7 A sinusoid fitting the consumption.

VII. Machine Learning

After doing all other works, we were left to do the task of linear regression to the data. Linear Regression is a technique of curve fitting that approximates a best fit line to the data. A general linear regression method makes use of Mean Squared Error to calculate the best line to the data. Till this point, we had cleaned the data, removed all the unwanted noise from the data and best of all we had fit a curve to the data.

Let all the original Y-axis are Y_{1i} and the curve that best fits the data has the points to be Y_{2i}. We plotted the points in the fashion (Y_{1i}, Y_{2i}) and the result were quite promising. From a curve, that looked like garbage, we went to a plot that showed that the curve that we plotted gives us a linear relationship between them.

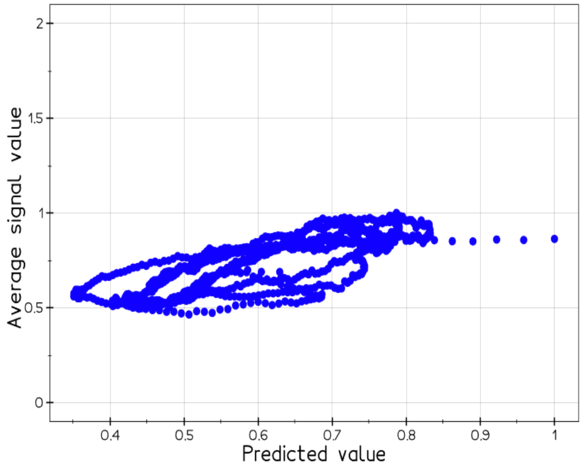

Fig.8 The plot of calculated output and the value generated by rolling mean of clean signal

Since it looked linear, we used Scikit-Learn to fit the data to the line. With the line been made, it showed about 67% accuracy in the test data and about 63% accuracy in the training data. Since linear regression makes use of MSE for error calculation, had we kept the original scale of data, it might have overflown and the line would be wrong. For this reason, we would scale the data down to being in range $(0, 1)$.Overall, we were happy with the result we got.

TABLE I.

Original Rolling Mean value	***Polynomial fit value***	***Sinusoidal fit value***
0.649	0.649	0.664
0.958	0.771	0.934
0.527	0.422	0.595
0.841	0.774	0.923
0.549	0.491	0.565

[a] The table shows the rolling mean, calculated value by polyfit and values calculated by sinusoidal fit.

VIII. References

[1]. Power consumption prediction in urban areas using Machine Learning as a strategy towards smart cities. AJBAR Vol1 (1), 2-22: 11-24, ISSN: 2811-2881

[2]. An overview of electricity demand forecasting techniques, Vol.3, No.2, 2013-National Conference on emerging trends in electrical, instrumental and communication engineering.

[3]. A deep learning architecture for power management in smart cities. Qin Xin a, Mamoun Alazab b, Vincente Garcia Diaz c, Carlos Enrique Montenegro-Marin d, Ruben Gonzalez Crespo e.

[4]. Strategies for predictive power: Machine Learning models in city-scale load forecasting, e-Prime-Advances in Electrical Engineering, Electronics and Energy, Vol.6 December 1023, 1—392

[5]. Machine Learning for modern power distribution systems: Progress and perspectives. Marija Markovic, Matthew Bossart, Bri-Mathias Hodge. Journal of Renewable and Sustainable energy, Vol. 15, Issued 3 May 2023.

[6]. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Energy and buildings, 147, 77-89.

[7]. Data driven prediction models of energy use of appliances in a low-energy house. Energy and buildings, 140, 81-97, Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017).

[8]. Forecasting the Usage of Household Appliances Through Power Meter Sensors for Demand Management in the Smart Grid, A.Barbato, A.Capone, M. Rodolfi, D. Tagliaferri Dipartimento di Elettronica e Informazione Politecnico di Milano, Italy.

[9]. A review on artificial intelligence-based load demand forecasting techniques for smart grid and buildings, Renew. Sustain. Energy Rev.50 (2015) 1352–1372, M.Q. Raza, A. Khosravi,

[10]. Time series forecasting for decision making on city-wide energy demand: a comparative study, in: 2022 International Conference on Decision Aid Sciences and Applications, DASA 2022, 2022, pp. 1706–1710, O. Nooruldeen, S. Alturki, M.R. Baker, A. Ghareeb,

[11]. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system, Mob. Netw. Appl. 26 (1) (2021) 234–25, W. Li.

[12]. Regression Based Peak Load Forecasting Using a Transformation Technique, IEEE Transaction on Power System, Vol.9, pp.1788–1794, 1994, T. Haida and S. Muto.

[13 Short Term Load Forecasting. Proceedings of the IEEE, Vol.75, pp. 1558-1573, 1987,]. G. Gross G. Gross, and F. D. Galiana,

[14]. Uncertainty management in electricity demand forecasting with machine learning and ensemble learning: case studies of COVID-19 in the US metropolitans, Eng. Appl. Artif. Intell. 123 (2023), 106350, M.R. Baker.