Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 9, No. 05, May, 2023

Prediction of Energy Consumption of an Administrative Building using Machine Learning and Statistical Methods

Meryem El Alaoui [1*], Laila Ouazzani Chahidi [2, 3], Mohammed Rougui [1], Abdeghafour Lemrani [1], Abdellah Mechaqrane [2]

[1] *LGCE, Civil Engineering and Environment Laboratory, High School of Technology (EST)-Sale, Mohammed V University, PO. Box 227, Rabat Sale, Morocco.*

[2] *SIGER, Intelligent Systems, Georesources and Renewable Energies Laboratory, Faculty of Sciences and Techniques, Sidi Mohamed Ben Abdellah University, PO. Box 2202, Fez, Morocco.*

[3] *LISAC, Computer Science, Signals, Automation and Cognitivism Laboratory, Faculty of Sciences Dhar Mehraz, Sidi Mohamed Ben Abdellah University, PO. Box 1796 Atlas, 30003, Fez, Morocco.*

Received 20 February 2023; Revised 23 April 2023; Accepted 27 April 2023; Published 01 May 2023

Abstract

Energy management is now essential in light of the current energy issues, particularly in the building industry, which accounts for a sizable amount of global energy use. Predicting energy consumption is of great interest in developing an effective energy management strategy. This study aims to prove the outperformance of machine learning models over SARIMA models in predicting heating energy usage in an administrative building in Chefchaouen City, Morocco. It also highlights the effectiveness of SARIMA models in predicting energy with limited data size in the training phase. The prediction is carried out using machine learning (artificial neural networks, bagging trees, boosting trees, and support vector machines) and statistical methods (14 SARIMA models). To build the models, external temperature, internal temperature, solar radiation, and the factor of time are selected as model inputs. Building energy simulation is conducted in the TRNSYS environment to generate a database for the training and validation of the models. The models' performances are compared based on three statistical indicators: normalized root mean square error (nRMSE), mean average error (MAE), and correlation coefficient (R). The results show that all studied models have good accuracy, with a correlation coefficient of $0.90 < R < 0.97$. The artificial neural network outperforms all other models (R=0.97, nRMSE=12.60%, MAE= 0.19 kWh). Although machine learning methods, in general terms, seemingly outperform statistical methods, it is worth noting that SARIMA models reached good prediction accuracy without requiring too much data in the training phase.

Keywords: Energy Management; Tertiary Sector; Energy Prediction; Machine Learning; Statistical Methods.

1. Introduction

The building sector is one of the most significant energy-consuming sectors in Morocco, representing 33% of definite energy utilization and keeping solid development in yearly energy utilization [1]. Integrating intelligent energy management strategies in this sector is among the switches that could assist with meeting the Kingdom's energy challenges and accomplishing its environmental change targets. Energy prediction in buildings is, therefore, essential for intelligent management. Real-time monitoring and decision-making are made possible by approaches based on artificial intelligence (AI), which can be particularly useful for reducing energy consumption in this sector. Substantial research has been carried out on this objective. Moreover, different machine learning methods can be used to predict building energy use.

* Corresponding author: meryem.elalaoui0@gmail.com

http://dx.doi.org/10.28991/CEJ-2023-09-05-01

Artificial neural networks (ANN) are among the most studied methods in this field [2]. A back propagation neural network has been used to predict cooling demands in a commercial building [3]. To identify the model parameters, Yokoyama et al. [3] used the Modal Trimming Method, which, according to them, allows for obtaining appropriate values. In another work, Ekici & Aksoy [4] used a back propagation neural network 3-3-1 to predict the building's heating energy needs using the building transparency ratio, orientation, and insulation thickness as inputs. An average deviation of about 1.48%–5.16% was obtained between calculated and predicted heating energy needs. Yalcintas & Akkurt [5] have developed a three-layer artificial neural network model (7-6-1) for the prediction of total building chiller plant power consumption in a multipurpose high-rise building. The authors stated that the used ANN model is able to successfully predict the outputs (average absolute error of about 10% in the testing phase) and that it could be very helpful for the modeling of heating, ventilation, and air conditioning (HVAC).

In order to enhance the estimation of energy consumption for residential structures during the early design stages, Elbeltagi & Wefki [6] provided a methodology based on an artificial neural network. The dataset used to generate the model is the outcome of simulating several design possibilities with random input variables and calculating the energy consumption. The created ANN model is assessed, validated, and applied to forecast energy usage with a respectable level of accuracy (5.36% and 0.98 for MAPE and R^2, respectively). Moreover, several works were devoted to reviewing and analyzing the studies that exploited the artificial neural network method to predict building energy use [7–9].

Furthermore, other researchers were interested in investigating other machine learning methods. The support vector machine (SVM) method was selected by Dong et al. [10] to predict the monthly energy consumed in four commercial buildings in Singapore. The effect of the model parameters was evaluated, and three weather parameters were considered as inputs: global solar radiation, outdoor temperature, and relative humidity. A coefficient of variance of less than 3% and a percentage error of around 4% were obtained. For the energy management of mixed-use buildings, Culaba et al. [11] used machine learning methods to generate a prediction model for energy usage. To illustrate the created approach, a new fusion of clustering (k-means) and regression (support vector regression) techniques was used to accurately classify and predict the energy use of 30 buildings. The authors stated that the model performs much better than statistical methods used in the literature and within building modeling requirements (with a coefficient of variation of the root mean square error of about 4.10% and a mean bias error of about 0.31%). Using the decision tree method, Yu et al. [12] predicted and classified buildings' energy use intensity levels for Japanese residential buildings. A good agreement was obtained between predicted and actual target values: 93% for the training and 92% for the testing phases. The authors have pointed out the advantages of this method compared to the other most commonly used methods, such as traditional regression and ANN methods. Edwards et al. [13] employed seven machine learning algorithms based on linear regression, support vector machines, and artificial neural networks to predict next-hour residential building consumption. The measurements were collected within 15 minutes from three west Knox County, Tennessee, residential buildings. Their results showed that Least Square Support Vector Machine (LS-SVM) is the best technique for predicting future electrical consumption. In another study, Li et al. [14] evaluated the accuracy of four machine learning models (back propagation neural network (BPNN), radial basis function neural network (RBFNN), general regression neural network (GRNN), and support vector machine (SVM)) in predicting the hourly cooling load of an office building in Guangzhou, China. According to their results, SVM and GRNN outperform the other methods in terms of accuracy and generalization. In order to predict the thermal loads of buildings, Papadopoulos et al. [15] evaluated the performance of random forests (RF), gradient-boosted regression trees (GBRT), and extremely randomized trees (ET). The authors compared their results with other studies' models (ANN, SVM, GP, and RF) and proved that the tree-based ensemble learning models accurately predict building energy loads. In another work, Borowski & Zwolińska [16] discussed cooling energy prediction methods for a hotel building in southern Poland throughout the summer, using neural networks and support vector machines. The input parameters used were meteorological data, time data, and occupancy level. To find the model with the highest accuracy, many configurations were tried using the input and output data that had been gathered. Results revealed that the application of neural networks led to a rise in prediction accuracy. The Weighted Absolute Percentage Error (WAPE) and the Coefficient of Variance (CV) of 19.93% and 27.03%, respectively, defined the best of the presented models.

In another work, Dong et al. [17] proposed an energy consumption prediction technique for buildings based on ensemble learning and classification of energy consumption patterns. For this research, hourly weather information from a meteorological station and energy usage information from a New York City office building were combined. Energy consumption data is first classified into appropriate categories using a decision tree to analyze consumption trends. The ensemble learning approach is then used to create energy consumption forecasting models for each pattern. In the end, the suggested method's prediction accuracy is compared to that of the other three approaches, namely, ensemble learning without energy consumption pattern categorization, SVR, and ANN. By contrasting the accuracy of various systems' predictions under varied training data loads, the robustness of different approaches is also evaluated. Results show that ensemble learning models with energy consumption pattern classification achieved the best results with Coefficient of the Variation of the Root Mean Square Error (CVRMSEs) of 17.7, 16.1, 15.4, 15.8, and 15.6% under the data availability thresholds of 20, 40, 60, 80, and 100%.

In addition to machine learning, there are statistical methods that deal with time series. One popular and widely used statistical method for energy prediction is the autoregressive integrated moving average (ARIMA) model. This method is suitable for analyzing and predicting time series data. It is characterized by its convenience and accurate prediction, simple computational process, and low data input requirement [18]. When the time series present seasonality, the adequate model for energy prediction could be seasonal ARIMA (SARIMA). Numerous studies have investigated ARIMA or SARIMA models for energy prediction. Jeong et al. [19] predicted the annual energy cost budget in South Korean educational establishments using SARIMA and a hybrid model combining SARIMA and ANN. The proposed hybrid model shows better accuracy with lower values of mean average error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) compared to the conventional SARIMA model. Based on historical data from 1973 to 2015, Camara et al. [20] predicted residential energy consumption in the U.S. using the Neural Networks (ANN) and statistical (SARIMA) approaches. Results show that both methods performed well in energy prediction, with slightly better performance using the ANN model. In another study, Tarmanini et al. [21] compared the performance of two forecasting techniques—Auto Regressive Integrated Moving Average (ARIMA) and Artificial Neural Network (ANN) for predicting electricity load. The mean absolute percentage error is used to compare the two forecasting techniques' performance. The research was based on daily real-load electricity data collected over an 18-month period in Ireland for 709 distinct households. The findings show that for non-linear load data, ANN provides superior results over ARIMA.

In order to predict the load demand for the upcoming month across the entire country and in each of Dubai's municipal areas, Sayed et al. [22] compared four predictive models: multiple linear regression (MLR), random forests (RF), artificial neural networks (ANNs), and automatic regression integrated moving average (ARIMA). This analysis made use of electricity consumption data from Dubai. The findings showed that ARIMA had an accuracy of about 93% when dealing with a single district, whereas ANN and RF consistently delivered good accuracy of about 97%.

Predicting building energy needs is of great importance in managing and controlling energy consumption, thus reducing potential energy and financial losses. According to the literature review, some machine learning methods (such as bagging and boosting trees) were not sufficiently exploited. Moreover, few studies were interested in comparing the predictive performances of multiple machine learning and SARIMA models. In this research, four machine learning methods (artificial neural network, bagging trees, boosting trees, and support vector machine) and 14 statistical methods (14 SARIMA models) are used to predict the heating energy consumption of an administrative building in Chefchaouen City, Morocco. Only three meteorological parameters (external temperature, internal temperature, and solar radiation), in addition to the factor of time, are used as model inputs. An evaluation and comparison between artificial intelligence and statistical models are established based on numerous statistical indicators. Emphasis is given to the size of the adequate training data for each approach.

2. Materials and Methods

2.1. Building Description

The building case study, located on the periphery of the city of Chefchaouen, northwest Morocco, is an administrative building constructed in 2006. The weather in this region is mountainous, with a minimum temperature of about 3°C reached in January and a maximum temperature of about 38°C recorded in July. The studied building consists of a single level with nine office rooms and other technical rooms (Figure 1). The building's total area is 468.17 m², with a conditioned area of about 393.6 m² (the set point temperature is 18° C in winter). The building is occupied from 8 a.m. to 4 p.m., five days a week, excluding holidays. Table 1 and Figure 2 summarize the case study building's localization, geometrical, and climatic parameters. The HDD (heating degree day) and CDD (cooling degree day) are calculated based on balance temperatures of 18 °C and 21 °C, respectively [23].

Figure 1. Case study building

Table 1. Building localization, geometrical and climatic parameters

Localization	
Location	Rural municipality of Laghdir, Douar Barhiouen, Chefchaouen City
Latitude	35°43'60"N
Longitude	5°52'60"O
Altitude	34 m
Dimensions	
Area	468.17 m²
Conditioned area	393.6 m²
Eave height	3.5 m
Roof top height	6 m
Climatic parameters	
HDD (balance temperature 18 °C)	896.14 °C
CDD (balance temperature 21 °C)	250.48 °C
Maximum temperature	38 °C
Minimum temperature	3 °C

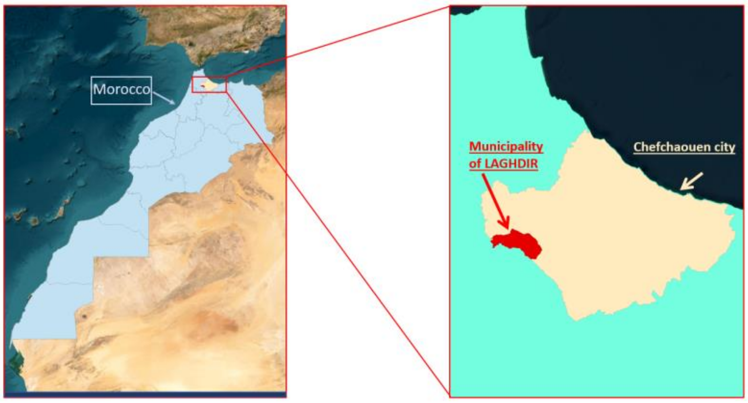

Figure 2. Case study localization

2.2. Methodology

The present study's aim is to compare machine learning (ANN, SVM, BG, and BT) and statistical (SARIMA) methods in predicting the heating energy consumption of the case study building. Due to the lack of measured data, the dataset used for the training and validation of the models is obtained from the energy modeling and simulations of the case study building in the TRNSYS environment, and the weather data of the studied location is imported from Meteonorm software. The TRNSYS (Figure 3) model is used to calculate the heating energy consumption of the building, used as an output in machine learning and statistical models. While external temperature, internal temperature, solar radiation, and the factor of time served as inputs.

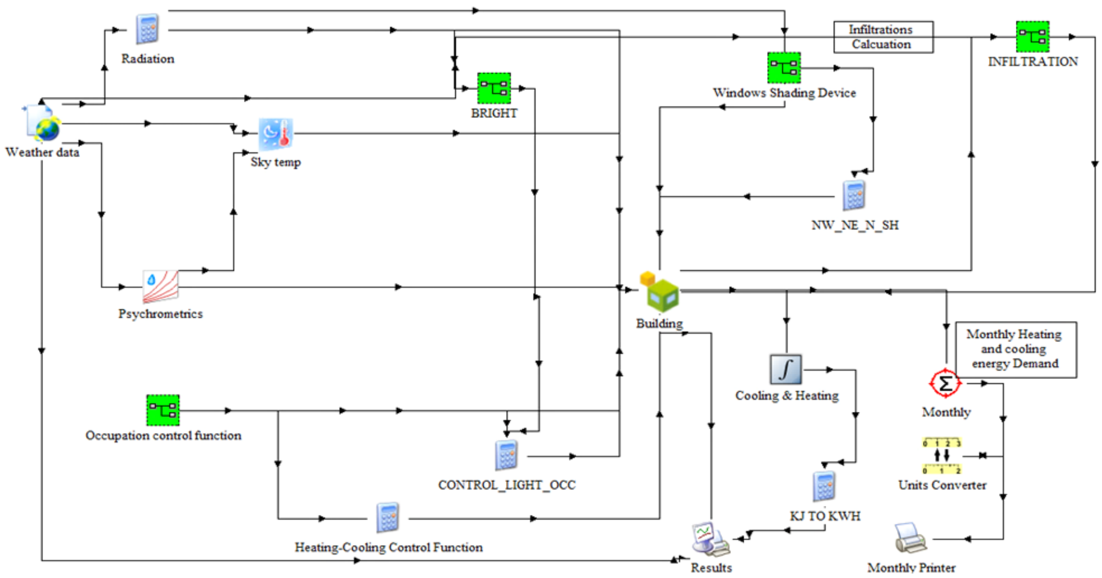

Figure 3. Schematic diagram of TRNSYS building model

The present methodology Flowchart (Figure 4) summarizes how the present study is designed and carried out and how the data is analyzed. First, building architecture and construction data were provided by Laghdir municipality in order to model the building in TRNSYS software. After that, inputs and output database are generated in order to train SARIMA and ML models. The rational selection of a particular set of parameters is based on previous studies (for ML models) and stationarity check and model diagnosis (for SARIMA models). More details related to models' setting are provided in sections below. The built models are utilized to predict building energy use in the ending stage, and performance comparisons based on three performance indicators are made.

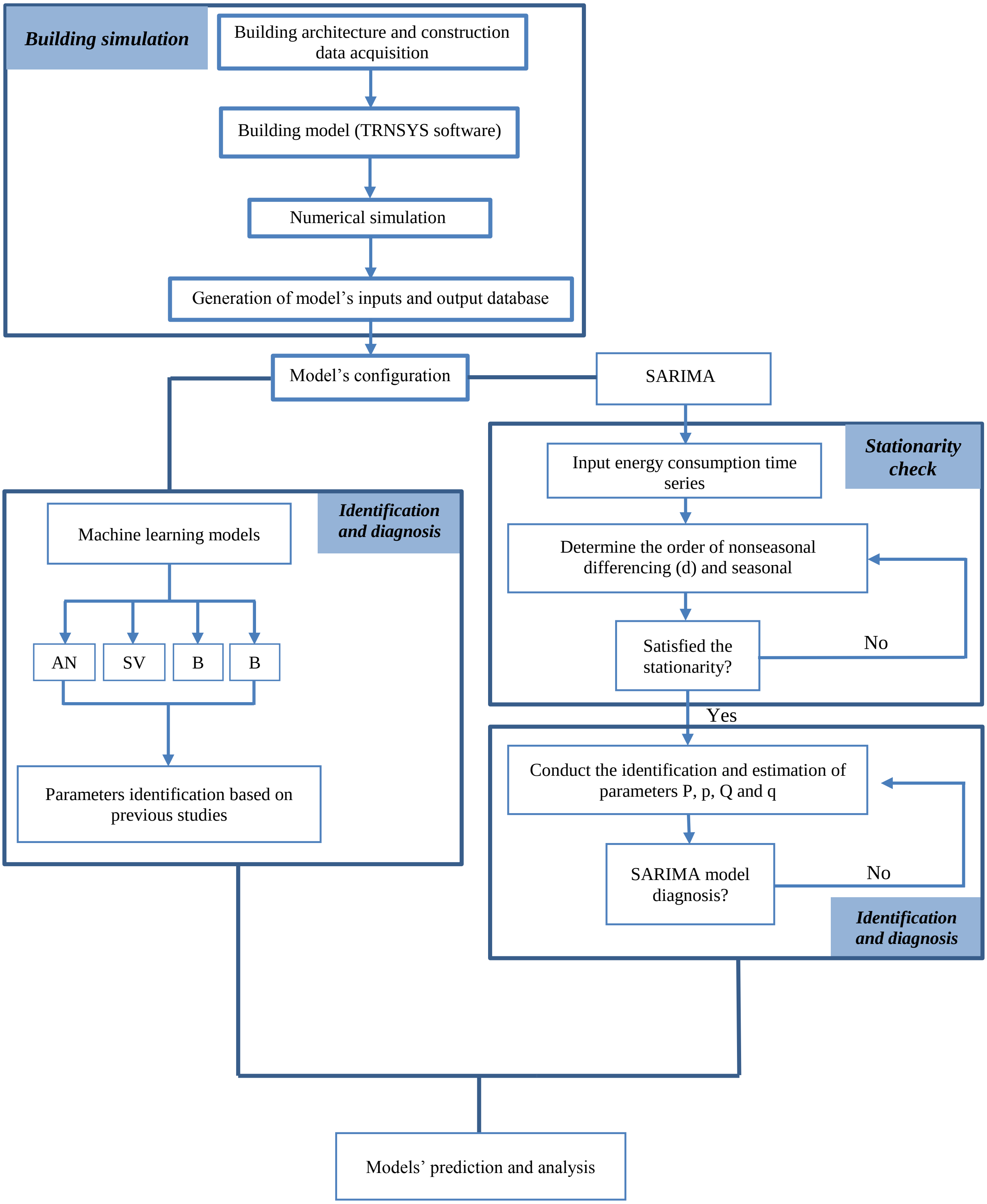

Figure 4. Flowchart of methodology

2.3. Models Presentation

2.3.1. Statistical Models

The statistical method deals with time series. This latter can be defined as a set of data collected sequentially, usually at fixed time intervals. The most popular and widely used statistical method is the ARIMA model, which Box-Jenkins introduced in 1976. The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. It combines the moving average (MA), the AR + MA (ARMA), and the autoregressive (AR) models. ARIMA model is used for non-stationary series, contrary to AR, MA and ARMA, which are used for stationary time series. When time series present seasonality, the appropriate used model is SARIMA.

The non-seasonal ARIMA model is defined as presented in the following Equation 1 [24]:

$$\nabla^d Y_t = \mu + \Phi_1 \nabla^d Y_{t-1} + \Phi_2 \nabla^d Y_{t-2} + \ldots + \Phi_p \nabla^d Y_{t-p} + U_t + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \ldots + \theta_q U_{t-q} \tag{1}$$

where ∇ is the difference operator; θ and Φ are the coefficients of the moving average and the autoregressive component, respectively; U_t and Y_t are the white noise and the actual values at time t, respectively; the numbers p, d, and q stand for the autoregressive term's order, the degree of difference, and the moving average's order (ARIMA (p,d,q)); and μ is the constant term.

ARIMA model can be written in the lag operator form as in the Equation below 2:

$$\Phi(L) \cdot (1-L)^d \cdot Y_t = c + \theta(L) \cdot U_t \tag{2}$$

where L is the backshift operator; θ (L) and $\Phi(L)$ are polynomials of orders q and p, respectively. The The coefficient c is calculated based on the following Equation 3:

$$c = \mu \cdot \left(1 - \Phi_1 - \Phi_2 - \ldots - \Phi_p\right) \tag{3}$$

ARIMA (p, d, q) (P, D, Q) is also used to describe seasonal ARIMA (SARIMA), where P specifies the number of seasonal autoregressive (SAR) terms, D the seasonal differences, and Q the number of seasonal moving average (SMA) terms.

The lag operator polynomial form of the SARIMA model is written as in Equation 4:

$$\Phi_p(L) \cdot \Phi_p(L^k) \cdot (1-L)^d \cdot (1-L^k)^D \cdot Y_t = \theta_q(L) \cdot \Theta_Q.(L^k) \cdot U_t \tag{4}$$

Some steps are followed to determine the best SARIMA model. First, a time plot of the data is created, and the time series' stationarity is tested. If the series are not stationary, data differentiating is required to make them stationary; then, the correlogram function is used to find potential models in the next stage. Finally, the seasonal and non-seasonal components of autoregressive and moving average parameters must be estimated until a suitable model is found [20, 25]. To find the optimum SARIMA model, two criteria must be used: "The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)" (Equations 5 and 6):

$$AIC = log \frac{\sum_{t=k}^{n}(x_t - x)^2}{n} + \frac{n+2k}{n} \tag{5}$$

$$BIC = log \frac{\sum_{t=k}^{n}(x_t - x)^2}{n} + \frac{k\, log(n)}{n} \tag{6}$$

where n denotes the sample size and k denotes the number of estimated parameters.

2.3.2. Artificial Neural Networks

The ANN model is a powerful and flexible machine learning technique inspired by the human brain's system structure. It provides an effective modeling algorithm even with non-linearity between output signals and feature variables [26]. Various neural network models were proposed in the literature, such as feed-forward, Elman, Hopfield, radial basis networks, self-organizing maps, and others [27].

In this paper, the feed-forward neural network is used to predict the heating energy consumption of an office located in Laghdir, near Chefchaouen City, Morocco. The Levenberg-Marquardt (LM) algorithm is selected as an optimization algorithm since it is the most widely used one [28]. Also, it was observed in the literature that the Levenberg-Marquardt algorithm and a feed-forward ANN provided good performance [29]. By training the model, neurons weights are updated, and the relationship between inputs and outputs can be established. The activation function used in this study is the tansig function [30] for hidden and output layers [31]. As shown in figure 5, a feed-forward neural network with one hidden layer composed of ten neurons is adopted.

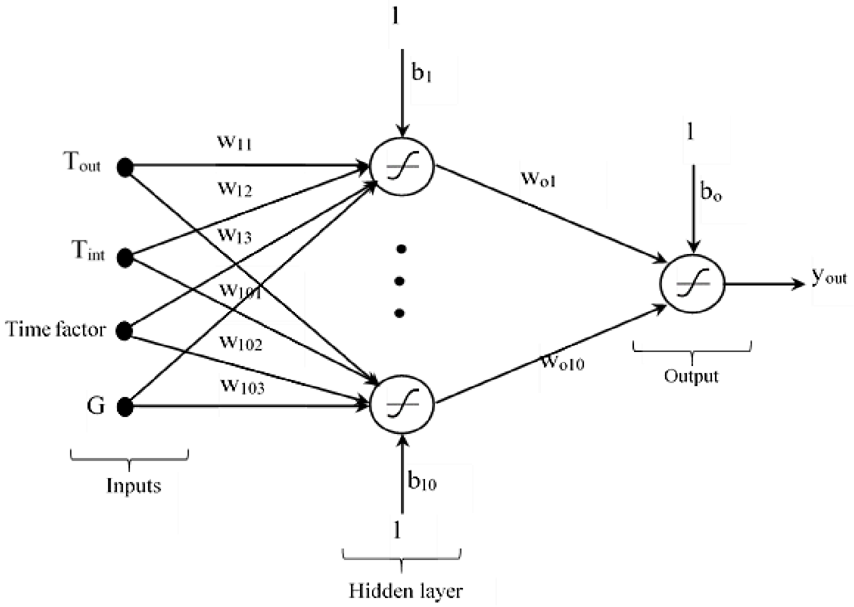

Figure 5. The Architecture of ANN Model

2.3.3. Ensemble Methods

The principle of ensemble methods is based on the combination of several learners in order to generate an optimal predictive model, using the same concept of strength in unity. This technique helps reduce noise, bias, and variance, which are the main causes of error in prediction. In this study, the prediction of heating energy consumption is conducted using Bagging and Boosting-based decision trees. Produced by random sampling with replacement, Bagging and Boosting get N learners from the original set. For Bagging, the models are built independently and trained in parallel. In contrast, for boosting, the new learner is built sequentially, taking into account the results of the previous regressor: The weights of data elements are updated after each training step. The incorrectly predicted cases are given increased weight during the next step [32]. In this study, a minimum leaf size of 8 and 30 learners are chosen for both ensemble methods. For boosting trees, the selected learning rate is equal to 0.1, which is a preferred option for the learning rate [33].

2.3.4. Support Vector Machine

The SVM model is also among the powerful machine learning algorithms used for approaching any multivariate function to any degree of accuracy. This method proved its efficiency in addressing general purpose classification and regression problems [34]. SVM aims to find a hyperplane function f that has at most ε deviation from the actual output vector y (Figure 6). This function is written as presented in Equation 7:

$$f(x) = \langle w, \phi(x) \rangle + b \tag{7}$$

Where w and b are the function parameters.

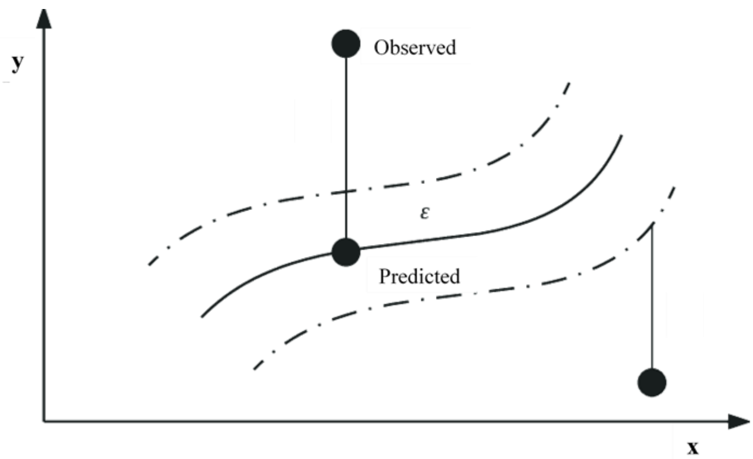

Figure 6. Support vector regression parameters

This function is written in a dot product form to enable the use of the "kernel function" $K(x_i, x_j)$, given the nonlinearity of the problem. After using the Lagrange multiplier, the dual formulation of the optimization problem is as in the Equations 8 and 9 below [35]:

Maximize:

$$\begin{cases} -\frac{1}{2}\sum_{i,k=1}^{1}(\alpha_i - \alpha_i^*)(\alpha_k - \alpha_k^*)\langle\phi(x_i), \phi(x_k)\rangle \\ -\varepsilon\sum_{i=1}^{1}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{1} y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{8}$$

Subject to:

$$\sum_{i=1}^{1} y_i(\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in 0, C \tag{9}$$

Where α_i and α_i^* are Lagrange multipliers.

After replacing the dot product with the kernel function, the new form of the function is presented in the following Equation 10:

$$f(x) = \sum_{j=1}^{1}(\alpha_j - \alpha_j^*) \cdot K(x_j, x) + d \tag{10}$$

The kernel function should be selected depending on the degree of nonlinearity between inputs and outputs. Kernel functions could be Gaussian, linear, polynomial, or other types. In this paper, the Gaussian function is selected [35] (Equation 11):

$$K(x_j, x_i) = exp\left(-\gamma\|x_j - x_i\|^2\right) \tag{11}$$

where γ is the kernel scale, considered in this study 1.7 [36].

2.3.5. Performance Indicators

In order to compare the performance of the previously cited machine learning and statistical methods, three statistical parameters are calculated (Equations 12 to 14): Correlation coefficient R, normalized root mean squared error (nRMSE) and mean absolute error (MAE).

$$R = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \tag{12}$$

$$nRMSE = \frac{1}{\bar{x}}\sqrt{\frac{1}{n}\sum_1^n (x-y)^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_1^n |x-y| \tag{14}$$

where x are the target values; y are the predicted values and n is the data sample size.

The strength of the correlation between the target and predicted values is indicated by the correlation coefficient R. Its value lies in the middle between 0 and 1. When the predicted and target values' curve trajectories are similar, the correlation coefficient R is close to 1, which denotes greater model performance. The mean absolute error (MAE) does not have a limit value because it depends on the calculated parameter. However, a lower value indicates better model performance. The normalized root mean square error (nRMSE) makes it possible to assess how widely apart the predicted values are from the desired ones. According to nRMSE, Table 2 shows various model accuracy levels [37].

Table 2. Model accuracy according to nRMSE

Classes	nRMSE (%)	State of precision
1	nRMSE < 10%	Excellent
2	10% < nRMSE < 20%	Good
3	20% < nRMSE < 30%	Fair
4	nRMSE > 30%	Poor

3. Results and Discussion

The major findings from the case study's prediction of the building's heating energy usage are presented in this section. The performance of each model in predicting the output is assessed for each studied method, namely, machine learning (ANN, SVM, bagging trees, and boosting trees) and statistical (SARIMA) methods.

3.1. SARIMA Models

In this section, a selection of SARIMA models is presented. To build the models, 231 values of the data are used for modeling and parameter estimation, while 91 values are utilized to test and evaluate the models' performance. At first, the stationarity of the time series is checked by applying a unit root test to the data. The results of the augmented Dickey-Fuller test show that the data is stationary, with p-values of 0.0044 smaller than 0.05. The correlograms of residuals are presented in Figures 7 and 8. In both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) graphs, the spikes show a significant seasonal distribution (of about 9). Subsequently, a logarithm transformation is applied to the data in order to stabilize the variance [25].

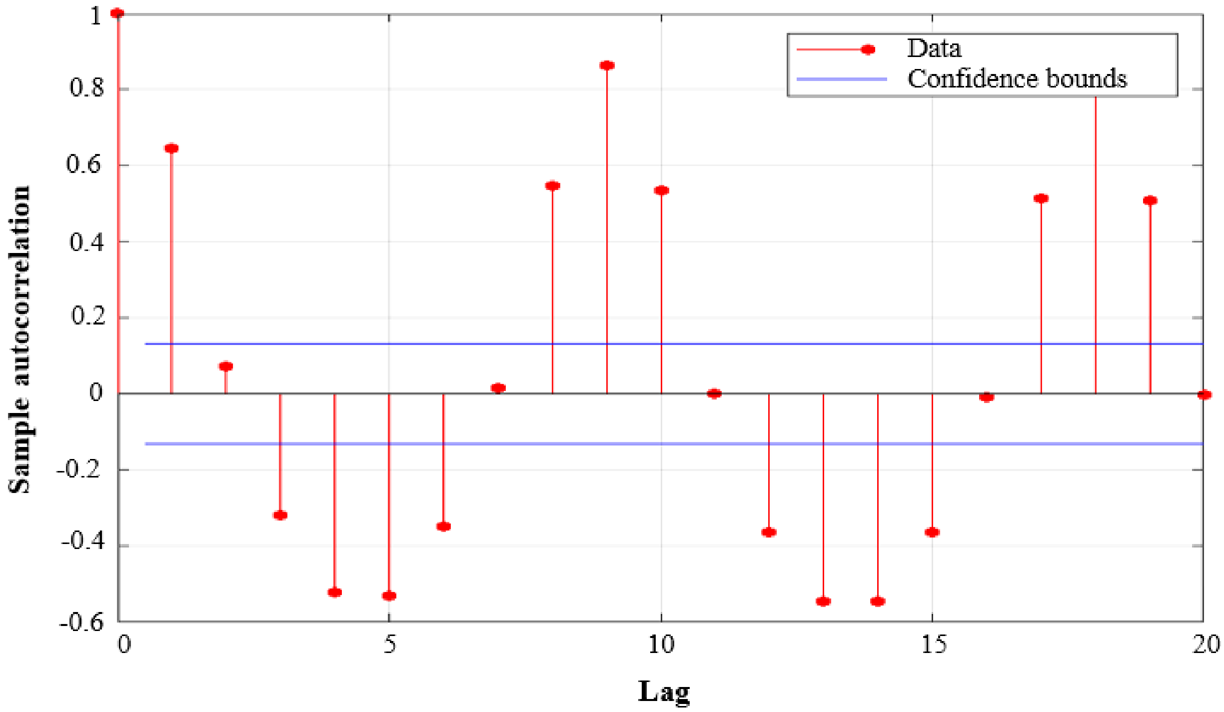

Figure 7. Sample autocorrelation function

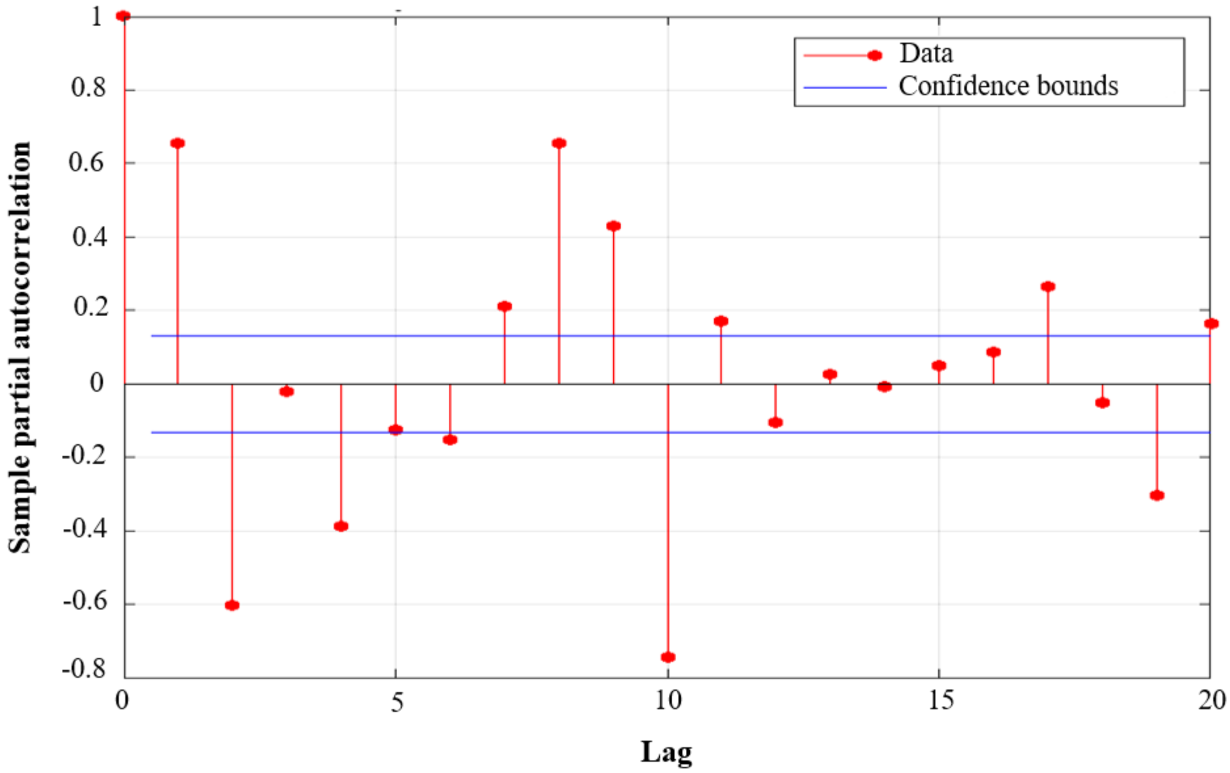

Figure 8. Sample partial autocorrelation function

Several possible SARIMA models of candidates are obtained. After the determination of the parameter values, the Akaike information criteria (AIC) and the Bayesian Information Criteria (BIC) are used to select the best models (Equations 5 and 6). 14 SARIMA models with close and lower AIC and BIC values are chosen as presented in Table 3.

Table 3. AIC and BIC of SARIMA models

SARIMA models	AIC	BIC
SARIMA (1,1,1)*(1,0,1)(9)	1.38	21.98
SARIMA (2,1,1)*(1,0,1)(9)	-1.81	22.19
SARIMA (3,1,1)*(1,0,1)(9)	-3.6	23.78
SARIMA (4,1,1)*(1,0,1)(9)	-4.52	26.25
SARIMA (5,1,1)*(1,0,1)(9)	-4.07	30.08
SARIMA (3,1,2)*(1,0,1)(9)	-5.26	25.5
SARIMA (3,1,3)*(1,0,1)(9)	-4.95	29.29
SARIMA (3,1,4)*(1,0,1)(9)	-6.7	30.9
SARIMA (2,1,3)*(1,0,1)(9)	-6.9	23.9
SARIMA (2,1,2)*(1,0,1)(9)	-1.7	25.72
SARIMA (4,1,3)*(1,0,1)(9)	-7.66	29.9
SARIMA (4,1,2)*(1,0,1)(9)	-8.37	25.8
SARIMA (5,1,3)*(1,0,1)(9)	-8.11	32.88
SARIMA (6,1,3)*(1,0,1)(9)	-6.8	37.5

SARIMA (3,1,1)*(1,0,1)(9) model is chosen randomly in order to examine the ACF and the residual Quantile-Quantile (Q-Q) plot as presented in Figures 9 and 10.

According to the correlogram of residuals (ACF) of the selected model, all the spikes are within the significance limits, and the residuals seem to have white noise. The plot of the residual Q-Q curve shows that the SARIMA model tracks the observed values with good accuracy. After SARIMA models pass the required checks, they are used to predict heating energy consumption. Table 4 presents the models' performance.

Table 4. SARIMA models' predicting performance

SARIMA models	nRMSE (%)	R	MAE (kWh)
SARIMA (1,1,1)*(1,0,1)(9)	20.70	0.90	0.36
SARIMA (2,1,1)*(1,0,1)(9)	21.07	0.90	0.37
SARIMA (3,1,1)*(1,0,1)(9)	21.28	0.90	0.38
SARIMA (4,1,1)*(1,0,1)(9)	21.40	0.90	0.38
SARIMA (5,1,1)*(1,0,1)(9)	21.60	0.90	0.38
SARIMA (3,1,2)*(1,0,1)(9)	21.40	0.91	0.37
SARIMA (3,1,3)*(1,0,1)(9)	22.13	0.92	0.38
SARIMA (3,1,4)*(1,0,1)(9)	21.90	0.91	0.38
SARIMA (2,1,3)*(1,0,1)(9)	22.16	0.91	0.38
SARIMA (2,1,2)*(1,0,1)(9)	21.20	0.91	0.37
SARIMA (4,1,3)*(1,0,1)(9)	22.20	0.91	0.38
SARIMA (4,1,2)*(1,0,1)(9)	21.50	0.91	0.37
SARIMA (5,1,3)*(1,0,1)(9)	22.00	0.91	0.38
SARIMA (6,1,3)*(1,0,1)(9)	21.90	0.91	0.38

According to the results presented in Table 4, the performance indicators show that the 14 SARIMA models have very close accuracy. In fact, the nRMSE, R, and MAE of all SARIMA models range between 20.7% and 22.20%, 0.90 and 0.92, 0.36 kWh, and 0.38 kWh, respectively. Based on the nRMSE, SARIMA (1,1,1)*(1,0,1)(9) slightly outperformed the other models (20.70%).

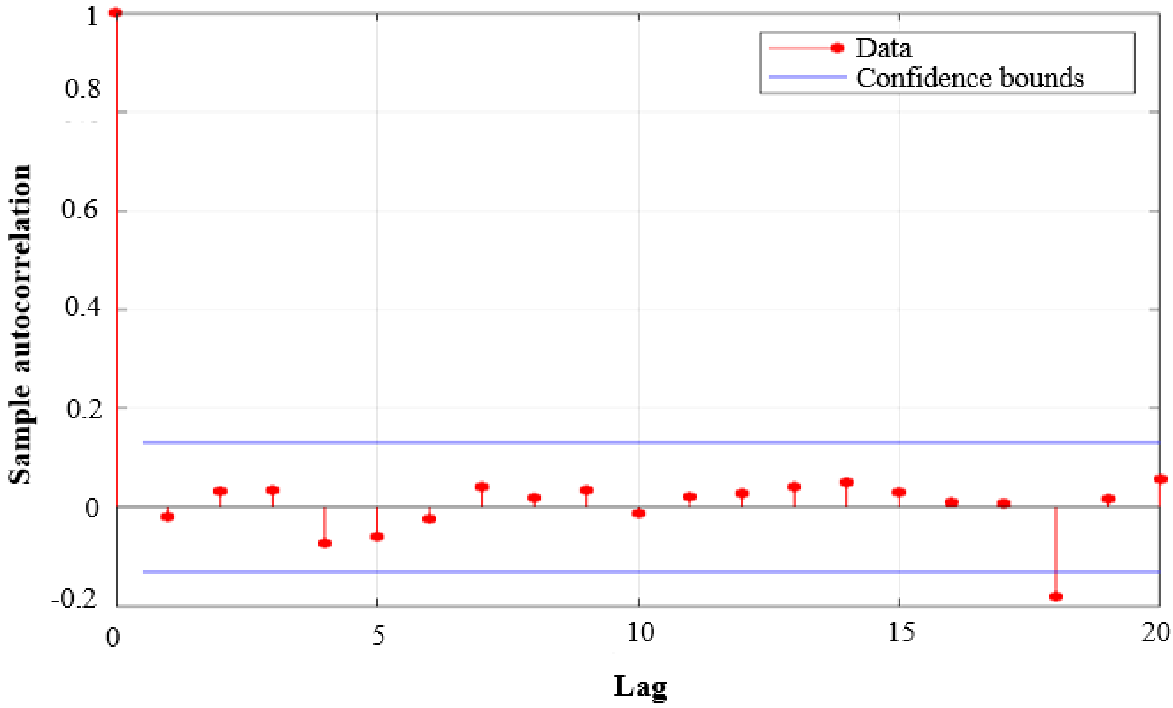

Figure 9. Residual sample autocorrelation function of SARIMA (3,1,1)*(1,0,1)(9) model

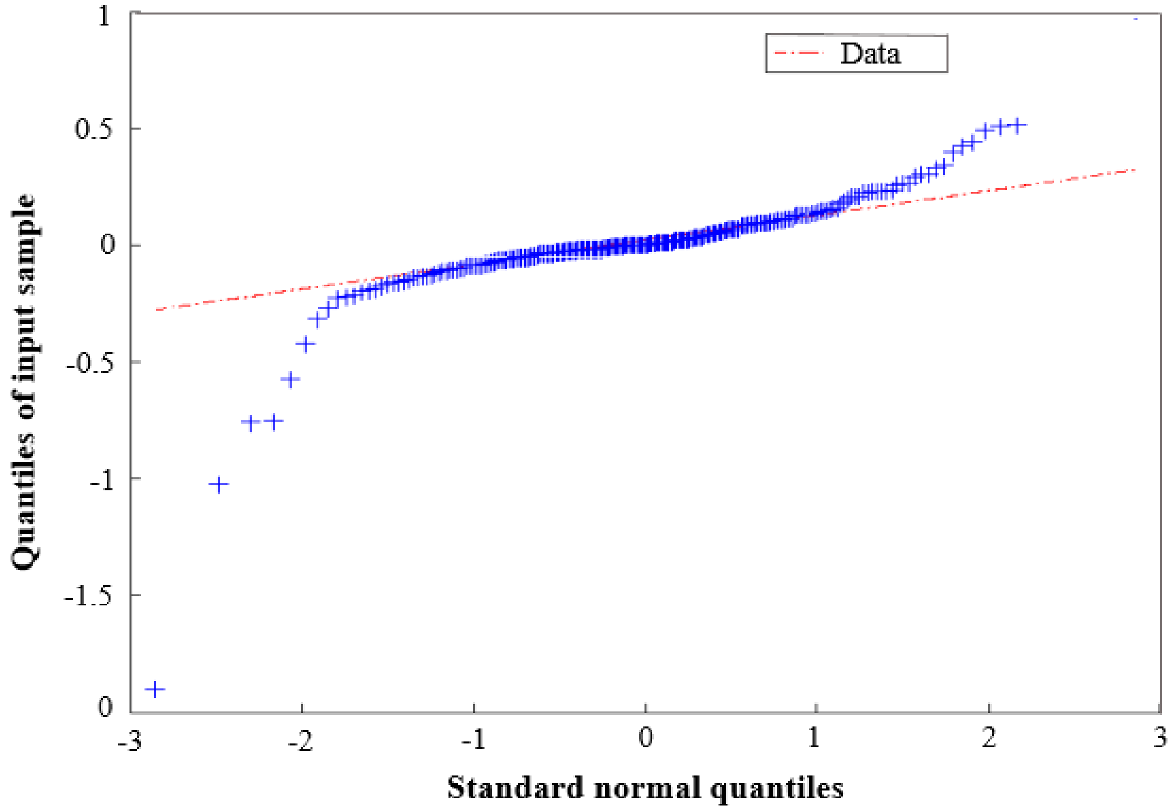

Figure 10. Residual Quantile-Quantile of SARIMA (3,1,1)*(1,0,1)(9) model and target values

3.2. Machine Learning Models

In this section, the performance of the four-studied machine learning algorithms (artificial neural network (ANN), bagging trees (BG), boosting trees (BT), and support vector machine (SVM)) is presented. About 1000 values for each input parameter are used to train the machine learning models. Some 91 data are kept for the testing phase. The models' performance is evaluated and compared based on the three statistical indicators: R, nRMSE, and MAE (section 2.3.5).

The results presented in Table 5 show that the studied machine learning models predict with good performance the heating energy consumption of the studied building [37–41]. The ANN model shows better performance compared to the other machine learning models. It has the lowest nRMSE and MAE (12.6% and 0.19 kWh, respectively) and the highest correlation coefficient (0.97). The ANN model is followed by SVM (nRMSE of about 16.7%, MAE of about 0.25 kWh, and R of about 0.95), then Bagging Trees (nRMSE of about 17.3%, MAE of about 0.26 kWh, and R of about 0.95), and finally, Boosting Trees (nRMSE of about 19.5%, MAE of about 0.32 kWh, and R of about 0.95).

Table 5. Machine learning models' predicting performance

	nRMSE (%)	R	MAE (kWh)
ANN	12.60	0.97	0.19
BG	17.30	0.95	0.26
BT	19.50	0.95	0.32
SVM	16.70	0.95	0.25

3.3. Evaluation and Discussion

Comparing the results presented in Tables 4 and 5, it is observed that the machine learning models outperform the SARIMA models. The Boosting Trees model, the last ranked model in terms of performance, outperformed all the SARIMA models and reached a higher correlation coefficient (0.95 against 0.92 reached by SARIMA models) and lower nRMSE and MAE (19.5% and 0.32 kWh against 20.7% and 0.36 kWh, respectively). These results are similar to the findings of Tarmanini et al. [21] and Sayed et al. [22]. In fact, according to their case studies, machine learning models (ANN, as an example) provide superior results over ARIMA models. For Tarmanini et al. [21], the MSE and MAPE results and the regression factor being closer to one all demonstrate that ANN had a lower error than ARIMA. According to the results, both ANN and ARIMA have the ability to predict consumption; however, ANN is superior at handling non-linear data. Moreover, ANN performs better when it comes to handling multiple tasks at once. For many predictive data mining applications, neural networks are the go-to technology because of their simplicity, effectiveness, and usability. According to Sayed et al. [22], when dealing with a single district, ARIMA had an accuracy of about 93%, whereas ANN and RF consistently provided good accuracy of about 97%.

When comparing the performance of the ANN model with the performance of the selected SARIMA (1,1,1)*(1,0,1)(9) model, it is clear that the prediction of the ANN model is much better. This outperformance is proved by a higher correlation coefficient R (0.97) and lower nRMSE (12.6%) and MAE (0.19 kWh) values. Figure 11 highlights the performance of the two competitive methods: ANN and SARIMA (1,1,1)*(1,0,1)(9). The ANN model approaches with very good accuracy the target values in most 91-testing data. On the contrary, the curve of the SARIMA model shows some deviations from the target values.

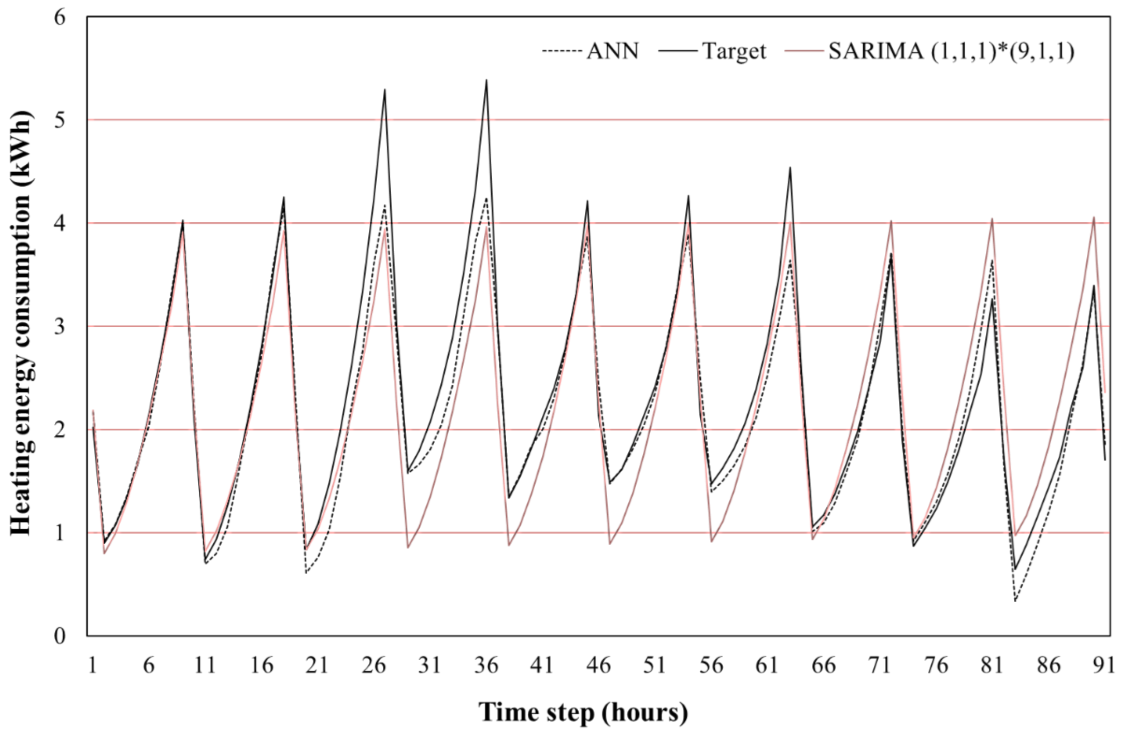

Figure 11. Graph of comparison between target values and heating energy consumption prediction of ANN and SARIMA(1,1,1)*(9,1,1) models

It is worth noting that, although the statistical approach shows less accuracy in energy prediction, SARIMA models have proven the efficiency of modeling the time series of heating energy consumption without requiring too much data in the training phase. In fact, with just 231 data introduced for training the model, the SARIMA model reached an nRMSE of about 20.7%, which can be considered an acceptable performance ($20\% < nRMSE < 30\%$). In addition, it is very close to the 19.5% reached by boosting trees, which required 1000 data points for each input parameter in the training phase. This result is of great interest. In fact, researchers who frequently struggle to obtain historical data may find it very useful to predict the energy consumption of their case studies using just few data in the training phase. However, in the case of having enough historical data, it is better to use machine learning models for energy prediction, so as to reach better precision in terms of predictions and to score better results on indicators of performance. When there is a lack of data, which is the main problem faced during experimental research, statistical models are recommended to be used and give acceptable prediction performances.

4. Conclusions

This study aimed to develop a predictive model for the heating energy consumption of an administrative building on the periphery of the city of Chefchaouen using statistical and machine learning methods.

The main goal of this study is to demonstrate the effectiveness of machine learning models in predicting energy and to emphasize the crucial role that statistical models play in providing precise predictions with a limited database. At the first stage, 14 competitive SARIMA models were selected based on lower values of AIC and BIC in the training phase. The 14 models have comparable results, with a slight outperformance of SARIMA (1,1,1)*(1,0,1)(9) that has a lower nRMSE value (20.7%). After that, the prediction of heating energy consumption was conducted using four machine learning models: ANN, SVM, BG, and BT. Results show that the ANN model outperformed all other models with the lowest nRMSE and MAE and the highest R (12.6%, 0.19 kWh, and 0.97, respectively).

The results of this study demonstrate that, in estimating heating energy usage, machine learning methods perform better than statistical methods. An extremely large data set is required for the training phase of machine learning models in order to execute the prediction step effectively. Even with a minimal amount of data, statistical models can still be used to construct an energy predictive model. This means that statistical models like the ARIMA and SARIMA models are the best option for energy prediction when there is a lack of historical data. The best prediction method is machine learning when there is sufficient data. As a result of this study, it is possible to assert that the issue of an insufficient database, which affects the majority of researchers, might be resolved because statistical models are, according to this research, capable of providing predictions with acceptable accuracy with a limited database.

5. Declarations

5.1. Author Contributions

Conceptualization, M.E. and M.R.; methodology, M.E., M.R., L.O., and A.L.; software, M.E. and A.L.; validation, M.R., L.O., and A.M.; formal analysis, M.E.; investigation, M.E. and M.R.; resources, A.L. and M.E.; data curation, M.E. and L.O.; writing—original draft preparation, M.E., L.O., and M.R.; writing—review and editing, M.E.; visualization, A.M.; supervision, M.R. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available in the article.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Acknowledgements

The authors would like to thank the municipal of Laghdir, Chefchaouen, for providing access to the building architecture and construction data.

5.5. Conflicts of Interest

The authors declare no conflict of interest.

6. Nomenclature

ANN	Artificial Neural Networks;	ACF	Autocorrelation Function;
AI	Artificial Intelligence;	AIC	Akaike Information Criterion;

ARIMA	Auto Regressive Integrated Moving Average;	BG	Bagging Trees;
BIC	Bayesian Information Criterion;	BPNN	Back Propagation Neural Network;
BT	Boosting Trees;	CDD	Cooling Degree Day;
$cov(x, y)$	Covariance;	CV	Coefficient of Variance;
CVRMSE	Coefficient of the Variation of the Root Mean Square Error;	ET	Extremely randomized Trees;
f	SVM hyper-plane function;	GBRT	Gradient Boosted Regression Trees;
GP	Gaussian Process;	GRNN	General Regression Neural Network;
HDD	Heating Degree Day;	HVAC	Heating, Ventilation and Air Conditioning;
K	Kernel function;	k	SARIMA number of estimated parameters;
L	Backshift operator;	LM	Levenberg-Marquardt algorithm;
LS-SVM	Least Square Support Vector Machine;	MAPE	Mean Absolut Error percentage;
MAE	Mean Absolute Error;	ML	Machine Learning;
MLP	Multilayer Perceptron;	n	Number of observations;
nRMSE	Normalized Root Mean Square Error;	PACF	Partial Autocorrelation Function;
R	Correlation coefficient;	RBFNN	Radial Basis Function Neural Network;
RF	Random Forest;	RT	Regression Trees;
SARIMA	Seasonal Auto Regressive Integrated Moving Average;	SAR	Seasonal Autoregressive;
SMA	Seasonal Moving Average;	SVM	Support Vector Machine;
SVR	Support Vector Regression;	U	White Noise;
WAPE	Weighted Absolute Percentage Error;	w, b	SVM hyper-plane function parameters;
x	Input variables;	y	Output variables;
	Greek letters		
α_i, α_i^*,	Lagrange multipliers;	∇	Difference operator;
Φ	Autoregressive component coefficients;	γ	Kernel scale;
σ	Standard deviation;	θ	Moving average coefficients;
μ	Constant term;	ε	Term of deviation;
	Subscripts		
d	Differencing degree;	D	Seasonal difference;
p	Autoregressive term order;	P	Number of seasonal autoregressive;
q	Moving average order;	Q	Number of seasonal moving average;
t	Time;		

7. References

[1] Buildings & Energy Efficiency. (2023). Department of Housing and Urban Policy, Rabat, Morocco. Available online: http://www.mhpv.gov.ma/?page_id=3605 (Accessed on April 2023).

[2] Zhao, H. X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16(6), 3586–3592. doi:10.1016/j.rser.2012.02.049.

[3] Yokoyama, R., Wakui, T., & Satake, R. (2009). Prediction of energy demands using neural network with model identification by global optimization. Energy Conversion and Management, 50(2), 319–327. doi:10.1016/j.enconman.2008.09.017.

[4] Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. Advances in Engineering Software, 40(5), 356–362. doi:10.1016/j.advengsoft.2008.05.003.

[5] Yalcintas, M., & Akkurt, S. (2005). Artificial neural networks applications in building energy predictions and a case study for tropical climates. International Journal of Energy Research, 29(10), 891–901. doi:10.1002/er.1105.

[6] Elbeltagi, E., & Wefki, H. (2021). Predicting energy consumption for residential buildings using ANN through parametric modeling. Energy Reports, 7, 2534–2545. doi:10.1016/j.egyr.2021.04.053.

[7] Mohandes, S. R., Zhang, X., & Mahdiyar, A. (2019). A comprehensive review on the application of artificial neural networks in building energy analysis. Neurocomputing, 340, 55–75. doi:10.1016/j.neucom.2019.02.040.

[8] Runge, J., & Zmeureanu, R. (2019). Forecasting energy use in buildings using artificial neural networks: A review. Energies, 12(17). doi:10.3390/en12173254.

[9] Lu, C., Li, S., & Lu, Z. (2022). Building energy prediction using artificial neural networks: A literature survey. Energy and Buildings, 262, 111718. doi:10.1016/j.enbuild.2021.111718.

[10] Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5), 545–553. doi:10.1016/j.enbuild.2004.09.009.

[11] Culaba, A. B., Del Rosario, A. J. R., Ubando, A. T., & Chang, J. S. (2020). Machine learning-based energy consumption clustering and forecasting for mixed-use buildings. International Journal of Energy Research, 44(12), 9659–9673. doi:10.1002/er.5523.

[12] Yu, Z., Haghighat, F., Fung, B. C. M., & Yoshino, H. (2010). A decision tree method for building energy demand modeling. Energy and Buildings, 42(10), 1637–1646. doi:10.1016/j.enbuild.2010.04.006.

[13] Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings, 49, 591–603. doi:10.1016/j.enbuild.2012.03.010.

[14] Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Conversion and Management, 50(1), 90–96. doi:10.1016/j.enconman.2008.08.033.

[15] Papadopoulos, S., Azar, E., Woon, W. L., & Kontokosta, C. E. (2018). Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. Journal of Building Performance Simulation, 11(3), 322–332. doi:10.1080/19401493.2017.1354919.

[16] Borowski, M., & Zwolińska, K. (2020). Prediction of cooling energy consumption in hotel building using machine learning techniques. Energies, 13(23), 6226. doi:10.3390/en13236226.

[17] Dong, Z., Liu, J., Liu, B., Li, K., & Li, X. (2021). Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. Energy and Buildings, 241, 110929. doi:10.1016/j.enbuild.2021.110929.

[18] Alsharif, M. H., Younes, M. K., & Kim, J. (2019). Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. Symmetry, 11(2). doi:10.3390/sym11020240.

[19] Jeong, K., Koo, C., & Hong, T. (2014). An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network). Energy, 71, 71–79. doi:10.1016/j.energy.2014.04.027.

[20] Camara, A., Feixing, W., & Xiuqin, L. (2016). Energy Consumption Forecasting Using Seasonal ARIMA with Artificial Neural Networks Models. International Journal of Business and Management, 11(5), 231. doi:10.5539/ijbm.v11n5p231.

[21] Tarmanini, C., Sarma, N., Gezegin, C., & Ozgonenel, O. (2023). Short term load forecasting based on ARIMA and ANN approaches. Energy Reports, 9, 550–557. doi:10.1016/j.egyr.2023.01.060.

[22] Sayed, H. A., William, A., & Said, A. M. (2023). Smart Electricity Meter Load Prediction in Dubai Using MLR, ANN, RF, and ARIMA. Electronics (Switzerland), 12(2), 389. doi:10.3390/electronics12020389.

[23] Lamrani, A., & Rougui, M. (2018). Parametric Study of Thermal Renovation of an Administrative Building Envelope in Region of Chefchaouen (Morocco). 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), Rabat, Morocco. doi:10.1109/irsec.2018.8702919.

[24] Kandananond, K. (2011). Forecasting electricity demand in Thailand with an artificial neural network approach. Energies, 4(8), 1246–1257. doi:10.3390/en4081246.

[25] Tatarkanov, A. A., Alexandrov, I. A., Chervjakov, L. M., & Karlova, T. V. (2022). A Fuzzy Approach to the Synthesis of Cognitive Maps for Modeling Decision Making in Complex Systems. Emerging Science Journal, 6(2), 368-381. doi:10.28991/ESJ-2022-06-02-012.

[26] Zhang, Z. (2016). A gentle introduction to artificial neural networks. Annals of Translational Medicine, 4(19), 370. doi:10.21037/atm.2016.06.20.

[27] Krenker, A., Bester, J., & Kos, A. (2011). Introduction to the Artificial Neural Networks. Artificial Neural Networks - Methodological Advances and Biomedical Applications. doi:10.5772/15751.

[28] Ranganathan, A. (2004). The levenberg-marquardt algorithm. Tutorial on LM algorithm, 11(1), 101-110.

[29] Rodrigues, F., Cardeira, C., & Calado, J. M. F. (2014). The daily and hourly energy consumption and load forecasting using artificial neural network method: A case study using a set of 93 households in Portugal. Energy Procedia, 62, 220–229. doi:10.1016/j.egypro.2014.12.383.

[30] Taki, M., Abdanan Mehdizadeh, S., Rohani, A., Rahnama, M., & Rahmati-Joneidabad, M. (2018). Applied machine learning in greenhouse simulation; new application and analysis. Information Processing in Agriculture, 5(2), 253–268. doi:10.1016/j.inpa.2018.01.003.

[31] Fidan, S., Oktay, H., Polat, S., & Ozturk, S. (2019). An Artificial Neural Network Model to Predict the Thermal Properties of Concrete Using Different Neurons and Activation Functions. Advances in Materials Science and Engineering, 2019, 1–13,. doi:10.1155/2019/3831813.

[32] Sutton, C. D. (2005). Classification and Regression Trees, Bagging, and Boosting. Data Mining and Data Visualization, 303–329, Elsevier, Amsterdam, Netherlands. doi:10.1016/s0169-7161(04)24011-1.

[33] MATLAB. (2023). Ensemble Algorithms - MATLAB & Simulink. Available online: https://www.mathworks.com/help/stats/ensemble-algorithms.html (accessed on April 2023).

[34] Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 86(10), 2249–2256. doi:10.1016/j.apenergy.2008.11.035.

[35] Fu, Y., Li, Z., Zhang, H., & Xu, P. (2015). Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices. Procedia Engineering, 121, 1016–1022. doi:10.1016/j.proeng.2015.09.097.

[36] Wang, W., Xu, Z., Lu, W., & Zhang, X. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. Neurocomputing, 55(3–4), 643–663. doi:10.1016/s0925-2312(02)00632-x.

[37] Bounoua, Z., Ouazzani Chahidi, L., & Mechaqrane, A. (2021). Estimation of daily global solar radiation using empirical and machine-learning methods: A case study of five Moroccan locations. Sustainable Materials and Technologies, 28, 261. doi:10.1016/j.susmat.2021.e00261.

[38] Laabid, A., Saad, A., & Mazouz, M. (2022). Integration of Renewable Energies in Mobile Employment Promotion Units for Rural Populations. Civil Engineering Journal, 8(7), 1406-1434. doi:10.28991/CEJ-2022-08-07-07.

[39] Li, M. F., Tang, X. P., Wu, W., & Liu, H. Bin. (2013). General models for estimating daily global solar radiation for different solar radiation zones in mainland China. Energy Conversion and Management, 70, 139–148. doi:10.1016/j.enconman.2013.03.004.

[40] Despotovic, M., Nedic, V., Despotovic, D., & Cvetanovic, S. (2016). Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. Renewable and Sustainable Energy Reviews, 56, 246–260. doi:10.1016/j.rser.2015.11.058.

[41] Chahidi, L. O., Fossa, M., Priarone, A., & Mechaqrane, A. (2021). Evaluation of supervised learning models in predicting greenhouse energy demand and production for intelligent and sustainable operations. Energies, 14(19). doi:10.3390/en14196297.