Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 23-33, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 23-33, December 2021

Research Article

Analysis of Different Regression Algorithms for the Estimate of Energy Consumption

Halit Çetiner[1*], İbrahim Çetiner[2]

[1*] Isparta University of Applied Sciences, Vocational School of Technical Sciences, Isparta, Turkey, (ORCID: 0000-0001-7794-2555), halitcetiner@isparta.edu.tr
[2] Mehmet Akif Ersoy University, Vocational School of Technical Sciences, Burdur, Turkey, (ORCID: 0000-0002-1635-6461), cetiner51@gmail.com

(First received 10 July 2021 and in final form 15 December 2021)

(**DOI:** 10.31590/ejosat.969539)

ATIF/REFERENCE: Çetiner, H., Çetiner, İ. (2021). Analysis of Different Regression Algorithms for the Estimate of Energy Consumption. *European Journal of Science and Technology*, (31), 23-33.

Abstract

Due to the increasing population, a great increase is observed in the number of different centers such as art, entertainment, and industry. The number of such centers is increasing day by day. These areas are naturally the centers where energy is needed at a high rate. Energy consumption data in the specified areas are increasing day by day. At this point, it has become a difficult problem to meet energy needs in all areas where people live and need energy, in addition to the mentioned centers. To eliminate this difficult problem, it has become a necessity to both meet the energy consumption and ensure the effective use of energy. It is observed that there is an increase in artificial intelligence supported housing systems consisting of electronic devices in order to minimize energy consumption in shelters and residences. Taking into account the increase in environmental factors such as global warming, greenhouse gas emissions, carbon dioxide, chemical solvents, and radiation, studies on the efficient use of energy should be increased. In line with the stated objectives and purposes, the data of the United States regional communications organization PJM Interconnection LLC (PJM) named Dominion Virginia Power (DOM) have been used. This dataset shows the hourly data consumption in Mega Watts of the Asian region. On this dataset, the energy estimation results of the recently popular XGBoost, LSTM algorithms, classical linear regression and RANSAC algorithms were compared.

Keywords: LSTM, XGBoost, RANSAC, Linear regression, energy consumption.

Enerji Tüketim Tahmini için Farklı Regresyon Algoritmalarının Analizi

Öz

Artan nüfusa bağlı olarak sanat, eğlence ve sanayi gibi farklı merkezlerin sayısında büyük bir artış gözlenmektedir. Bu gibi merkezlerin sayıları her geçen gün artmaktadır. Bu alanlar ise doğal olarak enerjiye yüksek oranda ihtiyaç duyulan merkezlerdir. Belirtilen alanlarda enerji tüketim verileri her geçen gün artış göstermektedir. Bu noktada sözü edilen merkezlerin yanı sıra insanların yaşadığı ve enerjiye ihtiyaç duyduğu tüm alanlarda enerji ihtiyacının karşılanması zor bir sorun haline gelmiştir. Bu zor problemi ortadan kaldırabilmek için hem enerji tüketimini karşılamak hem de enerjinin etkili kullanılmasını sağlamak bir zorunluluk olmuştur. Sığınak ve konutlarda enerji tüketimini en aza indirmek için elektronik cihazlardan oluşan yapay zeka destekli konut sistemlerinde artış olduğu gözlenmektedir. Küresel ısınma, sera gazı emisyonları, karbondioksit, kimyasal çözücüler, radyasyon gibi çevresel faktörlerin artması göz önüne alındığında, enerjinin verimli kullanımına yönelik çalışmaların mutlaka artırılması gerekmektedir. Belirtilen amaç ve amaçlar doğrultusunda Amerika Birleşik Devletleri bölgesel iletişim kuruluşu PJM Interconnection LLC'nin (PJM) Dominion Virginia Power (DOM) adlı verileri kullanılmıştır. Bu veri seti, Asya bölgesinin Mega Watt cinsinden saatlik veri tüketimini göstermektedir. Bu veri seti üzerinde son zamanlarda popüler olan XGBoost, LSTM algoritmaları, klasik Lineer regresyon ve RANSAC algoritmalarına ait enerji tahmin sonuçları karşılaştırılmıştır.

Anahtar Kelimeler: LSTM, XGBoost, RANSAC, Doğrusal regresyon, enerji tüketimi.

[*] Corresponding Author: halitcetiner@isparta.edu.tr

1. Introduction

The world population is increasing rapidly. Looking at the literature researches in recent years, it is expected that the world population will approach 10 billion towards 2040 or 2050. (Bahar et al. 2020). One of the main issues that a population increase of this magnitude makes us think about is how energy consumption will change. Consumption will increase as the energy demand increases, depending on the population. According to (Günay 2016), it is not sufficient to explain the increase in energy consumption only by population growth over the years. To understand the relationship between energy consumption data according to (Li 2019) and the factors that affect them, there is a disagreement about whether to use regression analysis or perform analyses over time series. In the studies, it has been observed that estimating the energy consumption from the data collected regularly at certain time intervals gives better results than the regression methods. This result, which was determined, presented a research that allowed the use of time series in the study. It is estimated that two-thirds of the world's population will live in urban areas away from rural areas in the next 30 years (Bahar et al. 2020). Considering that a population the size of the largest city of a country like China is added to the world population every four months, this reveals the true face of population growth.

Despite the stated population growth, it is getting more and more difficult to meet the energy demand by reducing harmful emissions such as CO_2, NO_x. It is reported that the emission values of two years ago exceeded 45% of the emission values of twenty years ago worldwide (Bahar et al. 2020). It is obvious that all kinds of energy, including clean electricity, will be needed more in the coming years for the reasons stated. Electricity demand is expected to double compared to previous years. In order to meet the energy demand by reducing the harmful emissions resulting from the increasing energy consumption due to the increasing population, clean energy production supported by accurate, conscious and smart systems is required When history is examined backward, it is seen that human beings have developed structures that are far from environmental effects to lead a more comfortable and easy life. The construction of these structures has been going on for centuries. Although people give these buildings different names such as workplace, office, house, student house, holiday home, their main task is to ensure that people live and work away from all kinds of influence unique to them. Due to the reasons stated, a large part of human life is spent in the specified buildings or residences. It is necessary to increase energy efficiency and reduce energy consumption in homes where this much time has passed. (García et al. 2021), evaluates people consuming energy in three different categories. In the first category, there are people who continue their normal consumption during their inactivity. The second category includes people who significantly reduce their consumption during the period of inactivity and continue to reduce their consumption after the restriction. The third category represents the group of people who reduced their consumption during the restriction period and increased their consumption after the restriction. When study (García et al. 2021) is examined, it is seen that the most accurate analysis results are obtained by analysing hourly energy consumption. Similar to this study, there are also studies that want to reduce the energy consumption in shelters and eliminate the disposal of harmful wastes to the environment (Guo et al. 2021).

It is becoming more and more important every year to abandon the production of classical buildings and to create scientifically based low energy-consuming houses and shelters. It is seen that studies are carried out by applying high investments in line with the goals and objectives stated in different parts of the world. They conducted studies investigating the impact of renewable energy consumption on reducing air pollution in Latin America over a long period from 1990 to 2016 (Koengkan, Fuinhas, and Silva 2021).

(Moreno et al. 2014) describes the work they developed for energy saving. In their study, they tried to determine which parameters should be taken into account by building designers in order to reduce energy consumption. As a result of their studies on effective energy use, they revealed that human life is closely related to energy consumption. They tried to identify the most relevant parameters that need to be controlled for efficient energy use. Environmental changes may be effective in determining the relationship between the correct and conscious use of energy in homes and human energy consumption time. It is stated that global changes due to environmental changes are the result of incorrect energy use. Climate change is said to have become a global problem due to environmental changes (Change 2014). It is stated that global warming due to environmental changes will disrupt floods, droughts, and food production. For this reason, it is stated that people can migrate to safer areas as a temporary solution. However, there is no guarantee that this problem will not occur in the target region. Exposure to excessive and irregular heat waves beyond what is normally expected can impair human health. Furthermore, it causes diseases to spread very quickly (Enn 2015).

According to (Change 2014), energy production accounts for a quarter of the total worldwide emissions value. This rate shows that the necessity of efficient, conscious, accurate and technology-supported energy consumption to be sustainable is becoming more important day by day. Academic studies are needed to prevent the extra energy use of all energy-consuming devices. One of the main purposes of studies such as estimating and classifying energy consumption data in buildings or shelters is to significantly improve residential performance, reduce harmful effects outside the living space, and guide people to more conscious energy consumption.

The contributions of this study, which was made for the purpose of energy consumption estimation, to the literature are presented below, respectively.

- It is an application of data normalization since the energy consumption data is a large data set.
- XGBoost, Linear regression and RANSAC algorithms were run on both normalized and normal raw data in order to accurately interpret the effect of data normalization on energy consumption estimation.
- Performance measurement results obtained as a result of experimental studies on both normalized and normal raw data are given comparatively according to the traditional criteria R^2 score, MSE, RMSE, MAE, MAPE.
- When the results are examined, the highest test success rate on normal raw data is 0.67 R^2 score. In the normalized data set, it is seen that a success rate of 0.9738 was obtained. This rate is quite high compared to the results obtained on normal raw data.

The remainder of the article is planned as follows. In the second chapter, similar studies in the literature are given. In the third chapter, the data set used and the methods by which we analysed this data set are explained. In the fourth chapter, the results obtained as a result of the experimental analysis studies are given. The results are discussed in the last section. Furthermore, information was given on the projects to be realized in the future based on this study.

2. Related Works

Predicting future energy consumption in shelters and residences is important for the correct use of energy and for governments to make energy planning. Developing effective and efficient energy consumption forecasting models is valuable for energy planning. For this reason, it is important for researchers to examine the data expressed as time series containing hourly energy data using statistical or machine learning techniques. It is not sufficient to examine time series data alone. It also needs to be examined quickly, accurately and effectively. It is seen that there are different studies in the literature that present a comprehensive analysis of energy consumption data consisting of time series (Deb et al. 2017). In the mentioned article, it is mentioned that time series should not be evaluated within a single time, but backwards with other time units (Deb et al. 2017). In this study, XGBoost, LSTM, RANSAC, Linear algorithms, which have gained popularity in recent years, apart from algorithms such as SVM and ANN that analyze popularly used time series data, are used. In the studies in the literature, it is seen that analyzes are carried out on the energy data on a regional, country basis or in a single research center. When the literature summaries are examined, it is seen that data sets consisting of time series collected at certain time intervals are used in the analysis(Bhati, Hansen, and Chan 2017) are looking for ways to reduce energy consumption in residences in Singapore. (Moletsane et al. 2018) analyzed the energy consumption data from a house controlled by more than one electronic system using linear regression method. These data include data on energy consumption between 2014 and 2016 within the scope of the housing project controlled by electronic systems. A linear regression model was used to analyze these consumption data. (Arghira et al. 2012) attempts to estimate the energy consumption of different appliances in the house. It describes a method that tries to estimate the next day's electricity consumption based on today's electricity consumption. (Edwards, New, and Parker 2012) analyzed hourly energy consumption data with seven different machine learning algorithms. It has been tried to determine which of the seven different algorithms used to estimate the energy consumed at a future date is the most successful. (Lü et al. 2015) offer a new methodology to solve the problem of inhomogeneity in energy modeling of houses. This method proposed a physical and statistical-based mathematical approach to generate homogeneity to improve prediction accuracy. (O'Neill and O'Neill 2016) developed a segmented model to determine residential energy performance. (Roldán-Blay et al. 2013) tried to estimate energy consumption data with an artificial neural network using an hourly data set.

(Bagnasco et al. 2015) tries to estimate the energy consumption of a medical clinic. They model this prediction process using an artificial neural network. It tried to estimate the electricity consumption by using the data in the attribute table created with the day definitions such as time of day, weekend, and holiday. (Antanasijević et al. 2015) propose a new neural network-based approach that predicts the relationship between harmful greenhouse gas emissions and energy consumption. (Jung, Kim, and Heo 2015) developed a least squares support vector machine to determine the energy consumed daily. (Protić, Fathurrahman, and Raos 2019) developed a mathematical model to determine future energy consumption based on energy data used in the Republic of Serbia. These developed models are compared with the artificial neural network model. (Li 2019), states that there is a close relationship between economic development, population, industrial relations, and energy consumption. It has been determined that many researchers use regression method to solve the relationship between these variables and energy consumption. In their study, they tried to predict the energy density of China using an LSTM-based neural network model based on both research groups. As a result of the experimental studies they carried out, it was concluded that the estimation results using the time series were much more successful than the other regression analysis. (Ilbeigi, Ghomeishi, and Dehghanbanadaki 2020) tried to develop a method to reduce energy consumption in houses and shelters. For this, a new residence suitable for research was designed. Input parameters used in energy consumption in this house have been determined. Monthly electricity consumption bills were collected for one year in the input parameters. These determined parameters were tested to be trained and tested using the multilayer artificial neural network model. It is reported that when creating a model in artificial neural networks, 70% of the data set is divided into training, 15% validation, and 15% testing. It is stated that the increase in the number of people living in houses and shelters increases the demand for energy. They concluded that the most effective factor after the number of occupants is the thickness of the outer wall of the house. These two factors declare that they are the most influential factors in energy consumption.

3. Material and Method

In this study, energy consumption data recorded at equal time intervals are used. Regular recording of energy consumption data at certain time intervals is called a time series. According to (Deb et al. 2017), the time series consists of two different categories. In the first category, it focuses on how time series are formed, their structure, and ground state. In the second category, it focuses on the kind of meaning it will make on the data obtained for the future. Time series analysis is used in many different areas from the change in the amount of diesel use during the epidemic to the population planning of the countries. It is known that there has been an increase in studies on monitoring, tracking and estimating energy consumption data in houses and shelters recently. Consistent with this information, energy consumption time series data from an eastern state region in the USA will be used in the study. The dataset consists of two columns. The first column contains date and time information and the second column contains the energy consumed. Although the data until January 2006 and 2015 were used as training data, the rest of the data were used as test data. After the training data are trained with different regression algorithms, it will be tested to be estimated on the test data that have never been seen before. The training and test data segments are shown in Figure 1.

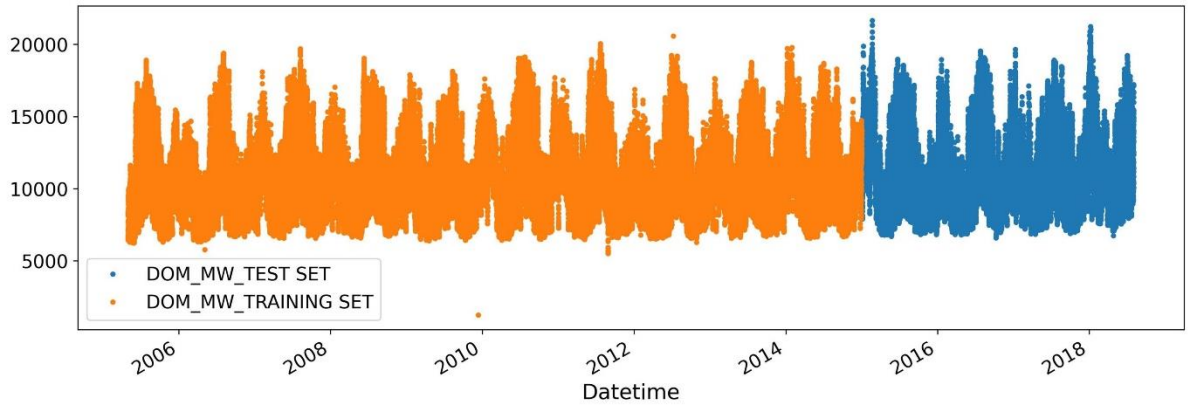

Figure 1. DOM MW training and test set

Features directly affect results in estimation and classification problems. For this reason, detailed time information extracted from raw data is used as an attribute in this article, since it provides a detailed detail about energy consumption. Table I shows 4 representative examples of the data set consisting of the mentioned time information and energy consumption data, from the data it starts to the data it ends.

Table I. Features used in energy consumption estimation

Datetime	Hour	Day of week	Quarter	Month	Year	Day of year	Day of month	Week of year	DOM_MW
2005-12-31 01:00:00	1	5	4	12	2005	365	31	52	9389
2005-12-31 02:00:00	2	5	4	12	2005	365	31	52	9070
….	…	…	….	…	…	….	….	…	…
2018-08-02 23:00:00	23	3	3	8	2018	214	2	31	12390
2018-08-02 23:00:00	23	3	3	8	2018	214	2	31	11385

In the first experimental studies, using XGBoost, RANSAC (Random Sample Consensus), Linear regression methods, eight different attributes such as day of the week, day of the month, day of the year were extracted from the date data of the first column and analyzed without any normalization function. Second, in experimental studies, the same data set was normalized using XGBoost, RANSAC, linear regression methods, and LSTM time series deep learning algorithms. Only the original data were used in the normalized data set. Unlike the first, the dataset used was not expanded by extracting features from the original data. In order to show the accuracy of the studies in the studies, the R^2 value is calculated to measure the relationship between the consumption result obtained by the algorithm and the actually consumed consumption result.

The energy data estimated in this way are compared with the actual energy consumed. In this comparison, the R^2 value is necessary to understand the relationship between the specified parameters. It is known that R^2 values, which get a result close to 1, have a great meaning in showing the relationship between them. For these reasons, the R^2 scoring value was used to compare the results of the used algorithms.

3.1. The Random Sample Consensus (RANSAC)

It is an algorithm put forward by the RANSAC algorithm (Fischler and Bolles 1981). It is a computational approach that deals with extreme values in input data. The RANSAC algorithm, which is used as a general estimation approach, is used to accurately predict the selected model parameters. It is a method that provides predictive solutions using the minimum number of data regions needed to predict certain model parameters. The most distinctive feature of this method is that it uses as much data as possible to obtain initial recommendations. After obtaining the initial suggestions, it creates a prediction set with consistent data structures by removing the extreme values.

$$1 - p = (1 - \mu^m)^N \quad (1)$$

In Equation (1), m shows the least number of points in the process repeated N times. μ represents the probability that any data region is the beginning. $v = 1 - \mu$ represents the probability of following an outlier (Derpanis 2010). The number of N iterations shown in Equation (2) must be high enough to ensure that the probability p is that one of the randomly generated sample sets is not outlier (Derpanis 2010).

$$N = \frac{\log(1 - p)}{\log(1 - (1 - v)^m)} \quad (2)$$

3.2. XGBoost

XGBoost is a method introduced by (Chen and Guestrin 2016). It is a method of transforming the weak into the strong in scalable learning based on the method named Gradient boosting machine (Friedman 2002; Zhou et al. 2019). While performing these, machine learning methods have features such as excessive learning, increasing the predictive power, fast processing power and managing empty data. Additionally, XGBoost prioritizes tree depth (Mitchell and Frank 2017). In the first step of the algorithm's operation, a certain value is created. The error value resulting from the operations is determined by looking at the difference between the actual value in the data set and the value as a result of the operations. Afterwards, similarity scores are calculated for each branch in the prediction trees. In this way, it is tried to find the best estimation result among the tree branches. After the best and strongest ones are determined, pruning is performed on weak branches. In the pruning process, a value called gamma should be selected. The gamma value is used in pruning as the threshold value. Weak branches in trees are destroyed according to the specified threshold value. Increasing the defined threshold value ensures that only the strongest branches remain in the tree and prevents overfitting. As the λ value increases, the similarity score calculated in the branches decreases. There is an inverse relationship between the value of λ and overfitting. As one increases, the other decreases. A high value of λ means that there must be many cycles to make the correct prediction. Cyclic operations mean that the operations will continue until the error reaches a certain value or a specified tree number criterion.

XGBoost partitions each value in the dataset. It works according to partitions. As the number of partitions increases, it will look at the smaller sub-partitions, resulting in better predictive results. Since this will increase the computational cost, the learning process will increase. In machine learning techniques, empty data in the data set is filled or the fields with empty data are removed from the data set. XGBoost algorithm can work with null values. Initially, the error values obtained using the default predictive value are assigned to the blank data. In the next assignment, the blank data is placed in all different branches and the earnings score is created. In which case the score is high, null values are placed in those branches.

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k} \Omega(k) \tag{3}$$

In equation (3), a general loss function is defined to measure the difference between the actual value y_i and the predicted value $\hat{y}_i$ for each sample i of the first term l (Hu et al. 2021). One of the evaluation functions used in this study is the mean square error. The Ω symbol represents the task of scoring and adjusting the complexity of each model.

$$\Omega(k) = \gamma.T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

In Equation (4), T shows the number of leaves found in decision trees. w_j represents the point value of each j tree leaf. γ and λ are the variables used to put the penalty score on a certain scale.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + q_t(x_i) \tag{5}$$

In Equation (5), the output of the i_{th} tree in the t_{th} loop is represented by $\hat{y}_i^{(t)}$. q_t is the sample to be added in the t_{th} loop. The specified optimization-based algorithm is transformed into a target function by Equation (6).

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + q_t(x_i)\right) + \gamma.T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

Then, as shown in Equation (7), a second-order Taylor expansion is performed in order to get closer to the target function and to make the optimization process easier.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{N} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i q_t(x_i) + \frac{1}{2} h_i q_t^2(x_i) \right] + \gamma.T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{7}$$

In equation (7), g_i and h_i are used to represent the first and second order derivatives of the objective objective function. When the $(t-1)$ loop is complete, $l\left(y_i, \hat{y}_i^{(t-1)}\right)$ is expressed as a fixed. This constant can be considered absent during optimization.

$$\hat{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma.T \tag{8}$$

In Equation (8) I_j shows that the sample set corresponding to j in the tree leaf node. In this equation, each sample is rewritten to correspond to each tree leaf node. In this way, the optimization of the targeted objective function is converted to a quadratic equation (Hu et al. 2021). At the same time, this function creates a decision tree. The depth dimension of each tree is defined by defining a certain threshold value. With the feature added to this function, the excessive compatibility problem is overcome.

3.3. Linear Regression

The XGBoost algorithm does not have the feature of filling the empty data with error. According to the XGBoost algorithm, it is an algorithm based on history. Linear regression is a linear model that represents the linear relationship between a single input parameter (x) and a single output parameter (y). The result of the input variable y can be calculated using a linear model of (x) values. As in Equation (9), the linear regression formula can be created simply by adding the bias coefficients.

$$y = B0 + B1 * x \tag{9}$$

3.4. LSTM (Long Short Term Memory)

Recurrent neural networks (RNN) are basically defined as neural networks that use information from the previous step to a forward transition process (Ketkar and Santana 2017). When applied to data forms where the input data is in the form of a time series, success values can be high. However, it loses a lot of success due to the fact that the data can be kept for a short time.

If the time series we want to process consists of long data, it will be difficult to carry information while each long data in these series is transferred to the next step. For this reason, RNN algorithms can cause data loss in the processing of long-dimensional data. Weights need to be updated in any neural network during training. This weight update must also be done backwards by gradients. It is undesirable for the gradient values updating the weights to get too close to zero during this process. When these values are close to zero, the learning rate may be insufficient. When the gradient value of the RNN algorithm approaches zero, it finishes the learning process and forgets the data in longer sequences and creates short-term memories.

Long short-term memory (LSTM) is defined as a customized version of the repetitive neural network model RNN (Li 2019). RNN and LSTM can be defined as an extension or extension of the classical artificial neural network model. While deep neural network models have thousands of hidden neural network layers, classical neural networks have no more than a few hidden neural network layers.

LSTM structures have structures called memory blocks in repetitive layers to solve the problem of error values that suddenly reset or become excessively high during training (Wang, Du, and Wang 2020). Figure 1 shows the block diagram of special units called memory blocks. There are three different gates in the LSTM method, which is proposed to eliminate the deficiencies in the RNN method. Doors with collision feature that temporarily save sequences are shown in Figure 1. Doors are structures with Forget, Input and Output tasks respectively. At the forget gate, it controls which information will be protected and which information will be destroyed. Information from hidden states and other input data is given as input to the activation function. Normalization of the data given as input to a value between 0 and 1 is provided. Here, if the value is close to 0, it semantically indicates that this data should be forgotten. In the other case, if the data value is close to 1, it means data protection. At the entrance gate, the previous hidden state and input data are inserted into the sigmoid activation function to update the cell state. As with the forget gate, the output values close to 1 means that the data is important. In addition, data is then transmitted to the tanh activation function as a hidden state and input. In this way, it is ensured that the network is organized. Then, the first activation output is multiplied by the tanh activation output. The sigmoid output values decide what information is important to retain from the tanh activation function output. The output gate checks for hidden states containing information about previous inputs. As with other gates, the previous hidden state and input values are inserted into the sigmoid function. Afterwards, the updated hidden state becomes an input to the tanh activation function, as in other gates. The tanh output is multiplied by the sigmoid output to detect important information which information is to be protected. The result obtained is defined as the latent state (Wang, Du, and Wang 2020).

The sigmoid function used in the gates is shown in Equation 10.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

In Equation (11), (13) and (15), i_t, f_t, o_t represent the input, forget and output gates, respectively. The $\widetilde{C}_t$ in Equation (12) is the intermediate value used during the calculation.

$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_t) \tag{11}$$

$$\widetilde{C}_t = tanh(W_z X_t + U_z h_{t-1} + b_z) \tag{12}$$

$$f_t = \sigma\left(W_f X_t + U_f h_{t-1} + b_f\right) \tag{13}$$

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \tag{14}$$

$$o_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \tag{15}$$

$$h_t = o_t * tanh(C_t) \tag{16}$$

W_i, U_i, W_z, U_z, W_f, U_f, W_o and U_o expressions in Equations (11), (12), (13), (15), respectively, indicate weight matrices, and b_t, b_z, b_f and b_o represent bias vectors. h_t in Equation (16) and h_{t-1} in Equation (11), (12), (13) and (15) are output values at previous time t and $t-1$, respectively. The hyperbolic tangent function used in LSTM gates is shown in Equation (17).

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{17}$$

A Dropout layer is successively added to each LSTM layer. This layer helps to prevent the undesirable overfitting by ignoring randomly selected neurons during training. Thus, memorization due to excessive learning in neurons is prevented.

4. Experiment Results

Using the settings given in Table II, the LSTM algorithm was run on both the training and test datasets on the normalized data.

Table II. LSTM setting parameters

Parameter	Value
Layers	3
Loss	Mean squared error
Optimizer	Adam
Epochs	7
Batch size	56
Activation	ReLU

$$x_{new} = \frac{x - x_{min}}{x - x_{max}} \tag{18}$$

Figure 3 shows not normalized data. Normalization is rescaling the data in the original range so that all values are in the new range of 0 and 1. All obtained features are scaled to the range of 0-1 using Equation (18) in the normalization process. Normalization allows us to know or accurately predict the minimum and maximum observable values. These values can be estimated from the available data. Figure 4 shows the normalized data. Five of the classically most common measures used to measure the accuracy of continuous variables were used to evaluate the results. These measurement values were determined based on studies in the literature.

The R^2 measurement value is used for battery estimation of uninterruptible power supplies in the operation of (Avkıran et al. 2020). (Özen, Saraç, and Koyuncu 2021) indicates that the outbreak uses Prophet, Polynomial Regression, ARIMA, Linear Regression, and Random Forest algorithms to predict the number of confirmed cases in the United States. The performance rates of these algorithms were determined by using MAPE, RMSE performance measurement values. In their study, (Şahin, Oktay, and Konar 2020) used a model that predicts thrust from propeller

and engine information. They used MSE measurement values to measure the performance results of this model. The prediction performances of the algorithms were measured using the measurement values given in Equations 19, 20, 21 and 22.

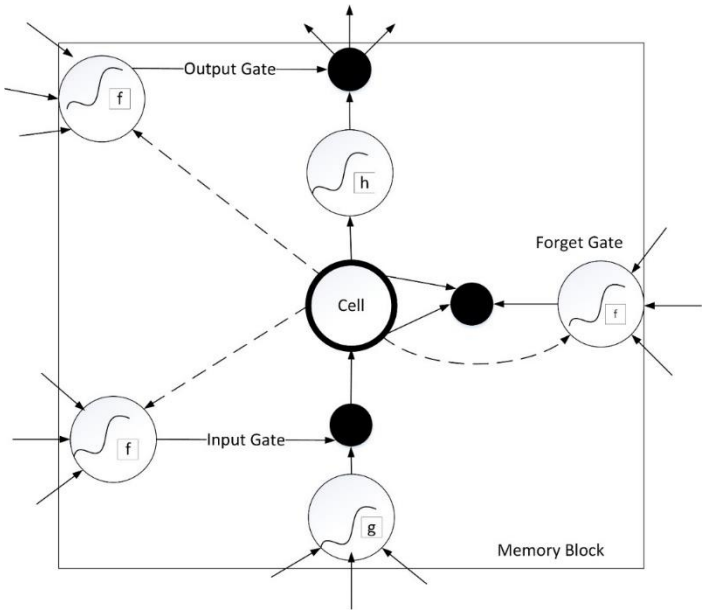

Figure 2. Memory block structure of LSTM architecture (Wang, Du, and Wang 2020)

The mean square error (MSE) is defined as the mean squared error, that is, the difference between the value estimated by the algorithm used and the original value in the data set used.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{Y}_i\right)^2 \quad (19)$$

The root mean square of error (RMSE) represents the square root of the second sample moment of differences between the result obtained by the tested algorithm and the original value in the data set, or the quadratic mean of these differences.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Y_i - \hat{Y}_i\right)^2} \quad (20)$$

The mean absolute error (MAE) determines the average size of the errors in the results obtained by the algorithm under test.

$$APE = \frac{100}{m}\sum_{i=1}^{m}\left[\frac{Y_i - \hat{Y}_i}{Y_i}\right] \quad (21)$$

The mean absolute percent error (MAPE) is often used to measure the accuracy of predictions in regression and time series models. If there is zero between the actual values, MAPE cannot be calculated because there will be division by zero.

$$R^2 = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum(Y_i - \bar{Y})^2} \quad (22)$$

In equations (19), (20), (21) and (22), $\hat{Y}$, $\bar{Y}$ show the predicted result and mean value of Y value, respectively. The R^2 score is the coefficient that shows how well the estimated values are in agreement with the original value. It can be said that the higher the result, the more successful the result.

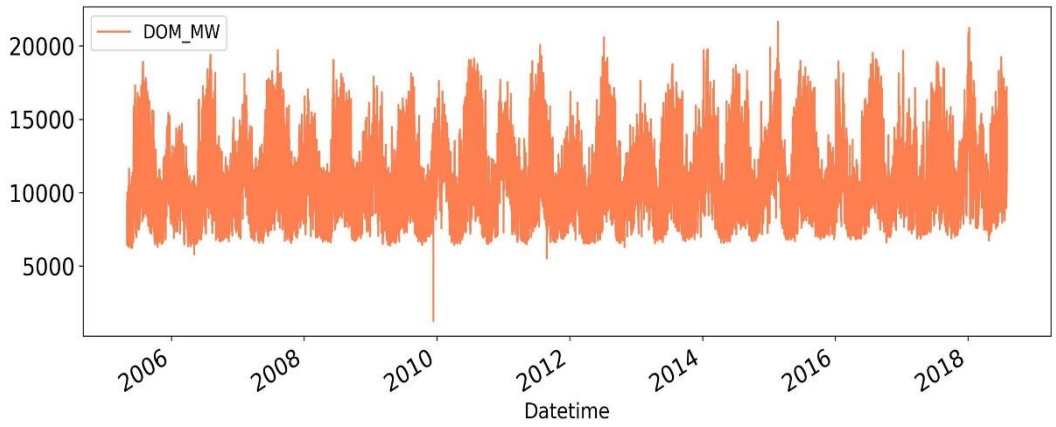

Figure 3. DOM_MW hourly power consumption data before normalization

The results obtained in Table III show the results obtained by using the hour, day of the week, month, year, day of the year, day of the month, week of the year features in the extended data set without normalization. Details of the specified features are given in Table I. Among the XGBoost, LSTM, Linear Regression, and RANSAC algorithms, only the LSTM algorithm has a condition to work on normalized data. There is no such requirement in other algorithms. Apart from the specified constraint, it can be found from the energy estimation without any constraint or condition in the LSTM algorithm.

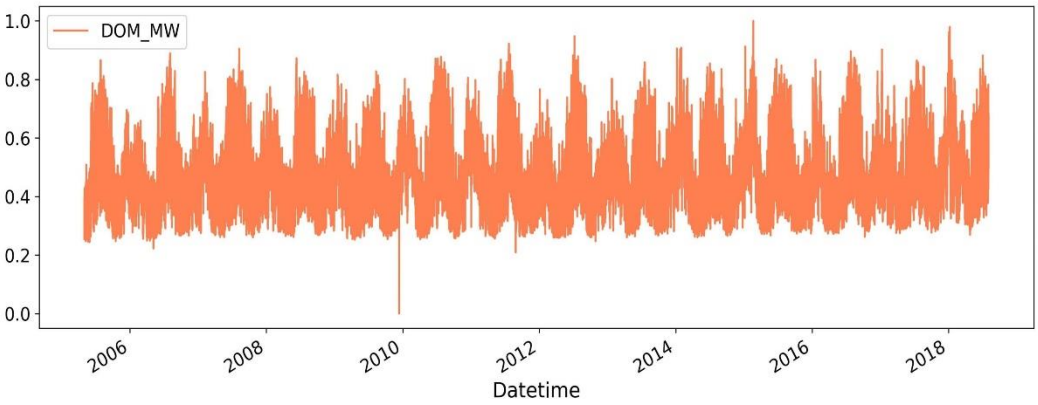

Figure 4. DOM_MW hourly power consumption data - after normalization

The results obtained in Table IV are the results obtained by normalizing the energy data to the 0-1 range, except for the time data in the first column of the data set. Only the original data were used in the results in Table III. There is no feature extraction from the original data.

Table III. Statistical results obtained without normalization

Algorithm	R^2 Score	MSE	RMSE	MAE	MAPE
Linear (Train)	0.28 59	3073545 9.1477	5553.02 29	4193.37 40	14.01 41
Linear (Test)	0.22 76	3517129 5.3672	5771.97 46	4588.82 27	15.90 20
Ransac (Train)	0.14 07	4707227 0.2538	6860.92 34	5301.94 89	16.88 57
Ransac (Test)	0.04 41	4301664 2.9497	6558.70 74	5188.48 13	17.67 47
XGBoost (Train)	0.79 62	9001379. 2930	2883.51 79	2224.85 59	7.259 8
XGBoost (Test)	0.67 44	1496091 3.3057	38736.4 306	2857.21 69	9.510 8

Looking at the results in Table III and Table IV, it was concluded that the normalization step should be done in the data set used for estimation. For this purpose, the data set to be used for regression was scaled to the range of 0-1 using Equation (18).

Table IV. Results obtained on the normalized data set

Algorithm	R^2 Score	MSE	RMSE	MAE	MAPE
Linear (Train)	0.9758	0.0003	0.0183	0.0115	0
Linear (Test)	0.96529	0.0005	0.0230	0.0130	2.6455
Ransac (Train)	0.9709	0.0004	0.0201	0.0116	0
Ransac (Test)	0.9567	0.0007	0.0256	0.0131	2.7165
XGBoost (Train)	0.9930	0.0001	0.0092	0.0060	0
XGBoost (Test)	0.9720	0.0004	0.0199	0.0102	2.0420
LSTM (Train)	0.9939	0.0001	0.0092	0.0060	0
LSTM (Test)	0.9738	0.0004	0.0199	0.0102	2.0302

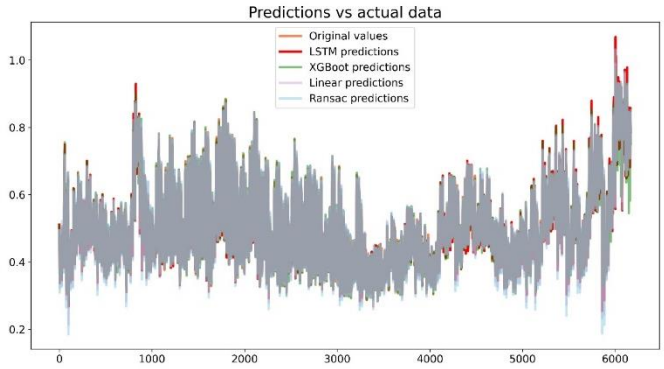

Figure 5. Prediction results made with LSTM, XGBoost, Linear, and RANSAC models

Figure 5 shows the estimation results obtained by the LSTM, XGBoost, Linear, and RANSAC methods. When we look at the results in Table IV, the results obtained are satisfactory. The differences between the estimates made with the specified models and the actual values are given in a table.

According to the table, it is seen that the prediction values of the model are very close to the true value. In Figure 6, it is seen that the decrease in training and test graphics decreases as the number of epochs increases. As can be seen in Figure 6, the test loss ratio performs the estimation process with a loss even less than the training loss ratio. In Figure 5, it is difficult to make an accurate interpretation because the time step in the x-axis of the graphs is close to the number 60000.

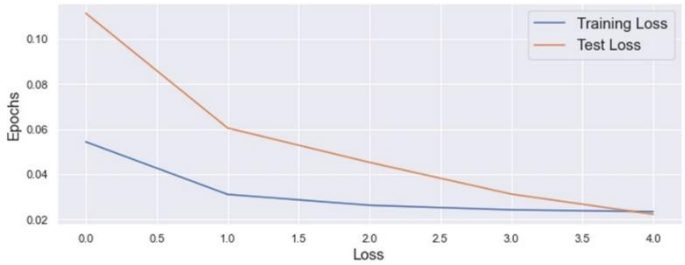

Figure 6. LSTM model training and test loss graph

For this reason, the result of the LSTM algorithm estimation, which provides the most successful model, is shown in Figure 7.

Figure 7 shows the results of the LSTM algorithm, which gave the highest success rate among the algorithms tested. The LSTM algorithm is a successful algorithm at the point of time series. When the results obtained within the scope of this study are examined, it is seen that the stated opinion is supported by the figures. In the LSTM architecture, the dropout process was performed after each layer creation. Just before the completion of the LSTM model creation, the dense layer was added.

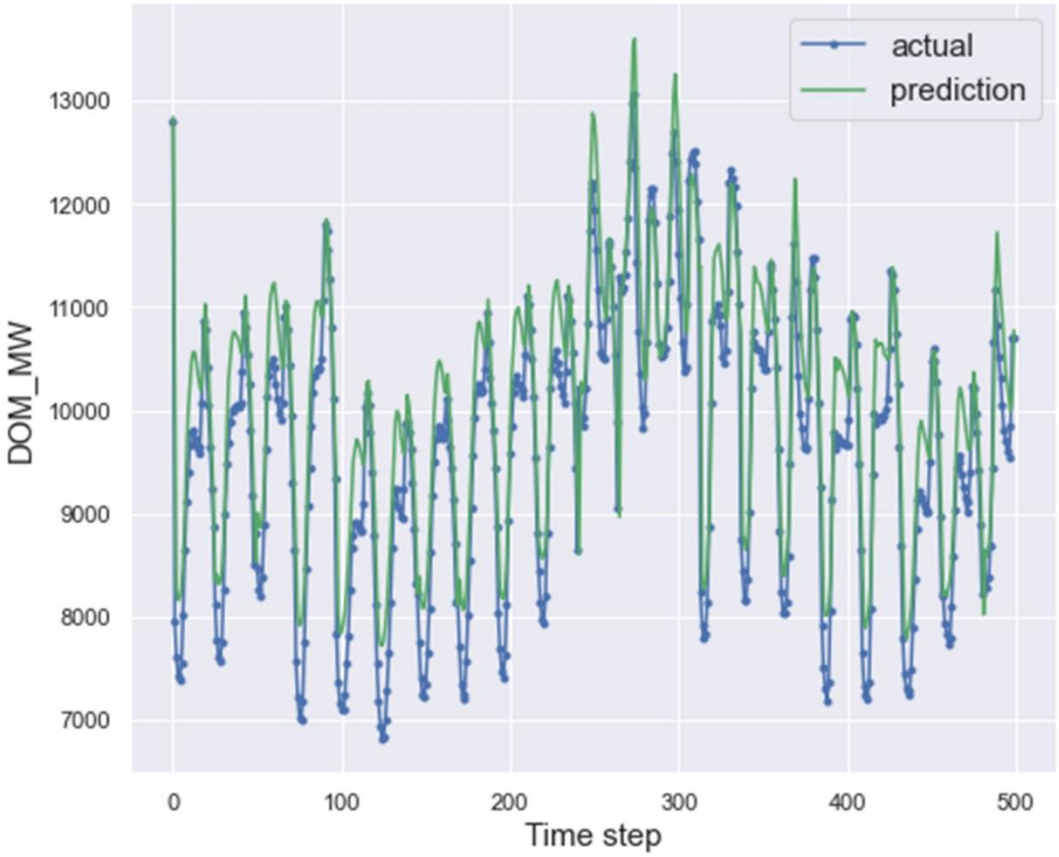

Figure 7. Predictions made by the LSTM model

5. Conclusion

Efforts are being made to facilitate future political and technological solutions for energy consumption in shelters and residences. In this sense, it is suggested that the structure of existing algorithms should be changed in order to calculate energy consumption accurately. In addition, it is mentioned that the algorithms data used by the developed algorithms should be normalized and optimized (Wei et al. 2018). Apart from these, the importance of the data sets used by the algorithms is gaining more and more meaning every day. Researchers recommend that smart meters, which enable high-particle measurement in a used electricity data set, be preferred for data collection. Datasets consisting of correct data are one of the important prerequisites for an accurate analysis. The time series consisting of hourly analyzes was used in the data set used in line with the suggestion of García et al. 2021.

In the development of this study, (Carvalho et al. 2021) studies that examine energy consumption regionally on a country basis and (García et al. 2021) studies that examine individual energy consumption on the basis of behavior were taken into consideration. It is seen that it is important to examine the events from many different windows in the energy consumption analysis studies to be made based on these studies. In this study, consumption data was examined in only one region. The analyzes were carried out without considering environmental effects and an extraordinary situation such as an epidemic. With this study, it has been provided to develop a study that can be a basis for examining the change in energy consumption data.

In the regression processes, the data should be normalized. It is seen that the R^2 scoring values of the algorithms used are close to each other. In this study, although other measurement models are given in a table, R^2 scoring is taken as a basis in the evaluation phase. According to the R^2 scoring values, the LSTM algorithm has been shown to have one-point higher accuracy than other

models. All regression algorithms used in this sense can be used in the energy consumption data set. Excluding holidays and public holidays will yield better results for better results. The LSTM algorithm took more time during the training phase than other algorithms. Here, faster regression operations can be performed with LSTM by extending the dataset over the date column and minimizing the attributes unrelated to the feature selection. Different results are obtained for different parameters for LSTM. In predictive modeling, the model parameters with the best results were used. As a result of various experimental studies, the LSTM tuning parameters were obtained. New algorithms can be developed for the automatic detection of these values. This study has also revealed an important study in terms of forming the basis of the automatic LSTM parameter determination algorithm, which will be used in the automatic parameter determination in the future.

References

Antanasijević, Davor, Viktor Pocajt, Mirjana Ristić, and Aleksandra Perić-Grujić. 2015. "Modeling of Energy Consumption and Related GHG (Greenhouse Gas) Intensity and Emissions in Europe Using General Regression Neural Networks." Energy 84 (May): 816–24. https://doi.org/10.1016/j.energy.2015.03.060.

Arghira, Nicoleta, Lamis Hawarah, Stéphane Ploix, and Mireille Jacomino. 2012. "Prediction of Appliances Energy Use in Smart Homes." Energy 48 (1): 128–34. https://doi.org/10.1016/j.energy.2012.04.010.

Avkıran, Metehan, Gül Vedat, Savaş Şahin, and İbrahim Tanağardıgil. 2020. "Data Acquisition Module Design for Remote Monitoring of Uninterruptible Power Supply and Regression Models of Battery Life." Avrupa Bilim ve Teknoloji Dergisi, 279–83.

Bagnasco, A, F Fresi, M Saviozzi, F Silvestro, and A Vinci. 2015. "Electrical Consumption Forecasting in Hospital Facilities: An Application Case." Energy and Buildings 103 (September): 261–70. https://doi.org/10.1016/j.enbuild.2015.05.056.

Bahar, Nur H.A., Michaela Lo, Made Sanjaya, Josh Van Vianen, Peter Alexander, Amy Ickowitz, and Terry Sunderland. 2020. "Meeting the Food Security Challenge for Nine Billion People in 2050: What Impact on Forests?" Global Environmental Change 62 (May): 102056. https://doi.org/10.1016/j.gloenvcha.2020.102056.

Bhati, Abhishek, Michael Hansen, and Ching Man Chan. 2017. "Energy Conservation through Smart Homes in a Smart City: A Lesson for Singapore Households." Energy Policy 104 (May): 230–39. https://doi.org/10.1016/j.enpol.2017.01.032.

Carvalho, Monica, Danielle Bandeira de Mello Delgado, Karollyne Marques de Lima, Marianna de Camargo Cancela, Camila Alves dos Siqueira, and Dyego Leandro Bezerra de Souza. 2021. "Effects of the COVID-19 Pandemic on the Brazilian Electricity Consumption Patterns." International Journal of Energy Research 45 (2): 3358–64.

Change, Climate. 2014. "Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.[Core Writing Team, RK Pachauri and LA Meyer." IPCC, Geneva, Switzerland.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–94.

Deb, Chirag, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. 2017. "A Review on Time Series Forecasting Techniques for Building Energy Consumption." Renewable and Sustainable Energy Reviews 74: 902–24.

Derpanis, Konstantinos G. 2010. "Overview of the RANSAC Algorithm." Image Rochester NY 4 (1): 2–3.

Edwards, Richard E, Joshua New, and Lynne E Parker. 2012. "Predicting Future Hourly Residential Electrical Consumption: A Machine Learning Case Study." Energy and Buildings 49 (June): 591–603. https://doi.org/10.1016/j.enbuild.2012.03.010.

Enn, Rosa. 2015. "Impact of Climate Change and Human Activity on the Eco-Environment. An Analysis of the Xisha Islands." Island Studies Journal 10 (2): 263–64.

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." Computational Statistics & Data Analysis 38 (4): 367–78.

García, Sebastián, Antonio Parejo, Enrique Personal, Juan Ignacio Guerrero, Félix Biscarri, and Carlos León. 2021. "A Retrospective Analysis of the Impact of the COVID-19 Restrictions on Energy Consumption at a Disaggregated Level." Applied Energy 287: 116547.

Günay, M Erdem. 2016. "Forecasting Annual Gross Electricity Demand by Artificial Neural Networks Using Predicted Values of Socio-Economic Indicators and Climatic Conditions: Case of Turkey." Energy Policy 90: 92–101.

Guo, Siyue, Da Yan, Shan Hu, and Yang Zhang. 2021. "Modelling Building Energy Consumption in China under Different Future Scenarios." Energy 214 (January): 119063. https://doi.org/10.1016/j.energy.2020.119063.

Hu, Liyang, Chao Wang, Zhirui Ye, and Sheng Wang. 2021. "Estimating Gaseous Pollutants from Bus Emissions: A Hybrid Model Based on GRU and XGBoost." Science of The Total Environment 783: 146870.

Ilbeigi, Marjan, Mohammad Ghomeishi, and Ali Dehghanbanadaki. 2020. "Prediction and Optimization of Energy Consumption in an Office Building Using Artificial Neural Network and a Genetic Algorithm." Sustainable Cities and Society 61: 102325.

Jung, Hyun Chul, Jin Sung Kim, and Hoon Heo. 2015. "Prediction of Building Energy Consumption Using an Improved Real Coded Genetic Algorithm Based Least Squares Support Vector Machine Approach." Energy and Buildings 90 (March): 76–84. https://doi.org/10.1016/j.enbuild.2014.12.029.

Ketkar, Nikhil, and Eder Santana. 2017. Deep Learning with Python. Vol. 1. Springer.

Koengkan, Matheus, José Alberto Fuinhas, and Nuno Silva. 2021. "Exploring the Capacity of Renewable Energy Consumption to Reduce Outdoor Air Pollution Death Rate in Latin America and the Caribbean Region." Environmental Science and Pollution Research 28 (2): 1656–74.

Li, Yan. 2019. "Prediction of Energy Consumption: Variable Regression or Time Series? A Case in China." Energy Science & Engineering 7 (6): 2510–18. https://doi.org/10.1002/ese3.439.

Lü, Xiaoshu, Tao Lu, Charles J Kibert, and Martti Viljanen. 2015. "Modeling and Forecasting Energy Consumption for Heterogeneous Buildings Using a Physical–Statistical Approach." Applied Energy 144 (April): 261–75. https://doi.org/10.1016/j.apenergy.2014.12.019.

Mitchell, Rory, and Eibe Frank. 2017. "Accelerating the XGBoost Algorithm Using GPU Computing." PeerJ Computer Science

3 (July): e127. https://doi.org/10.7717/peerj-cs.127.

Moletsane, Phenyo Phemelo, Tebogo Judith Motlhamme, Reza Malekian, and Dijana Capeska Bogatmoska. 2018. "Linear Regression Analysis of Energy Consumption Data for Smart Homes." In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 0395–99. IEEE. https://doi.org/10.23919/MIPRO.2018.8400075.

Moreno, M, Benito Úbeda, Antonio Skarmeta, and Miguel Zamora. 2014. "How Can We Tackle Energy Efficiency in IoT BasedSmart Buildings?" Sensors 14 (6): 9582–9614. https://doi.org/10.3390/s140609582.

O'Neill, Zheng, and Charles O'Neill. 2016. "Development of a Probabilistic Graphical Model for Predicting Building Energy Performance." Applied Energy 164 (February): 650–58. https://doi.org/10.1016/j.apenergy.2015.12.015.

Özen, Nur Selin, Selin Saraç, and Melik Koyuncu. 2021. "COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları Ile Tahmini: Amerika Birleşik Devletleri Örneği." Avrupa Bilim ve Teknoloji Dergisi, no. 22: 134–39.

Protić, Milan, Fahman Fathurrahman, and Miomir Raos. 2019. "Modelling Energy Consumption of the Republic of Serbia Using Linear Regression and Artificial Neural Network Technique." Tehnicki Vjesnik - Technical Gazette 26 (1): 135–41. https://doi.org/10.17559/TV-20180219142019.

Roldán-Blay, Carlos, Guillermo Escrivá-Escrivá, Carlos Álvarez-Bel, Carlos Roldán-Porta, and Javier Rodríguez-García. 2013. "Upgrade of an Artificial Neural Network Prediction Method for Electrical Consumption Forecasting Using an Hourly Temperature Curve Model." Energy and Buildings 60 (May): 38–46. https://doi.org/10.1016/j.enbuild.2012.12.009.

Şahin, Hüseyin, Tugrul Oktay, and Mehmet Konar. 2020. "Anfis Based Thrust Estimation of a Small Rotary Wing Drone." Avrupa Bilim ve Teknoloji Dergisi, no. 18: 738–42.

Wang, Jian Qi, Yu Du, and Jing Wang. 2020. "LSTM Based Long-Term Energy Consumption Prediction with Periodicity." Energy 197: 117197.

Wei, Yixuan, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, and Xiaoyun Zhao. 2018. "A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption." Renewable and Sustainable Energy Reviews 82: 1027–47.

Zhou, Jian, Enming Li, Mingzheng Wang, Xin Chen, Xiuzhi Shi, and Lishuai Jiang. 2019. "Feasibility of Stochastic Gradient Boosting Approach for Evaluating Seismic Liquefaction Potential Based on SPT and CPT Case Histories." Journal of Performance of Constructed Facilities 33 (3): 4019024.