Hindawi
Mobile Information Systems
Volume 2021, Article ID 5729630, 7 pages
https://doi.org/10.1155/2021/5729630

Research Article

Using Hybrid Machine Learning Methods to Predict and Improve the Energy Consumption Efficiency in Oil and Gas Fields

Jun Li,[1] **Yidong Guo**,[1] **Xiangyang Zhang**,[2] **and Zhanbao Fu**[2]

[1]*Northwest Branch of Research Institute of Petroleum Exploration and Development of CNPC, Beijing 100083, China*
[2]*Northwest Branch of Research Institute of Petroleum Exploration and Development of CNPC, Lanzhou 730020, China*

Correspondence should be addressed to Jun Li; lijun_xb@petrochina.com.cn

Received 22 July 2021; Accepted 30 August 2021; Published 14 December 2021

Academic Editor: Sang-Bing Tsai

Oil and gas will remain essential to global economic development and prosperity for decades to come, and the oil and gas industry is an energy-intensive industry. Thus, enhancing energy efficiency for producing oil and gas in oil and gas companies is an important issue. The intelligent energy consumption prediction method with the ability to analyze energy consumption patterns and to identify targets for energy saving proved itself as an effective approach for energy efficiency in many industrial domains. Moreover, prediction of energy consumption enables managers to scientifically plan out the energy usage of energy production and to shift energy usage to off-peak periods. However, it still remains a challenging issue to some degree with the unpredictability and uncertainty caused by various energy consumption behaviors, and this phenomenon is becoming more obvious in the oil and gas company. To this end, in our work, we primarily discussed the forecasting of the energy consumption in the oil and gas company. Firstly, four different forecasting models, support vector machine, linear regression, extreme learning machine, and artificial neural network, were trained on the training dataset and then evaluated by the test dataset. Secondly, in order to enhance the energy consumption prediction accuracy, the combinations of all these four models were examined with the RMSE value by taking the average of two models' outputs. The outcomes show that these four different models are able to predict energy consumption with good accuracy, but the hybrid model—artificial neural network and extreme learning machine—would present higher accuracy. In addition, the hybrid model is installed in the energy management system of the oil and gas industry to manage oil field energy consumption and improve the efficiency.

1. Introduction

Oil and gas will remain essential to global economic development and prosperity for decades to come, and in the 2018 Energy Outlook from the British Petroleum Company, it is said that the absolute consumption of oil and gas would have a steady growth trend in 2040 [1]. Moreover, global concerns about climate change are leading to a focus on the amount of energy it takes to produce these hydrocarbon-based fuels and the advent of more unconventional sources and methods that continue to further increase the energy intensity of production. In the face of these challenges, the industry recognizes that energy efficiency and conservation can make a major contribution to both environment protection and energy supply.

Meanwhile, in 2016, the 13th Five-Year Plan for National Economic and Social Development of the People's Republic of China was released. This statement also specifically presented concrete indicators for energy consumption, including electricity sector, renewable energy, hydro, wind, solar, and biomass energy. For instance, it is said that power consumption is expected to be 6800–7200 TWh at an average annual growth rate of 3.6–4.8% [2]. Thus, energy companies must pay more efforts to the reasonable investments of energy over energy production and to cut the energy cost.

Commonly, the most potential for energy conservation lies with end users, and the energy efficiency is a challenging problem for oil and gas companies, which can contribute by implementing improvements in their operations, planning, and investments.

Nowadays, with the increasing momentum of big data, the huge advancement in advanced metering infrastructure, and the rise of the Internet of Things (IoT), large amounts of energy consumption and production data are collected and stored, and many researchers used machine learning techniques based on statistical theory to explore these data and analyzed the energy consumption behavior [3–5]. As a result, the intelligent energy prediction method has proved itself as an efficient approach for enhancing energy efficiency in some industrial domains: authors in [6] introduced the forecasting tool with neural networks and the regression model to predict the day-ahead power output for small-scale solar photovoltaic electricity generators and achieved high forecast accuracy; Prema and Rao proposed time series models for one day-ahead solar power prediction and presented the peak performance with 9.28% error [7]; some authors also considered the ensemble method for short-term probabilistic solar power prediction and the fuzzy method for global solar radiation prediction [8, 9]. Concerning the prediction of building heating energy consumption, various artificial neural networks are explored as well: Guo et al. used the machine learning-based model to predict the energy demand for the indoor heating system [10]; on the contrary, it is worth noting that some researchers used the regression model straightforwardly to calculate the energy consumption, and the widely used regression methods that achieved promising performance include weighted support vector regression and multilinear regression [11–13]; and these methods would ensure more stable energy supply and optimize the management of energy demand for energy enterprises.

In short, the main topic issued in this paper is "how to accurately predict the total energy consumption for one oil and gas company to produce a certain amount of crude oil and gas and help the company improve the energy consumption efficiency?" To deal with this question, this paper discussed four different forecasting models, support vector machine, linear regression, extreme learning machine, and artificial neural network, which were able to accurately forecast the final energy consumption and to improve the energy consumption efficiency. For the sake of higher performance, the hybrid models were discussed as well by taking the average of two models' outputs.

The main innovations of this paper include the following:

(1) The forecasting of the energy consumption in the oil and gas company at a company level, instead of the oil operation level, was presented and proved to be helpful for the company managers and policy makers to make informed decisions concerning the energy usage

(2) Four different prediction methods and 6 hybrid models were discussed and analyzed for the company energy consumption prediction

(3) Four kinds of the performance evaluation indexes were discussed to evaluate the prediction models, and a smaller value of the model's index denotes the better performance of the model

(4) The outcome of this paper can be used in other oil and gas companies to predict the energy consumption

The remaining part of this paper is organized as follows: the commonly used or traditional methods to improve the energy efficiency in the domains of oil and gas from the literature are discussed in Section 2. Section 3 summarizes the proposed model. Section 4 elaborates the methodology of the aforementioned models. The conclusion and the discussion are shown at the end.

2. Traditional Energy Efficiency Improvement Method

In the oil and gas domain, a large variety of opportunities do exist to cut total energy consumption of production as producing the same or more amount of crude oil and natural gas, and the improvement of the efficiency can be affected by the mechanical, chemical, and other physical parameters. The most common techniques for energy efficiency improvement of production include changing efficient production equipment and improving the production process. For example, in the petroleum refinery, the studies from several companies have revealed that there is a strong potential to improve the energy efficiency, and the major areas include utilities, fired heaters, process optimization, heat exchangers, motor and motor applications, and other areas [14].

The following cases from the literature studies are very representation of the widely used energy efficiency improvement methods: in [15], Ping et al. improved the thermal efficiency of the vacuum furnace and achieved the requirements of energy saving via new technologies with the furnace flue gas waste heat recovery technological applications. In [16], to reduce the energy consumption of the oil transportation line, genetic algorithm was used to predict and optimize the oil pipeline. In [17], the hybrid modeling method of comprehensive energy consumption for the oil and gas production process, as is shown in Figure 1, was investigated; they specifically discussed the whole oil and gas production process, including the mechanical extraction system and gathering and transferring process, analyzed every important factor of the process chain, and finally established the LS-SVM model to predict the comprehensive energy consumption in some oil recovery operation areas.

In addition to the optimization of operating conditions and the technological changes in equipment, changing the management of energy usage for production is also one of the most successful ways to cut energy cost of production, and it can be implemented by building an organization-wide energy management program [14], and many oil and gas companies already have designed and built their own energy management system with the strong commitment to boot energy efficiency.

Nowadays, the smart fields also present novel opportunities for the energy efficiency in the oil and gas fields by

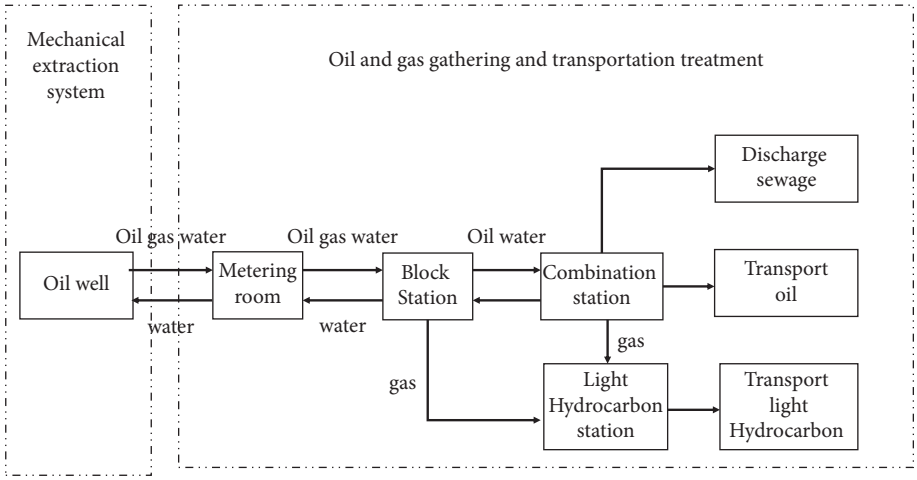

Figure 1: The production process chain of oil and gas.

means of saving operation cost. For instance, in the smart oil field, both workers and engineers have access to the data and the documents related to their operations. When complex problems occurred on-site and were outside the scope of worker's knowledge, they were able to easily cooperate with the experts to solve the problems. Operation costs would dramatically reduce via the quick and improved decision-making. Meanwhile, the smart fields would also be able to offer significant opportunities to improve the existing condition of oil field assets in the context of reservoir engineering and production performance and would finally reduce the energy consumption of oil exploration [18].

In this paper, we investigated the energy efficiency improvement merely via the data-driven energy consumption prediction method with the ensemble machine learning model at a company scale instead of the oil operation station, and the highest-performance model would be used in the practical production environment.

3. Proposed Model

The proposed model consists of three stages, and they are shown in Figure 2. The first stage is designed to divide the whole dataset into several different subsets or clusters, and a clustering method, namely, KNN, is considered. Meanwhile, this clustering method is also applied to remove the nonstationarity or erroneous samples or observations from the whole dataset.

The second stage is constructed to forecast the energy consumption with the two-dimensional input of crude oil and nature gas production over each subdataset, and in this stage, four different forecasting machine learning methods including support vector machine, linear regression, extreme learning machine, and artificial neural network are investigated, respectively.

The last stage is designed for enforcing the final prediction results via the ensemble model. In [19], it was concluded that the generalization ability and accuracy of a machine learning model may be enhanced by an ensemble model with respect to the single model. Meanwhile, many researchers already proved that the simple combination of many models' outputs was beneficial for better performance [20, 21]. In general, the ensemble model performs better in the case where the individual models present good performance over the dataset and their errors distribute on different spaces. In our paper, we tried 6 combinations by taking the average of the two models' outputs from the second stage and then evaluated their performance.

4. Methodology

4.1. Dataset. In this paper, we collect the energy consumption and production data over two decades in four oil and gas companies. Usually, in the oil and gas company, these data would be reported by the energy management system, so the collection frequency is once per month, and then the company manager would adjust the energy usage and production plan over the next month. Thus, every year, we collect 12 observations, and in total, the dataset consists of 960 data samples. We divided these data samples into the training dataset and test dataset with the rate 6 : 4. Concerning the energy consumption in oil and gas industries, it usually includes two parts: the industrial part and the nonindustrial one. The industrial part is the direct energy consumption to produce the oil and gas, and the main contribution of the nonindustrial part is from the oil worker's daily life, such as heating, cooking, and transportation. In our work, due to the complexity and the diversity of the nonindustrial part, we just consider the industrial part, and these datasets are represented in the triple form (x, y, z) where x denotes the crude oil production, y the natural gas production, and z the industrial energy consumption. However, as is shown in Figure 3, the energy consumption fluctuation through the entire observed data shows a weak relation with the time. Thus, just the oil production and gas production are designed as the input for the prediction of energy consumption.

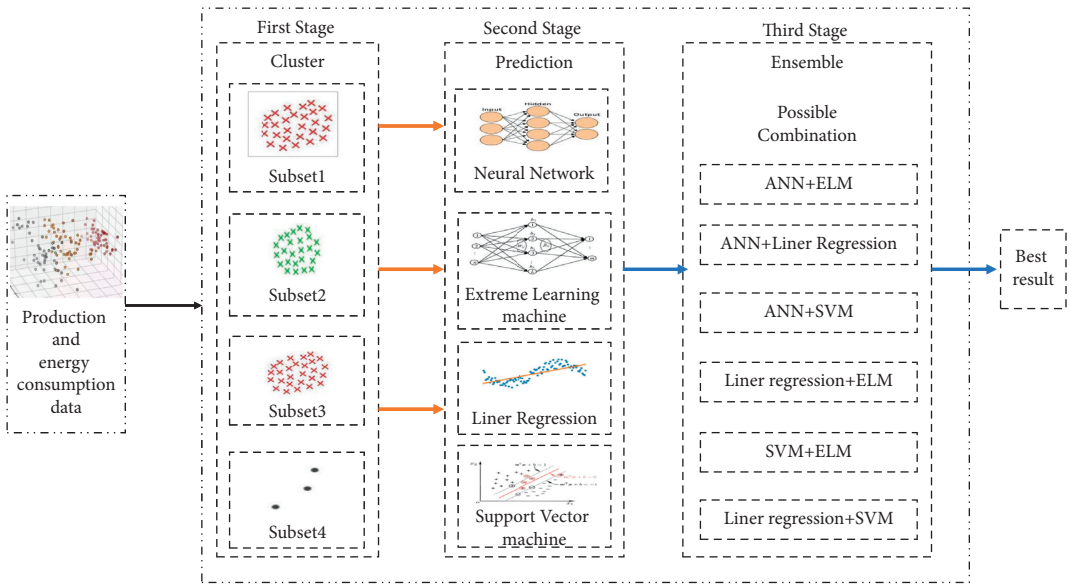

Figure 2: Main steps of the proposed method.

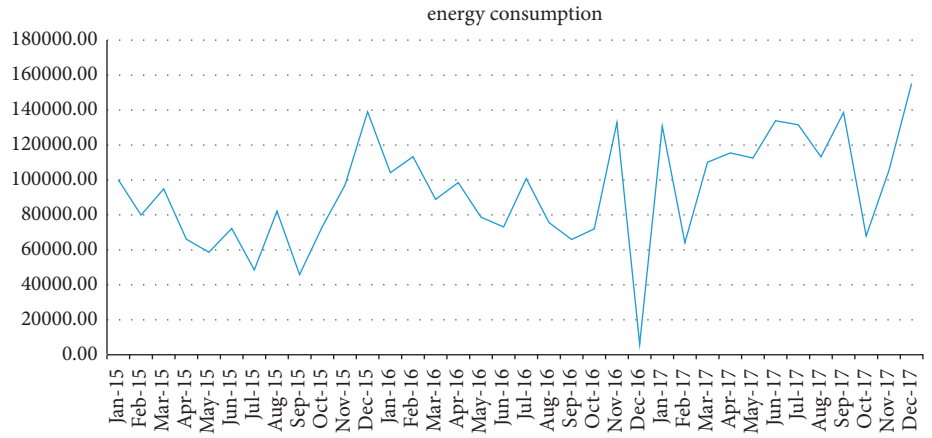

Figure 3: Monthly recorded energy production and consumption data sample.

4.2. Prediction Metrics for Model Performance. In order to evaluate the prediction ability of these models as equidistant as possible, four kinds of metrics are introduced. Firstly, the prediction accuracy is the ability of a metric to forecast with the minimum error, and it can be directly measured by the following three metrics:

$$\text{RMSE} = \sqrt{\frac{\sum_i^N (v_i - v_i')^2}{N}},$$

$$\text{MAE} = \frac{\sum_{i=1}^N |v_i - v_i'|}{N}, \tag{1}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^N \frac{(v_i - v_i')}{v_i} * 100,$$

where N denotes the total number of samples, v_i is the real value, and v_i' is the predicted value. The model's performance is measured by RMSE, MAE, and MAPE. The smaller values of them indicate better performance of these models. Secondly, the correlation coefficient (R) is considered to measure the linear dependence between the real value and the predicted one. The definition is as follows:

$$R(v, v') = \frac{E(v - \mu_v)(v' - \mu_v')}{\sigma_v \sigma_v'}, \tag{2}$$

where E is the expected value operator with standard deviations σ_v and σ_v' and the means μ_v and μ_v'. The correlation coefficient may take any value with the range [−1, 1]. When the correlation coefficient is close to zero, there is no strong linear dependence over the real value and the predicted one. The confidence in a relationship is formally determined not just by the correlation coefficient but also the number of pairs in our data. If there are few pairs, then the correlation coefficient needs to be very close to 1 or −1. More information about the coefficient can be found in [10].

4.3. Clustering. The first step of the proposed model is to classify all these data samples into different subsets and then build a specific prediction model for each data subset. From a statistical point of view, the clustering in this paper is considered as an unsupervised classification problem, and the subset number is unknown. To this end, the following questions should be taken into account: which clustering method should be used to classify these data? How many subsets should be created?

However, there are many advanced approaches to both challenges, and in the light of the previous conducted work, the *K*-means seems to be the right model in this case.

The *K*-means, as a centroid-based clustering method, mainly follows a two-step and iterative calculation: firstly, every observation is assigned to its nearest subset, and then the means are adjusted so that the total within-cluster sum of squared distance is minimized. This process is kept repeated until means remain unmoved.

In our work, before using the *K*-means algorithm, it is a must to confirm the number of clusters in advance, and there are many techniques that can be used to find the optimal count of clusters by calculating the within-cluster and between-cluster distance or using the *R*-squared index to validate the cluster by measuring the homogeneity of clusters. We use the *K*-means to divide the similar dataset into the same group so that the sum of within-cluster distance should be considered. After conducting many processes and with respect to the small dataset size, the best count of subsets is set as 3, and the final number of each subset's member is shown in Table 1.

4.4. Models' Forecasting Result. This section aims to shed light on the results of using regression algorithms to predict the energy consumption, and there are multiple regression models available for regression analysis. In this paper, four different regression models, namely, support vector machine, linear regression, extreme learning machine, and artificial neural network, are discussed with the input of oil production and gas production to predict the energy consumption in the oil and gas company.

The main methodology of the linear regression (LR) is to find the functional relationship with a linear equation among the output and input. Regarding the number of input variables, there are two kinds of linear regression models: simple linear regression and multiple linear regression. In our work, the input variables include oil production and gas production, so we choose the multiple linear regression with equation (1) to fit the relationship between energy production and energy consumption.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n, \quad (3)$$

where Y denotes the output variable, $X_1, X_2, \ldots, X_n$ mean the input variable, n represents the number of input variables, and $\beta_0, \beta_1, \ldots$, are the coefficient of the linear regression. After the training process with the training dataset and the least squares method, the result of the test dataset is presented in Table 2.

Table 1: The number of each subset member.

Subset number	Subset 1	Subset 2	Subset 3
3	329	457	174

Table 2: Test dataset results of linear regression over all subsets.

Linear regression	RMSE	MAE	MAPE	R_2
Subset 1	6.51	5.33	10.32	0.4215
Subset 2	6.52	5.34	10.69	0.4214
Subset 3	6.52	5.34	10.92	0.4121

The artificial neural network (ANN) method has achieved huge success in the field of energy prediction in buildings and shows strong performance in the case of handling nonlinear regression over the input and output. The artificial neural network model captures the feature lying in the dataset through multiple levels of processing layers. In this part, a four-layer artificial neural network is developed to predict the energy consumption, and the input layer includes 2 neurons and is equal to the dimension of input feature vector, the first hidden layer and second hidden layer have 4 and 5 neurons, respectively, and the output layer consists of 1 neuron with respect to the energy consumption. The active function used is sigmoid. Table 3 shows the result of the test dataset.

Extreme learning machine (ELM), as a new learning algorithm, is widely applied for classification and regression with its simplicity and good generalization ability. Compared to the artificial neural network, the extreme learning machine consists of a single hidden layer and requires low computational cost. Concerning the case of energy prediction, many researchers also picked this model to study its performance. In our work, we build the extreme learning machine model with 4 neurons in the hidden layer to predict the energy consumption. Table 4 shows the result of the test dataset with this model.

Support vector machine (SVM), as one of the most robust and accurate data mining algorithms, has been widely applied in many fields such as classification and regression. Due to the strong ability for nonlinear function approximation, SVM is increasingly used to solve nonlinear regression estimation problems. Nowadays, many researchers used the SVM to predict indoor electricity consumption, and the error analyses show that the SVM presents better performance and higher accuracy than the most data mining methods. In this paper, we also tried the SVM to discuss the prediction of energy consumption, and the polynomial kernel function was used to build the SVM model. Table 5 shows the result of the test dataset with the SVM model.

4.5. Hybrid Model Forecasting Result. Many energy consumption prediction problems seem to be too complex with a single machine learning model, so we tried the hybrid method to explore the prediction accuracy. Usually, the hybrid model would be able to compensate the individual model's drawbacks and to offer better performance. Thus, we discuss 6 possible combinations by taking the average of the two models' outputs. To reduce the calculation

Table 3: Test dataset results of the artificial neural network over all subsets.

Artificial neural network	RMSE	MAE	MAPE	R^2
Subset 1	0.42	0.33	4.42	0.9426
Subset 2	0.49	0.37	4.64	0.9377
Subset 3	0.55	0.37	4.74	0.9189

Table 4: Test dataset results of the extreme learning machine over all subsets.

Extreme learning machine	RMSE	MAE	MAPE	R^2
Subset 1	0.51	0.36	4.89	0.9275
Subset 2	0.53	0.38	4.87	0.9192
Subset 3	0.56	0.39	4.90	0.9123

Table 5: Test dataset results of the support vector machine over all subsets.

Support vector machine	RMSE	MAE	MAPE	R^2
Subset 1	3.96	2.83	12.15	0.7150
Subset 2	3.98	2.84	12.26	0.7094
Subset 3	3.98	2.89	12.35	0.7089

Table 6: The RMSE value of these 6 combinations over all subsets' test dataset.

Ensemble	ANN, ELM	ANN, SVM	ANN, LR	SVM, LR	ELM, SVM	ELM, LR
Subset 1	0.41	2.65	4.23	5.34	2.76	4.46
Subset 2	0.48	2.73	4.56	5.21	2.91	4.59
Subset 3	0.55	2.89	4.91	5.97	3.02	5.03

Table 7: Test dataset results of the hybrid model over all subsets.

Ensemble	RMSE	MAE	MAPE	R^2
Subset 1	0.41	0.34	4.41	0.9473
Subset 2	0.48	0.37	4.58	0.9391
Subset 3	0.55	0.36	4.72	0.9201

consumption, we just selected the most representative and commonly used metric, namely, RMSE, rather than four metrics, to evaluate these 6 combinations' performance.

As is shown in Table 6, the hybrid model of the ANN and ELM over all datasets achieves the best performance, and the combination's result is always better than any individual's result. Finally, we choose the hybrid model—ANN and ELM—as our final model, and the hybrid model's results over all subsets' test dataset are shown in Table 7.

5. Conclusion

This paper firstly discussed the four commonly used machine learning methods for energy consumption prediction in the oil and gas industry. The experiment results reveal that most of the analyzed four models can present good performance for energy consumption prediction, and the coefficient of determination for these models is in the range of 0.70 and 0.95 over all the datasets. For the linear regression model, it shows the worst performance with the following reason: the energy consumption dataset is too complex and presents a strong nonlinear relation. The SVM model also shows good performance with the polynomial kernel function, and the coefficient of determination (R^2) is around 0.70 for all these datasets. The ANN and ELM have the highest accuracy and demonstrate the feasibility for the case of energy consumption prediction in oil and gas fields. In order to improve the final energy consumption prediction accuracy, we tried the hybrid method by taking the average of their outputs, and it is also trained with the same training datasets. The final results show the hybrid model would be able to achieve better prediction accuracy than the individual's accuracy. And the hybrid model of the ANN and ELM presents the best performance and is chosen as our final model to predict the total energy consumption in oil and gas fields, and we also installed this model in the energy management system of the oil and gas industry to help the company predict the total energy consumption and improve the energy consumption efficiency.

Data Availability

Some or all data, models, or codes generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] 2019. https://www.bp.com/en/global/corporat e/energy-economics/energy- outlook/introduction/overvie w.html.

[2] J. Gosens, T. Kåberger, and Y. Wang, "China's next renewable energy revolution: goals and mechanisms in the 13th five year plan for energy," *Energy Science and Engineering*, vol. 5, no. 3, 2017.

[3] F. Kaytez, M. C. Taplamacioglu, E. Cam, and F. Hardalac, "Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 431–438, 2015.

[4] M. Raatikainen, J.-P. Skön, K. Leiviskä, and M. Kolehmainen, "Intelligent analysis of energy consumption in school buildings," *Applied Energy*, vol. 165, pp. 416–429, 2016.

[5] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, p. 452, 2018.

[6] Y. Zhang, M. Beaudin, R. Taheri, H. Zareipour, and D. Wood, "Day-ahead power output forecasting for small-scale solar photovoltaic electricity generators," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2253–2262, 2015.

[7] V. Prema and K. U. Rao, "Development of statistical time series models for solar power prediction," *Renewable Energy*, vol. 83, pp. 100–109, 2015.

[8] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied Energy*, vol. 157, pp. 95–110, 2015.

[9] Z. Ramedani, M. Omid, A. Keyhani, B. Khoshnevisan, and H. Saboohi, "A comparative study between fuzzy linear regression and support vector regression for global solar radiation prediction in Iran," *Solar Energy*, vol. 109, pp. 135–143, 2014.

[10] Y. Guo, J. Wang, H. Chen et al., "Machine learning-based thermal response time ahead energy demand prediction for building heating systems," *Applied Energy*, vol. 221, pp. 16–27, 2018.

[11] F. Zhang, C. Deb, and S. E. Lee, "Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique," *Energy and Buildings*, vol. 126, pp. 94–103, 2016.

[12] S. Asadi, S. S. Amiri, and M. Mottahedi, "On the development of multi-linear regression analysis to assess energy consumption in the early stages of building design," *Energy and Buildings*, vol. 85, pp. 246–255, 2014.

[13] N. Fumo and M. A. Rafe Biswas, "Regression analysis for prediction of residential energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332–343, 2015.

[14] E. Worrell and C. Galitsky, "Energy efficiency improvement in the petroleum refining industry," *Lawrence Berkeley National Laboratory*, 2005.

[15] W. Ping, X. Changfang, X. Shiming, and G. Yulin, "Application of energy-saving technology on furnaces of oil refining units," *Procedia Environmental Sciences*, vol. 12, pp. 387–393, 2012.

[16] E. Liu, C. Li, and L. Yang, "Research on the optimal energy consumption of oil pipeline," *[J]*, vol. 36, no. 4, pp. 703–711, 2015.

[17] L. Tan, G. X. Wen, and W. L. Na, "Hybrid modeling method of comprehensive energy consumption for oil and gas production process," *Journal of Northeastern University*, vol. 34, no. 11, pp. 1525–1528, 2013.

[18] C. Temizel, C. H. Canbaz, and Y. Palabiyik, "A comprehensive review of smart/intelligent oilfield technologies and applications in the oil and gas industry," in *Proceedings of the SPE Middle East Oil and Gas Show and Conference. Society of Petroleum Engineers*, Manama, Bahrain, March 2019.

[19] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[20] D. West, S. Dellana, and J. Qian, "Neural network ensemble strategies for financial decision applications," *Computers & Operations Research*, vol. 32, no. 10, pp. 2543–2559, 2005.

[21] C. Potes, S. Parvaneh, and A. Rahman, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proceedings of the 2016 Computing in Cardiology Conference (CinC)*, Columbia, Canada, September 2017.