Fatemi Bushehri, Seyyed Mohammad Mehdi; Dehghan Khavari, Saeed; Mirjalili, Seyed Hossein; Babaei Meybodi, Hamid; Sardari Zarchi, Mohsen

Article — Published Version

Energy Consumption Prediction in Iran: A Hybrid Machine Learning and Genetic Algorithm Method with Sustainable Development Considerations

Environmental Energy and Economic Research

Suggested Citation: Fatemi Bushehri, Seyyed Mohammad Mehdi; Dehghan Khavari, Saeed; Mirjalili, Seyed Hossein; Babaei Meybodi, Hamid; Sardari Zarchi, Mohsen (2022) : Energy Consumption Prediction in Iran: A Hybrid Machine Learning and Genetic Algorithm Method with Sustainable Development Considerations, Environmental Energy and Economic Research, ISSN 2676-4997, Iranian Association for Energy Economics, Tehran, Vol. 6, Iss. 2, https://doi.org/10.22097/EEER.2022.307251.1224

Environmental Energy and Economic Research 2022 6(2): S034
DOI 10.22097/EEER.2022.307251.1224

Research Article

Energy Consumption Prediction in Iran: A Hybrid Machine Learning and Genetic Algorithm Method with Sustainable Development Considerations

Seyyed Mohammad Mehdi Fatemi Bushehri [a], **Saeed Dehghan Khavari** [b,*], **Seyed Hossein Mirjalili** [c], **Hamid Babaei Meybodi** [d], **Mohsen Sardari Zarchi** [e]

[a] ICT Center of Yazd University, Yazd, Iran
[b] Department of Economics, Meybod University, Meybod, Iran
[c] Faculty of Economics, Institute for Humanities and Cultural Studies, Tehran, Iran
[d] Department of Management, Meybod University, Meybod, Iran
[e] Department of Computer Engineering, Meybod University, Meybod, Iran

Received: 15 October 2021 /Accepted: 14 February 2022

Abstract
Ensuring energy security is a major concern of policymakers and economic planners. This objective could be achieved by managing the energy supply and its demand. The latter has received less attention, especially in developing countries. Neglect of energy consumption and its accurate forecasting leads to potential outages and also unsustainable development. Nonlinear methods that are consistent with the nature of energy consumption have led to better results. Therefore, in the present study, both aspects of sustainable development in the determinants of energy demand and the nonlinear hybrid method have been used. We introduced a model based on sustainable development indicators to forecast energy consumption in Iran in which the relevant indicators are specified by the determination phase. To forecast energy consumption, we provided a new standard dataset for energy consumption in Iran (IREC) based on the data extracted from the World Bank and Ministry of Energy dataset in Iran. The highlight of this research is that it provided the most efficient features from the dataset using the genetic algorithm and five forecasting approaches based on machine learning methods. The algorithm was able to select 14 features as the most effective indicators in predicting energy consumption from all the 104 ones in the IREC with 500 repetitions. The empirical results indicated that the model can provide important indicators for energy consumption forecasting. The experiment result of the model using the GA-Based feature selection indicates that the hybrid model has had better results and GA-SVM and GA-MLP have the best result respectively.
Keywords: Energy consumption prediction, Sustainable development, Predictive model, Machine learning, Data mining.

[*] Corresponding author E-mail: saeed.khavari@meybod.ac.ir

Introduction

In developing countries, the energy demand has been incremental in the last two decades. Besides, due to the macroeconomic uncertainty, energy consumption has a chaotic and nonlinear trend (Haouraji et al, 2020). Accordingly, energy configuration and its estimation have been a crucial issue for developing countries such as Iran.

In this regard, energy consumption prediction is needed for energy planning, formulating strategies, and energy policies. But the prediction is challenging for developing countries where data, suitable models, and required institutions are challenging. Predicted energy consumption typically deviated from the real demands due to the limitations in the model and assumptions (Delmastro et al., 2016).

Therefore, careful prediction of energy consumption is required. Policymakers need to predict more precisely how much energy they will need in a given period. On the one hand, the underestimation of energy demand will lead to higher operating costs, which cannot meet the development needs of the local economy. The overestimation of energy demand also results in the waste of energy and expenses (Fatemi Bushehri and Sardari zarchi, 2017).

The same issue was raised on the prediction method. The machine learning method is the one that might be used in predicting energy demand (Pino-Mejias et al, 2017).

In this paper, we apply a hybrid machine learning method. In a diagnostic model, training data play an important role. Moreover, a standard dataset (containing train and test data) is used. Typically, in datasets with a large number of features, some features have a greater impact on the process of diagnosis and classification (Banadkooki et al., 2020). In the machine learning method, there are approaches to selecting effective features. Evolutionary algorithms, such as the Artificial Bee Colony (ABC) (Jamadi et al., 2016), Ant Colony Optimization (ACO)(Tabakhi et al., 2014), Particle swarm optimization (PSO) (Xue et al., 2012), and Genetic Algorithm (GA)(Fazelpour et al., 2016) are well-known methods in feature selection. In these algorithms, feature sets are usually randomly selected from all features and then their effectiveness in predicting results is measured by criteria set assigned by the model designer.

In the next step, new sets are created and re-evaluated using the resulting feature sets and their effectiveness. By repeating these steps, these algorithms try to evolve the selected set of features and achieve features to optimize the accuracy of the predictor. Due to the number of effective features in energy consumption, we utilized the Genetic Algorithmin machine learning to select the best subset of the most effective features. Genetic Algorithm inspired by a method of selecting genes in nature to achieve the desired response (Chambers, 2019 and Gen and Lin, 2007).

Careful selection of factors is one of the most important aspects of forecasting methods, especially in energy consumption, which genetic algorithms can fulfill this issue. The selection of influencing factors by this algorithm can have a positive effect on the accuracy of the results of the next step methods (i.e. machine learning). In this paper, we examined a combination of two methods that can be effective in improving the accuracy of prediction (reducing standard error).

The remainder of the paper is organized as follows: In section 2, literature review and proposed model validation are provided. Section 3 is devoted to model specification. In section 4, the IREC dataset is provided. In Section 5 the prediction results are presented. section 6 is devoted to the interpretation of the method and results. Finally, section 7 concludes the paper by conclusion.

Literature Review and Proposed Model Validation

Evolutionary algorithms such as Artificial Bee Colony, Ant Colony Optimization, Particle swarm optimization, and Genetic Algorithm play an important role in the machine learning-based models to find effective features in predicting or classification. Feature selection is an integral part of prediction evolutionary algorithms. In fact, in cases where many features are involved in a prediction, feature selection helps to determine the most effective features, because it collects information on some features instead of collecting information for all features, to classify or forecast a process (Li, 2017 and Chandrashekar and Sahin, 2014). The following review on machine learning forecast indicates how evolutionary algorithms and feature selection appeared in the literature on energy consumption prediction.

Kazemi and Hosseinzadeh (2020) decomposed changes in energy consumption and highlighted its structural changes in Iran during 2001-2011 using Input-Output Structural Decomposition Analysis. The results show that structural changes in intermediate inputs caused increased energy consumption. The final demand had the main contribution on increase energy consumption. Among the final demand components, an increase in the level of investment and household consumption was the main driver of energy consumption increment.

Somu and Ramamritham(2020) utilized a hybrid model for building energy consumption forecasting using long and short term memory networks. One feature of the model is the use of optimization algorithms. In this model, memory networks and optimization algorithms are used, which is called ISCOA-LSTM model, an optimized algorithm model based on long- and short-term memory. The results show that the prediction was better than other models in terms of prediction criteria such as prediction absolute error. The prediction was made for both short-term and long-term. They conducted experimental tests which indicated that the predictions made for energy demand are reliable.

Daryaei et al. (2019) examined among others the impact of fossil fuel energy consumption on nitrogen dioxide (NO2) emissions in Iran using time series data for the period 1978–2012. They applied the autoregressive distributed lag (ARDL) bounds testing approach. Findings indicate that energy consumption stimulates NO2 emissions in the long run.

Wei et al. (2019) in a review article compared Conventional models with artificial intelligence-based models for energy consumption prediction. They used different criteria to select the best methods. The regression method was used to predict annual energy consumption and national consumption levels. Accordingly, the nonlinear regression method got the lowest value in the MAPE index.

Hu et al. (2020) predicted China's energy consumption using an enhanced bagged echo state network. In this model, differential algorithms were employed which examined the determinants of energy consumption such as GDP, population, and urbanization rate. Also, among 6 different models and compared to other models, the MAPE of this model is 0.215% which shows less error in predicting and selecting the determinants of energy consumption.

Xiao et al. (2018) predicted China's energy consumption over the period 2015-2020 using a hybrid model based on a selective ensemble. The model, which is a combination of four nonlinear models, is predicted with less error than the other four models individually. The results indicated that using more complex nonlinear models can predict with less error.

Zhao and Luo (2018) Forecasted fossil energy consumption structure toward low-carbon economy in China using ARDL method to examine short- and long-term relationships. They also explored the causal relationship between economic growth and fossil fuels consumption in the Chinese economy. The results show a positive and significant relationship between economic

growth and energy consumption in the short and long term. However, this relationship is stronger in the long run than in the short run which shows an increase in energy consumption in the long run. The results of Granger causality also show a two-way relationship between model variables in the long run. The results show a decrease in the share of coal and oil in China's energy consumption and an increase in the share of natural gas.

Deb et al. (2017) examined time series forecasting techniques of a machine learning model to predict energy consumption for the short, medium, and long term. Although each model has its advantage, composite models made from a combination of two or more individual models have stronger predictive power. According to the authors, providing different forecasting methods as well as different hybrid models can be useful for strong and accurate forecasting in different fields. This paves the way for further research.

Daut et al. (2017) predicted electrical energy consumption using conventional and artificial intelligence methods. They assessed techniques for forecasting electricity consumption especially hybrid of two forecasting techniques such as swarm intelligence (SI) technique and Artificial Intelligence. The results show that the hybrid of SVM and SI techniques has indeed offered superior performance for forecasting the consumption of electricity.

Rostami and Kaveh (2021) developed a combined approach of optimization and machine learning to select optimal features for SAR image classification using biogeography-based optimization, artificial bee colony and support vector machine. The paper aims to maximize the accuracy of classification using a minimum number of features. In their proposed method, they used an optimization support vector machine to classify images using the features designated by the ant colony algorithm as effective features and were able to provide an optimal model for this classification.

Niu et al. (2010) Developed a model for predicting power load using a support vector machine and ant colony optimization method to process large amounts of data and remove redundant and ineffective data. They were able to overcome the problem of slow processing of big data for forecasting and to reduce the cost and time of forecasting by choosing effective features in addition to achieving high accuracy.

Son and Kim (2015) predicted one-month residential electricity consumption by employing support vector regression and fuzzy-rough feature selection with particle swarm optimization. Based on the particle swarm optimization algorithm, they identified 21 features out of a total of 39 features as the most effective features in the prediction.

Vieira et al. (2021) Used a genetic algorithm for feature selection in a model to predict flood. As there are many features related to flood prediction, they used a genetic algorithm to speed up the process. They determined the most effective features in predicting river flow, using parameters in the genetic algorithm based on real data. The value of these parameters is as input to the model which is represented by three indices of coefficient of determination, mean square root error, and mean absolute error. The results indicated that the model predicted river flow with 98% accuracy.

In this paper, we employed the genetic algorithm method to select effective features in forecasting energy consumption, due to its advantage in prediction. Using genetic algorithm method can provide better results than other studies to better simulate the complexities of the two stages of determining factors as well as forecasting.

Based on the results of previous research, in this paper, an energy consumption model based on Iran’s energy data is provided using artificial intelligence methods. For the proposed model, we designed a new standard dataset based on Iran’s energy data called IREC. IREC characteristics are explained in the next section. In the proposed model by utilizing evolutionary

algorithms, novel and most effective features in predicting Iran's energy consumption are provided. The proposed model is in such a way that can be used with different predictors, therefore it has no dependence on one type of predictor.

In the next section, this combined method is examined in detail, in which we clarify how the combination of genetic algorithm and machine learning can be effective in selecting effective factors and improving error criteria.

Model specification

We developed a hybrid genetic algorithm model to predict energy consumption based on its determinants. The model consists of two main phases: determination and forecast. In the determination phase, a Genetic Algorithm-based approach is employed to extract a subset of the dataset that are determinants of energy consumption. In the forecast phase, a machine learning method is used to predict energy consumption.

For the elaboration of the two aforementioned phases, let's simulate the chromosome and gene's structure. Inside the nucleus of cells of all living organisms, the DNA molecule is packaged in a complex, coil-like structure called a chromosome. Each chromosome has a compact region called the centromere in its structure that divides the chromosome into two parts or two short and long arms. The centromere position gives a specific shape to each chromosome and can be used to describe the position of genes. Chromosomes are very different in number and shape in different living organisms. Genes on chromosomes determine the different characteristics of organisms.

For its survival, nature tries to select the best set of genes and pass them on to future generations in the form of chromosomes to create future generations. In this way, nature selects a set of the best genes over time and passes it onto the next generations. Accordingly, by selecting the primary genes, evaluating them, and gradually eliminating the weaker genes over time, nature could achieve a stronger and more optimal set of genes. Knowledge of how this smart process works led to the introduction of machine-learning-based models called Genetic Algorithms. Genetic Algorithms are widely used in feature selection. In such cases, since there are a set of features, predictions utilize these features. The Genetic Algorithm selects the best and most effective traits as genes. It then classifies these genes to form a set of chromosomes. In the next step, the effect of the selected features on estimating the response is measured using a criterion called the fit function. In the next step, considering the impact of the selected features in achieving the desired result, the feature is selected again from the features of the previous step, and then a new set is created and re-evaluated. This process continues until the evaluation result is acceptable. Then, at the end of the last set, the selected features are considered the best features.

In the determinant phase, we designed a determiner based on GA by customizing the fitness function. In this model, a predictor regression was employed to evaluate the efficiency of features selection. In this phase, first of all, 100 subsets of the dataset's features are created randomly. In this step, using a random function, numbers between 1 and the total number of features are generated. Then features that have these numbers are selected and random subsets are generated. Then, we evaluated its accuracy in the prediction process using the fitness function. In the model, the fitness function tries to minimize the cost index. The cost index is defined as presented in equation 1.

$$Cost = Variance/Bias^2 \quad (1)$$

A regression predictor model, whether using maximum likelihood, or least squares for prediction, when there are limited data sets can cause over-fitting. Therefore, determining the right size of the biases functions to prevent over-fitting, increases the flexibility of the model against different data (Bishop, 2006). Therefore, we considered the cost function in such a way that to reduce the ratio of variance to the bias. In this step, if the value of the fitness function is not optimal, the genetic algorithm generates new combinations from the previous stage subsets and the new generation is evaluated by fitness function again. This process is repeated and continues until it reaches an optimal point. In the end, the subset that reaches the optimal point is considered as the subset of the selected features.

In the forecast phase, a regression classifier was employed to evaluate selected features and to learn and configure predicting energy consumption. In this paper to figure out the best forecaster for predicting energy consumption, several successful algorithms were evaluated such as MLP, SVM, and Random Forest. The efficiency of the forecasting is measured by 5 parameters: Correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. Figure 1 shows the model structure.

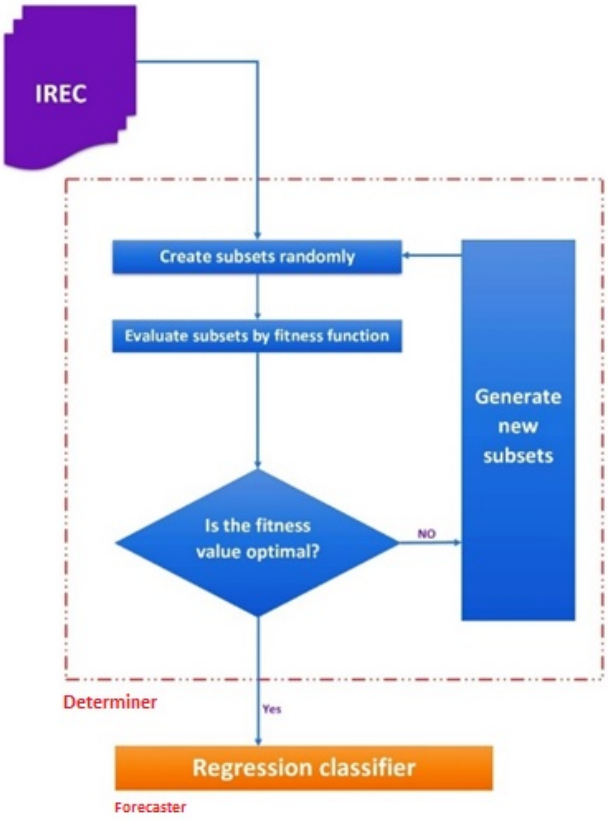

Figure 1. The model structure

IREC dataset

Forecasting models need a standard dataset to train and for configuration. Hence, a standard dataset was developed for Iran's energy consumption based on the Energy Balance Sheet Report that was published annually by the Ministry of Energy and called IREC.

Attention has been paid to sustainable development in the selection of indicators and factors. This goal has been achieved by using the opinions of development economics experts for the initial filtering of factors and out of 521 factors, 104 factors have been selected according to the sustainable development and scoring of economic experts. These 104 factors are then used as input to the algorithm.

In IREC, 104 factors for Iran's energy consumption are considered as features. Also, Iran's energy consumption between 1988 and 2017 has been considered as a target.

Table 1. Features of Iran's energy consumption

Indicator	
Access to clean fuels and technologies for cooking (% of the population)	Imports of goods and services (constant 2010 $US)
Access to electricity, rural (% of rural population)	Individuals using the Internet (% of the population)
Access to electricity, urban (% of urban population)	Industry (including construction), value added (annual % growth)
Adjusted net national income per capita (constant 2010 US$)	Industry (including construction), value added (constant 2010 $US)
Age dependency ratio (% of working-age population)	Inflation, consumer prices (annual %)
Agricultural irrigated land (% of total agricultural land)	International tourism, expenditures (current $US)
Agricultural methane emissions (thousand metric tons of CO2 equivalent)	Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate)
Agricultural nitrous oxide emissions (thousand metric tons of CO2 equivalent)	Life expectancy at birth, total (years)
Agriculture, forestry, and fishing, value added (constant 2010 US$)	Machinery and transport equipment (% of value-added in manufacturing)
Air transport, freight (million ton-km)	Manufacturing, value added (% of GDP)
Air transport, passengers carried	Manufacturing, value added (annual % growth)
Alternative and nuclear energy (% of total energy use)	Manufacturing, value added (constant 2010 $US)
Aquaculture production (metric tons)	Medium and high-tech exports (% of manufactured exports)
Chemicals (% of value-added in manufacturing)	Medium and high-tech Industry (including construction) (% of manufacturing value-added)
CO2 emissions (kg per 2010 $US of GDP)	Merchandise exports (current $US)
CO2 emissions from manufacturing industries and construction (% of total fuel combustion)	Merchandise imports (current $US)
Consumer price index (2010 = 100)	Merchandise trade (% of GDP)
Domestic credit to the private sector by banks (% of GDP)	Mobile cellular subscriptions
Electric power consumption (kWh per capita)	Net investment in nonfinancial assets (% of GDP)
Electric power transmission and distribution losses (% of output)	Nitrous oxide emissions in energy sector (% of total)
Electricity production from coal sources (% of total)	Official exchange rate (LCU per US$, period average)
Electricity production from hydroelectric sources (% of total)	Other manufacturing (% of value-added in manufacturing)
Electricity production from natural gas sources (% of total)	Population ages 65 and above (% of the total population)
Electricity production from nuclear sources (% of total)	Population density (people per sq. km of land area)
Electricity production from oil sources (% of total)	Population growth (annual %)

Electricity production from oil, gas and coal sources (% of total)	Population in the largest city
Employers, total (% of total employment) (modeled ILO estimate)	Population, total
Employment in agriculture (% of total employment) (modeled ILO estimate)	Rail lines (total route-km)
Employment in industry (% of total employment) (modeled ILO estimate)	Renewable electricity output (% of total electricity output)
Employment in services (% of total employment) (modeled ILO estimate)	Renewable energy consumption (% of total final energy consumption)
Energy imports, net (% of energy use)	Self-employed, total (% of total employment) (modeled ILO estimate)
Energy intensity level of primary energy (MJ/$2011 PPP GDP)	Services, value added (% of GDP)
Energy use (kg of oil equivalent) per $1,000 GDP (constant 2011 PPP)	Services, value added (constant 2010 $US)
Exports of goods and services (constant 2010 $US)	Textiles and clothing (% of value-added in manufacturing)
Final consumption expenditure (constant 2010 $US)	Trade (% of GDP)
Financial intermediary services indirectly Measured (FISIM) (constant LCU)	Unemployment, total (% of the total labor force) (modeled ILO estimate)
Fixed telephone subscriptions	Urban population (% of the total population)
Food, beverages and tobacco (% of value-added in manufacturing)	Literacy rate (% of the population)
Forest area (% of land area)	Consumption of Electricity Household (million kWh)
Fossil fuel energy consumption (% of total)	Consumption of Electricity Industrial (million kWh)
GDP (constant 2010 $US)	Consumption of Electricity Public (million kWh)
GDP growth (annual %)	Consumption of Electricity Agricultural (million kWh)
GDP per capita (constant 2010 $US)	Consumption of Electricity Street Lighting (million kWh)
GDP per unit of energy use (constant 2011 PPP $ per kg of oil equivalent)	Number of Establishment Permits Issued for Newly Established Manufacturing (item)
General government final consumption expenditure (constant 2010 $US)	Number of Operation Permits Issued for Newly Established Manufacturing (item)
GNI (constant LCU)	Selected Industrial and Mining Products: Steel (thousand tons)
GNI per capita (constant 2010 US$)	Petrochemical Products (thousand tons)
Government expenditure on education, total (% of GDP)	Cement (thousand tons)
Gross capital formation (constant 2010 $US)	Selected Industrial and Mining Products: Automobile (unit)
Gross domestic savings (% of GDP)	Private Sector Investment in New Buildings in Urban Areas (billion Rials)
Gross fixed capital formation (constant 2010 $US)	Construction Permits Issued by Municipalities in Urban Areas (item)
Households and nonprofit institutions serving households (NPISHs) Final consumption expenditure (constant 2010 $US)	Production Index of Large Manufacturing Establishments

Target (Iran's energy consumption)

Iran's Energy Consumption dataset (IREC) is a standard dataset, which is composed of 30 rows (1 for each year) and 104+1 columns, where the last column is the target and the rest of the columns are the value of the features of Iran's energy consumption. Each row shows features of a year extracted from the World Development Indicators (WDI). Table 1 provides the features of Iran's energy consumption and Figure 2 shows Iran's energy consumption between 1988 to 2017.

Figure 2. Iran's energy consumption between 1988 to 2017

The Prediction Results

For model specification, we employed several experiments. In the first step of the experiments, the regression classifiers are employed without the determiner. These experiments aim to evaluate the performance of classifiers. 10-Fold cross-validation is used to achieve more accurate results. The performance of the model is measured with 5 parameters as follows:

Correlation coefficient: This parameter indicates the amount of coordination predicted value by the correct value. This value range is between -1 and +1, in which -1 indicates the lowest disagreement and +1 indicates the highest agreement (Boughorbel et al, 2017).

Mean absolute error: Measures the comparison of predictions with their final results and calculates errors between a couple of observations that indicate the same phenomenon (Myttenaere et al, 2016).

Root mean squared error: It is the deviation of the prediction errors and measures of how data points are far from the regression line. Root mean squared error is a measure of how these residuals are spread out (McClendon and Meghanathan, 2015).

Relative absolute error: This measures normalized total absolute error. It calculates by mean of the absolute value the actual forecast errors/mean of the absolute values of the naive model's forecast errors (Dietterich and Kong, 1995).

Root relative squared error: This measure is similar to Relative absolute error, except that it calculates its root (Karegowda et al, 2010). Table 2 provides experiments results.

By collecting the data and combining them from 1988 to 2017 and setting the goal, and then using the opinions of development economics experts, finally, we obtained a data set with 104 features and a goal.

In the first phase, the model is estimated using the machine learning method. The experiments without feature selection were conducted on the original data and the 10-Fold Cross-validation method in the training and evaluation phase. The results of the first step are given in table 2. It enables us to compare the results with the results of the second step.

Table 2. The experiments results without the determiner (first phase)

Method	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
M5P Tree	0.9486	84.9159	114.2397	26.20%	31.60%
Decision Table	0.9287	78.682	133.2566	24.28%	36.86%
M5 Rules	0.9733	65.7126	83.7211	20.28%	23.16%
SVM (SMOreg)	0.9953	23.4591	34.4043	7.24%	9.52%
MLP	0.9934	29.7754	40.5066	9.19%	11.20%

In the second phase, estimates and experiments were conducted using the data obtained from feature selection by Wrapper method and genetic evolutionary algorithm on the main data. Then the cost function is implemented using ANN and the cost in this algorithm is calculated based on equation 1. Then, using this algorithm, 14 features listed in table 3 were selected as the most effective features.

Therefore, In the second phase, the model is evaluated again using the proposed determiner. The proposed determiner can extract a subset of features with the highest impact on energy consumption. In this step, the proposed model selects the most efficient features from the dataset using the genetic algorithm, as described in the previous section. This algorithm was able to select 14 features from all the features in the IREC with 500 repetitions. Using these features instead of all features in the forecasting process can reduce the time required to achieve the forecast result. Table 3 shows the features selected by the model.

Table 3. Features selected by the model

NO.	Indicator
1.	Access to electricity, rural (% of rural population)
2.	Adjusted net national income per capita (constant 2010 $US
3.	Employers, total (% of total employment) (modeled ILO estimate)
4.	Energy use (kg of oil equivalent) per $1,000 GDP (constant 2011 PPP)
5.	Financial intermediary services indirectly Measured (FISIM) (constant LCU)
6.	Fixed telephone subscriptions
7.	GDP (constant 2010 $US)
8.	Life expectancy at birth, total (years)
9.	Manufacturing, value added (constant 2010 $US)
10.	Official exchange rate (LCU per $US, period average)
11.	Self-employed, total (% of total employment) (modeled ILO estimate)
12.	Number of Establishment Permits Issued for Newly Established Manufacturing
13.	Selected Industrial and Mining Products: Steel (thousand tons)
14.	Cement (thousand tons)

By determining the subset of features, the previous step experiments are double-checked with this subset. Table 4 shows the second phase experiment result of the model using the GA-Based on feature selection. Figure 3 shows that in all models, the error criterion in the combined method is less than the first step method.

Table 4. The second phase experiments results

Method	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
M5P Tree	0.9829	51.0918	65.319	15.77%	18.07%
Decision Table	0.9829	54.6038	65.002	16.85%	17.98%
M5 Rules	0.9786	54.9752	73.3417	16.96%	20.29%
SVM (SMOreg)	0.9953	23.2703	34.2837	7.18%	9.48%
MLP	0.9956	24.7396	35.1993	7.63%	9.74%

For elaboration of the experiment results, the measurements of the parameters are mapped to a chart for comparison. Figure 4 indicates that in the mean absolute error SVM and MLP have the best result respectively.

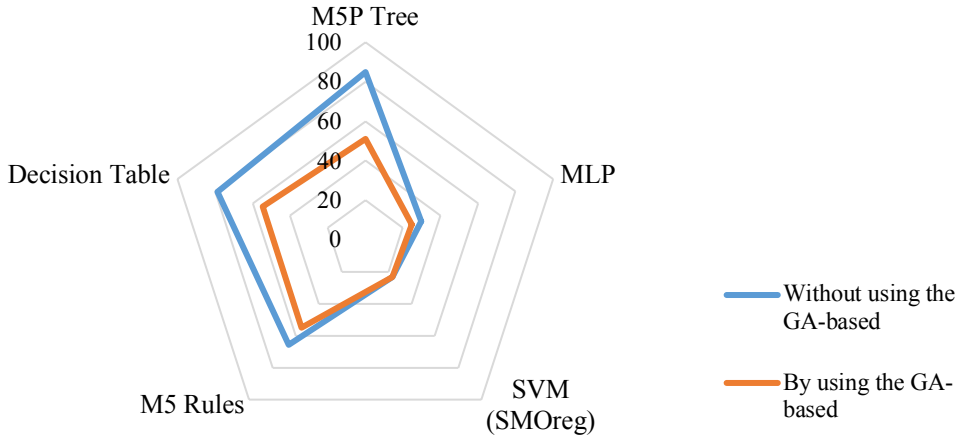

Figure 3. The mean absolute error (Comparison between two phases)

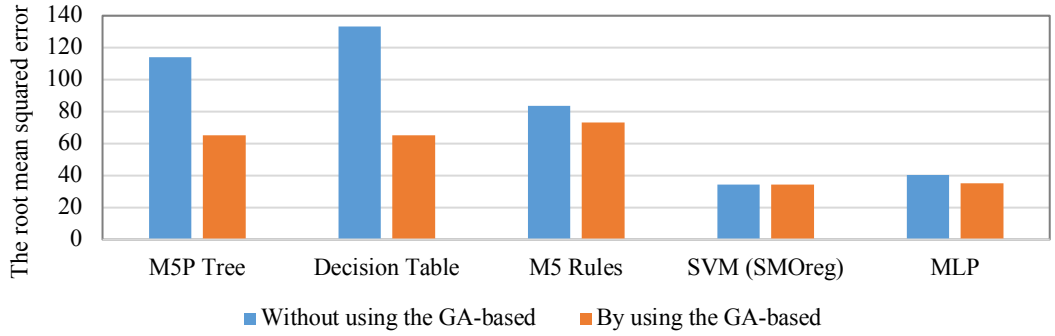

Figure 4. The root mean squared error (Comparison between two phases)

Also, Figure 5 indicates that the SVM mean error and MLP mean error is lower than other methods. Finally, Figure 6 illustrates that SVM and MLP have a lower prediction error than other methods.

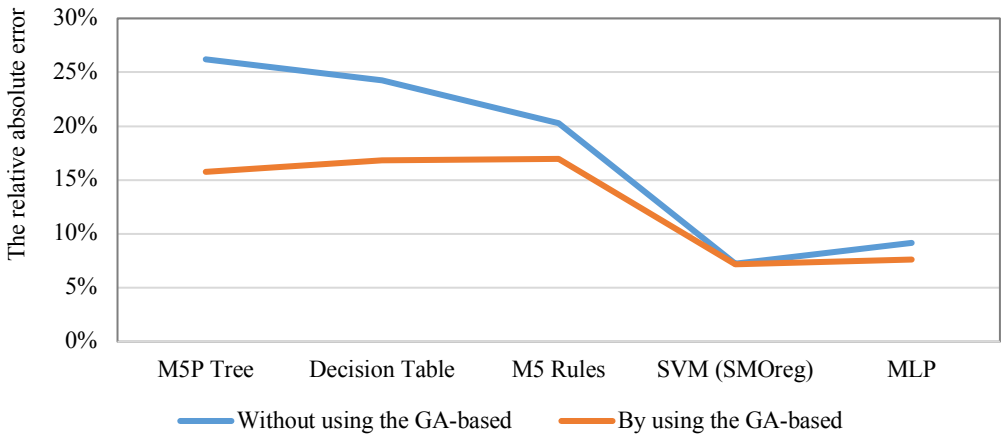

Figure 5. The relative absolute error (Comparison between two phases)

Figure 5 shows the comparison of the relative absolute error criterion between the models of the first and second steps. According to the results, this error criterion has been reduced in all models in the second method, which reduces the error criterion of M5P Tree and Decision Table models.

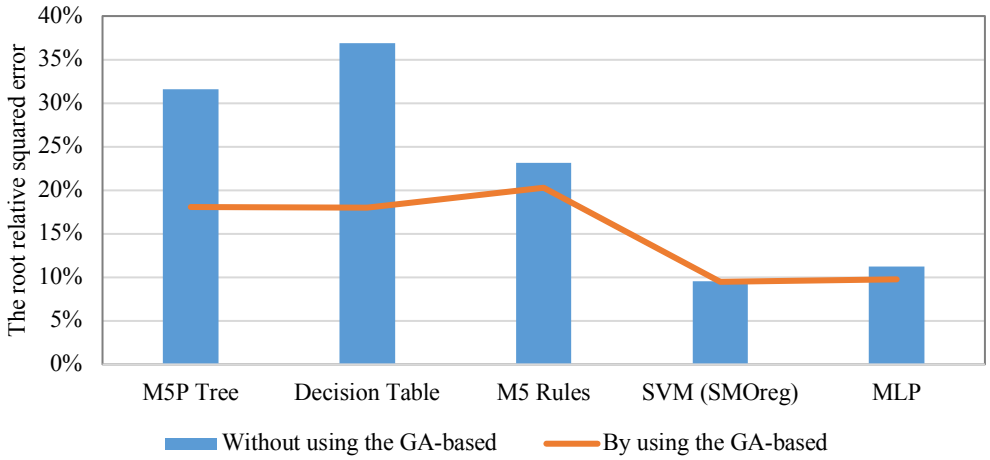

Figure 6. The root relative squared error (Comparison between two phases)

The experiment results indicated that SVM performs better than other methods. Previous research also has shown that this method works well (Yujun et al, 2016). Because this method is based on statistical theory, due to the high complexity of time series prediction problems, they are more reliable than traditional methods. Moreover, SVMs predict fewer errors (Jaramillo et al, 2017). On the other hand, the experimental results show that the model based on evolutionary algorithms can predict energy consumption with less error, using the feature selection method.

Interpretation of the results

The results obtained from the previous section can be examined from different aspects. First, our proposed method has been effective in reducing the error criteria. Estimation of factors affecting

energy consumption was conducted in two steps to compare the error criteria of both steps. As the results show in the second phase, i.e. the use of two-stage and combined model, the error criteria have been reduced, which shows that the combined model performs better than the one-step model.

The second is to use filtering for the determinants of energy consumption. Predicting energy consumption in Iran using the machine learning method resulted in the articulation of the impact of 104 features on energy consumption.

By using the algorithm, the number of factors is reduced to 14, which can facilitate the estimation time of the second step. Analysis of factors will also be easier for policymakers and planners. The impact of features on predicting energy consumption in Iran can be divided into two categories.

First, features that directly affect energy consumption or a criterion for measuring energy consumption in Iran which are: access to electricity in rural areas, adjusted net national income per capita, energy use per $1,000 GDP, GDP itself, value-added of manufacturing, number of establishments permits issued for newly established manufacturing, industrial and mining products of steel and cement.

The results show that the manufacturing industry and the contribution of some industries are higher in energy consumption. Industries like steel are among them. This can be due to the high proportion of these industries. GDP at the highest level and value-added of manufacturing, number of establishments permits issued for newly established manufacturing at the second level, and products of steel and cement at the third level indicate the importance of specific industrial sectors in energy consumption. Access to electricity in rural areas, adjusted net national income per capita and energy use per $1,000 GDP are key features affecting energy consumption.

The second category is the features that indirectly affect energy consumption. There are employers, financial intermediary services, fixed telephone subscriptions, life expectancy at birth, official exchange rate, and self-employed in this category. As an example, employment affects energy consumption due to its role in production. Also, the exchange rate is effective because it affects price indices like the price of energy. The effect of the rate of exchange in different directions affects consumption in general and energy consumption in particular. An overview of the general direction can be investigated that the devaluation of the country's currency (increasing the rate of exchange) will lead to a decrease in the purchasing power of the low-income people and consequently reduce their consumption. It is also low-income people who have a more marginal propensity to consume. Therefore, money depreciation will lead to a reduction in their consumption. But on the other hand, and in the production sector, the rise in the rate of exchange will increase the cost of imports. The exchange rate will also affect the equipping and modernization of industries technologically and reduce energy consumption. But on the other hand, the rise in the exchange rate can lead to a decline in capital goods and thus adversely affect capital formation and production.

Also, human capital features such as education and health conditions are important factors in the country's GDP. Therefore, this factor can directly affect energy consumption. It is also believed that there is a positive relationship between living standards and energy consumption and that living standards are higher, such as life expectancy in countries that have more energy consumption. Financial intermediary services may have different effects on energy consumption. On the one hand, in the national accounts system, financial institutions are formed from two main sectors of banking and insurance. In the Iranian national accounts system, bank operations include the activity of all banks including central banks, commercial banks, and specialized banks. The insurance activity also includes the activity of all insurance companies and agents.

Both sides have a significant share of the services sector in the GDP. On the other hand, e-banking services will lead to easier urban transportation and reduce energy consumption in public transportation and urban traffic.

IT has three effects on energy consumption. It can reorganize production processes to make energy consumption more efficient, and hence, reducing costs (substitution effect). In contrast, the provision of new products and services, and information capital in the technology sector, leads to an additional demand for energy (income effect). Because fixed telephone subscriptions are considered, the first effect will be stronger than the second effect. But overall, IT is an important feature in energy consumption. Also, it can be effective in reducing energy consumption due to increased communication and decreased urban traffic.

Conclusion

The results show improvement in the performance of the hybrid model compared to the model without determiner. The conclusion is that among the methods used in the second phase, MPL and SVM methods have been able to have the lowest error criteria and are more efficient for predicting energy consumption. Also, the results are consistent with the results of Somu and Ramamritham (2020), Hu et al. (2020), Wei et al. (2019), and Xiao et al. (2018). Utilizing the combined method has been able to achieve better results than one-step methods.

Also, the results confirm the conclusion of Zhao and Luo (2018) which indicated a strong relationship between GDP growth and energy consumption, but in addition to this variable, we find factors such as employers, Fixed telephone subscriptions, manufacturing, etc.

Also, compared to Son and Kim (2015), which identified 21 out of 39 factors, using a two-step approach, and recognized fewer factors, our primary factors are more comprehensive. Moreover, considering the opinions of experts in the initial filter of factors in terms of sustainable development is another distinct aspect of the results of our paper relative to Son and Kim (2015).

References

Banadkooki, F.B., Ehteram, M., Ahmed, A.N., Teo, F.Y., Ebrahimi, M., Fai, C. M., Huang, Y. F., and El-shafie, A. (2020). Suspended sediment load prediction using artificial neural network and ant lion optimization algorithm. Environ Sci Pollut Res, 27, 38094–38116.

Bishop, C. M. (2006). Pattern recognition and machine learning: Springer.

Boughorbel, S., Jarray, F. and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, 12(6), e0177678.

Chambers, L. D. (2019). Practical handbook of genetic algorithms: complex coding systems: CRC press.

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

Chen, Y. (2017). A feature-free 30-disease pathological brain detection system by linear regression classifier. CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders), 16(1), 5–10.

Daryaei, A., Bajelan, A., and Khodayeki, M. (2019). The Impact of Stocks Traded-Total Value, Foreign Direct Investment, Number of Students and Fossil Fuel Energy Consumption on NO2 Emissions in Iran. Environmental Energy and Economic Research, 3(4), 335-348.

Daut, M. A. M., Hassan, M., Abdullah,Y. H., Rahman, H. A., Abdullah, M. P., and Hussin, F. (2017). Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. Renewable and Sustainable Energy Reviews, 70, 1108–1118.

Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74, 902–924.

Delmastro, C., Mutani, G., and Schranz, L. (2016). The evaluation of buildings energy consumption and the optimization of district heating networks: a GIS-based model. Int J Energy Environ Eng ,7, 343–351.

Dietterich, T. G. and Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University.

Encyclopædia Britannica, Inc. https://cdn.britannica.com/16/166816-050-C8D9D729.jpg

Fatemi Bushehri, S. M., and Sardari zarchi, M. (2017). A Proposal for a Model for Diagnosis and Classification of Exceptional Children with learning Disabilities by Using Intelligent Expert Systems. Middle Eastern Journal of Disability Studies,.7, 19.

Fazelpour, F., Tarashkar, N., and Rosen, M. A. (2016). Short-term wind speed forecasting using artificial neural networks for Tehran, Iran. International Journal of Energy and Environmental Engineering, 7(4), 377–390.

Gen, M., and Lin, L. (2007). Genetic algorithms. Wiley Encyclopedia of Computer Science and Engineering, 1–15.

Haouraji, C., Mounir, B., Mounir, I., and Farchi, A. (2020). A correlative approach, combining energy consumption, urbanization, and GDP, for modeling and forecasting Morocco's residential energy consumption. International journal of energy and environmental engineering, 11(1), 163-176.

Hu, H. Wang, L., Peng, L., and Zeng, Y. (2020). Effective energy consumption forecasting using enhanced bagged echo state network. Energy, 193, 116778.

Jamadi, M., Merrikh-Bayat, F. and Bigdeli, M. (2016). Very accurate parameter estimation of single-and double-diode solar cell models using a modified artificial bee colony algorithm. International Journal of Energy and Environmental Engineering, 7(1), 13–25.

Jaramillo, J., Velasquez, J. D., and Franco, C. J. (2017). Research in financial time series forecasting with SVM: Contributions from literature. IEEE Latin America Transactions, 15(1), 145–153.

Karegowda, A. G., Manjunath, A. S., and Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation-based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271–277.

Kazemi, H., Hosseinzadeh, R. (2020). Decomposition analysis of Changes in Energy Consumption in Iran: Structural Decomposition Analysis. Environmental Energy and Economic Research, 4(3), 231-239.

Li, J. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 1–45.

McClendon, L., and Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ), 2(1), 1–12.

Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean absolute percentage error for regression models. Neurocomputing, 192, 38–48.

Niu, D., Wang,Y., and Wu, D. D. (2010). Power load forecasting using support vector machine and ant colony optimization. Expert systems with Applications, 37(3), 2531–2539.

Pino-Mejias, R., Pérez-Fargallo, A., Rubio-Bellido, C., and Pulido-Arcas, J. A. (2017). Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO2 emissions. Energy, 118, 24–36.

Rostami, O., and Kaveh, M. (2021). Optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC) and support vector machine (SVM): a combined approach of optimization and machine learning. Computational Geosciences, 25(3), 911–930.

Somu, N., MR, G. R., and Ramamritham, K. (2020). A hybrid model for building energy consumption forecasting using long short term memory networks. Applied Energy, 261, 114131.

Son, H., and Kim, C. (2015). Forecasting short-term electricity demand in residential sector based on support vector regression and fuzzy-rough feature selection with particle swarm optimization. Procedia Engineering, 118, 1162–1168.

Tabakhi, S., Moradi, P., and Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 32, 112–123.

Wei, N., Li, C. Peng, X., Zeng, F., and Lu, X. (2019). Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. Journal of Petroleum Science and Engineering, 181, 106187.

Vieira, A. C. (2021). Improving flood forecasting through feature selection by a genetic algorithm-experiments based on real data from an Amazon rainforest river. Earth Science Informatics, 14(1), 37–50.

Xiao, J., Li, Y., Xie, L., Liu, D., and Huang, J. (2018). A hybrid model based on selective ensemble for energy consumption forecasting in China. Energy, 159, 534–546.

Xue, B., Zhang, M., and Browne, W. N. (2012). Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE transactions on cybernetics, 43(6), 1656–1671.

Yujun, Y., Yimei, Y., and Jianping, L. (2016). Research on financial time series forecasting based on SVM. in 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 346–349.

Zhao, X., and Luo, D. (2018). Forecasting fossil energy consumption structure toward low-carbon and sustainable economy in China: Evidence and policy responses. Energy strategy reviews, 22, 303–312.