Research Article
International Journal of Data Science and Applications (JOINDATA) 7(1), 39-47, 2024
Received: 19-Aug-2024 Accepted: 19-Nov-2024
homepage: https://dergipark.org.tr/tr/pub/joindata

SAKARYA UYGULAMALI BİLİMLER ÜNİVERSİTESİ

Forecasting of Natural Gas Supplied for EU in Gazprom with ARIMA and Machine Learning Methods

Tuba KOÇ[1], Emre DÜNDER[2], Haydar KOÇ[3*]

[1] Statistics Department, Science Faculty, Çankırı Karatekin University, e-mail: tubakoc@karatekin.edu.tr
[2,] Statistics Department, Science Faculty, Ondokuz Mayıs University, e-mail: emre.dunder@omu.edu.tr
[3 *] Statistics Department, Science Faculty, Çankırı Karatekin University, e-mail: haydarkoc@karatekin.edu.tr

ABSTRACT

Natural gas, as an environmentally friendly energy source, is gaining importance with its capacity to respond to the increasing energy demand worldwide. Its wide range of uses and low carbon emissions place it at the heart of sustainable energy policies. This critical resource in the energy sector has a dynamic and complex structure under the influence of political, economic, socio-cultural and technological factors. In particular, the forecasting of natural gas supply is of great importance for energy planning and strategic decision making. This study examines different modelling techniques for forecasting Gazprom's natural gas supply. ARIMA, ELM and MLP models are used, and their performance is compared. ARIMA is a classical method often preferred in time series analysis and makes predictions based on past values of the data. ELM is a model based on artificial neural networks and has a fast-learning capability. MLP is a deep learning method with the ability to model complex relationships thanks to its layered structure. As a result of the comparisons, the MLP model was found to perform best. MLP has the lowest error criterion and is more successful than other models in forecasting natural gas supply. This can be attributed to the complexity of the MLP and its strong learning ability.

Keywords: Gazprom; Natural gas; ARIMA; Machine learning.

Introduction

One of the most basic requirements for the economic and social development of countries is energy. For this reason, countries must find energy in uninterrupted, reliable, clean and cheap ways and to diversify these sources. Today, the interdependence between energy, economy and the environment has brought various dimensions to energy. It is very important in terms of environmental policies that natural gas, which creates less pollution than other fossil fuels, is preferred in reducing global warming and climate changes caused by energy use [1]. The role of natural gas in the energy market is very critical in the development of world countries. The entry of natural gas energy, which is the main input of development for the countries of the world, to the sector, took place after the oil crisis. The search for a new energy source in the international energy sector has focused on natural gas, and as a result, the demand for natural gas has increased day by day. Natural gas is one of the most preferred energy sources today because it is a cleaner energy source than other fossil fuels and because of its high reserves in the global arena. Natural gas is transported to consumption centers far from production areas by pipelines or tankers. Pipeline transportation for landlocked countries and tanker transportation for overseas natural gas importing countries is a necessity [2]. A large amount of natural gas is also consumed in Europe. The high demand for natural gas in Europe attracts the attention of major gas supplier countries such as Russia, the United States of America (USA), and Norway [3]. However, among these suppliers, the European Union (EU) has been buying natural gas from Russia for years, because Russia has high reserves, Russian gas is cheap, and Russia is geographically close to Europe. Gazprom, Russia's largest gas company, still exports many of its exports to European countries.

Natural gas has become an important energy source that shapes the world economy. Therefore, accurate natural gas production and consumption estimations are very important for countries to save costs and increase efficiency. In recent years, different natural gas production and consumption estimation methods have been applied continuously. Beyca et al. [4] used multiple linear regression model (MLR), artificial neural network approach (ANN) and support vector regression (SVR) machine learning methods to predict natural gas consumption in Istanbul. They revealed that SVR is superior to ANN technique for time series estimation of natural gas consumption. Aydin [5] used various modelling approaches to model global natural gas production and showed that the S regression model had the best prediction performance among the applied models. Al-Fattah and Startzman [6] presented a new approach to predict natural gas production for the United States using an artificial neural network. Nguyen-Le et al. [7] used two multivariate approaches, a multivariate polynomial approach and a surface methodology approach to develop prediction models for shale gas production. Zheng et al. [8] proposed the confirmatory factor analysis and the Bernoulli equation into the nonhomogeneous gray model to predict natural gas production and consumption in North America. Sen et al. [9] used logarithmic, linear and nonlinear multiple regression methods to estimate natural gas consumption in Turkey. Zhang and Yang [10] used the Bayesian Model Averaging method to estimate natural gas consumption in China. They compared Bayesian Model Averaging method with gray estimation model, Linear regression model and Artificial neural networks and as a result they showed that Bayesian Model Averaging method is a more flexible method for estimating natural gas consumption. Xue et al. [11] used the multi-objective random forest regression method to predict the dynamic production behaviour in shale gas reservoirs. According to Manigandan et al. [12], using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Seasonal ARIMA with eXogenous Factors (SARIMAX) models, estimated monthly natural gas production and consumption in United States-Evidence until 2025. Li et al. [13] proposed a new gray seasonal model with particle swarm optimization to predict monthly natural gas production in China. Liu et al. [14] estimated the natural gas production in China by the optimized nonlinear gray Bernoulli forecasting model and made some reasonable suggestions according to the development trend of natural gas production. On the other hand, Ma et al. [15] used nonlinear autoregressive model, support vector machine, Gaussian process regression and ensemble tree model machine learning methods to predict and compare gas load over the last 3 years based on gas load data in a particular region. Bassey et al. [16] used Random Forest and Artificial Neural Network models to create forecasts for natural gas production. Mao et al. [17] developed educed-order models for upper hydrostatic stress in depleted natural gas reservoirs by utilizing deep neural networks and leveraging extensive reservoir simulation datasets. Singh [18] employed methods like ANN and SVM to estimate the annual natural gas consumption in the USA. In this paper, the focus is on forecasting natural gas supplied from Gazprom using machine learning methods. Since machine learning methods are capable of processing large datasets efficiently, they are well adapted to natural gas production forecasting using a large dataset of daily natural gas supplied from Gazprom. Yamkin et al. [19] used machine-learning models to optimize the process of selecting candidate wells, while using common technologies such as hydrochloric acid treatment of the formation zone at the bottom of the well to increase oil recovery and intensify inflow in oil fields. Anani et al. [20] review the application of machine learning to predict coal and gas explosions in underground mines using a mixed method approach. Schlüter et al. [21] compared the forecasting performance of machine learning methods with ARMA models featuring conditional heteroskedasticity, as well as copula-based time series models, in the context of rising and volatile gas prices in Europe. The focus was particularly on price interval forecasting. This study highlights the growing importance of natural gas as an environmentally sustainable energy source and its ability to meet global energy demand. With its wide range of uses and low carbon emissions, natural gas is at the heart of sustainable energy policies. However, this critical resource has a dynamic and complex structure affected by political, economic, socio-cultural and technological factors. Accurate forecasting of natural gas supply is essential for energy planning and strategic decision-making. The main reason for conducting this study is the increasing uncertainty in energy markets, which highlights the strategic importance of forecasting natural gas supply. In this context, natural gas forecasts were first estimated using ARIMA methods, and then machine learning methods such as ELM and MLP were applied to increase forecast accuracy. The results demonstrate the superior forecast performance of the MLP machine learning algorithm, providing energy managers with actionable insights to optimize Gazprom's supply strategies and increase resilience to fluctuations in the European energy market. The remainder of this paper is organized as follows: Section 2 provides a brief introduction to ELM and MLP machine learning methods and ARIMA. Section 3 presents results and analyses based on machine learning methods and ARIMA. Section 4 concludes the paper.

Research Methodology

1.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA models are models that are applied to non-stationary series but converted to stationary by taking difference. If the degree of the autoregression parameter is *p* and the degree of the moving average parameter is *q* and the difference is made *d* times, this model is called the *(p,d,q)* degree autoregressive integrated moving average model and is expressed as ARIMA *(p,d,q)* [19].

The general ARIMA(p,q,d) model is as follows:

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + \delta + a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \cdots - \Theta_q a_{t-q} \tag{1}$$

Here $Z_t, Z_{t-1}, \ldots, Z_{t-p}$ indicates the d-order differenced observation values, $\Phi_1, \Phi_2, \ldots, \Phi_\mathrm{p}$ are the coefficients for d-order differenced observation values, δ is constant value, $a_t, a_{t-1}, \ldots, a_{t-q}$ are the error term and $\Theta_1, \Theta_2, \ldots, \Theta_q$ shows the coefficients related to the error terms.

In this study, the following machine learning techniques were used for the forecasting time series data.

1.2 Extreme Learning Machine (ELM)

Extreme learning machines are a recommended method for training feedforward neural networks with a single hidden layer [20]. In ELM, the input layer weights, and threshold values are randomly assigned, and the output layer weights are calculated analytically, so it can be trained faster than feedback-learning ANNs and many methods such as support vector machines, and as a result, it has higher generalization ability [21]. The ELM model is as follow

$$\sum_{i=1}^{M} \beta_i g(W_i X_k + b_i) = Y_k\,,\ \ k = 1,2,\ldots,N \tag{2}$$

Where β_i is the output weight, W_i is the input weight matrix, X_k is the input the b_i is the threshold values in the *i*. neuron and Y_k is the output of the network [22].

1.3 Multilayer Perceptron (MLP)

Multilayer Perceptron consists of input, intermediate and output layers. Unlike the single-layer perceptron, the interlayer acts as a bridge between the input layer and the output layer. The interlayer evaluates the inputs from the input layer according to the problem before sending them to the output layer. As a result of the evaluation, a better decision is made according to the problem. The number of intermediate layers can be increased according to the situation of the problem [23]. For MLP networks to be used in time series estimation, the structure of the network must be determined. The relationship between the output value y_t and the inputs $(y_{t-1}, y_{t-2}, \dots, y_{t-N})$ consisting of the past lagged values of the series is as follows

$$y_t = w_0 + \sum_{j=1}^{p} w_j f\left(w_{0j} + \sum_{i=1}^{N} v_{ij} y_{t-i}\right) + e_t \tag{3}$$

Where w_j, v_{ij} are weight values between neurons, p is the number of hidden neurons, f represents the nonlinear activation function used in the hidden layer.

Application

In this study, actual natural gas supplied for EU in Gazprom data was used. The data used is from 1st of July 2016 to 5th of February 2022. Data were obtained from three different supply regions. These are as follows, Supply from Russia and Belarus to Ukraine, Supply from Russia to the EU and Supply from Belarus to the EU.

Table 1: *Descriptive statistics for Actual gas supplied in Gazprom*

Metrics	N	Mean	Std.	Min.	Max.	Skew.	Kurt.	Jarque-Bera	p-value
Actual natural gas supplied in Gazprom (mln, m^3)									
Russia and Belarus to Ukraine	2047	206.71	68.13	38.5	323.4	-0.37	-0.97	534.93	0.0000
Russia to the EU	2047	173.47	36.43	4.1	229.9	-2.22	7.03	5889.8	0.0000
Belarus to the EU	2047	99.69	23.73	4.9	125.7	-2.35	0.05	130.98	0.0000

In Table 1, mean, standard deviation, minimum and maximum, skewness and kurtosis, Jarque-Bera and probability values are given for Actual gas supplied in Gazprom. If the series are normally distributed, the skewness measuring the symmetry must be zero. Otherwise, values greater than zero represent skewness to the right, values less than zero represent skewness to the left. For normally distributed series, the kurtosis value describing the thick-tailed or light-tailed should be three. In Table 1, there is not a dataset whose skewness value is zero and kurtosis value is three. In addition, according to the Jarque-Bera test, the datasets do not show normal distribution. In Figure 1, time series plot of the actual natural gas supplied for EU in Gazprom is given.

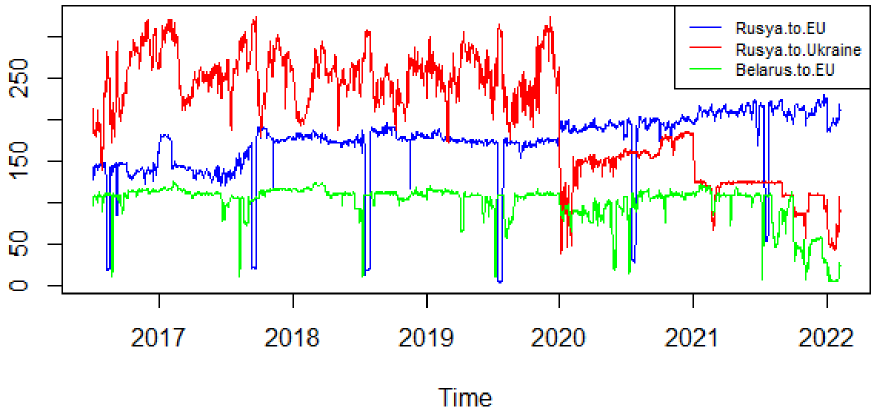

Figure 1: Time series plot of the Actual natural gas supplied for EU in Gazprom

In Figure 1 that the gas supplied from Supply from Russia and Belarus to Ukraine has decreased significantly after 2020. Firstly, ARIMA method and ELM and MLP methods from Machine learning methods were applied to Gazprom datasets. Evaluation metrics for ARIMA, ELM and MLP models were given in Table 2.

In this study, three evaluation metrics are used to compare accuracy performances: Mean Squared Error (MSE), Mean Absolute Error (MAE), and The Root Mean Square Error (RMSE) and their formulations are given below, respectively:

$$MSE = \frac{1}{N}\sum(y_i - \hat{y}) \tag{4}$$

$$MAE = \frac{1}{N}(\sum|y_i - \hat{y}|) \tag{5}$$

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y})} \tag{6}$$

where $\hat{y}$ is the predicted values and y_i is the actual values.

Table 2: *Model Performance Metrics for three different supply regions in Gazprom*

Methods	Russia and Belarus to Ukraine(mln, m^3)			Russia to the EU (mln, m^3)			Belarus to the EU (mln, m^3)		
	MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE
ARIMA	1180.58	11.15	34.35	1131.21	11.21	33.63	503.97	7.84	22.44
ELM	768.25	9.66	27.71	682.14	16.00	26.11	236.09	6.64	15.36
MLP	571.26	7.69	23.90	606.14	14.18	24.62	278.32	7.75	16.68

In Table 2, MLP model reported the lowest MSE of 571.26, the lowest MAE of 7.69 and the lowest RMSE of 23.90 for Russia and Belarus to Ukraine supply dataset. Highest MSE, MAE and RMSE are reported by the ARIMA model. For the Russia to EU supply dataset, MLP has the lowest MSE of 606.14, the lowest MAE of 14.18 and the lowest RMSE of 24.62. For the Belarus to the EU supply dataset, ELM has the lowest MSE of 236.09, the lowest MAE of 6.64 and the lowest RMSE of 15.36. Like the results in other datasets, again the ARIMA has highest MSE, MAE and RMSE. As a result, when the prediction capabilities of ELM, MLP and ARIMA models were compared, the MLP method showed the best performance for all three data sets.

The 90-day forecasting for Gazprom datasets are as given in the time series graphs of Figures 2, 3 and 4.

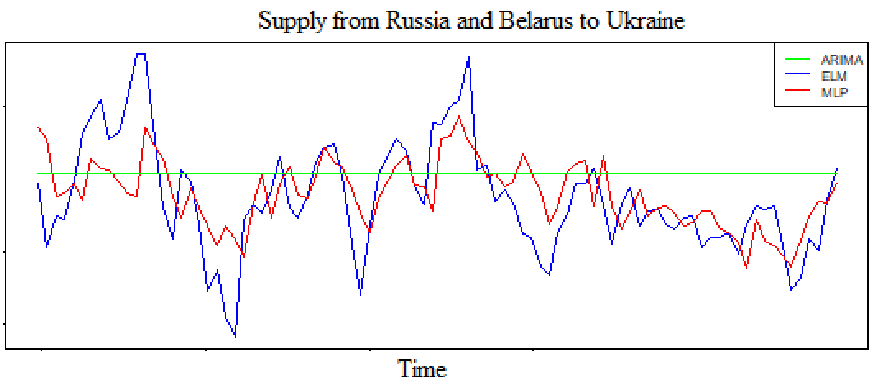

Figure 2: Time series plot of 90-day forecasts for natural gas supply from Russia and Belarus to Ukraine

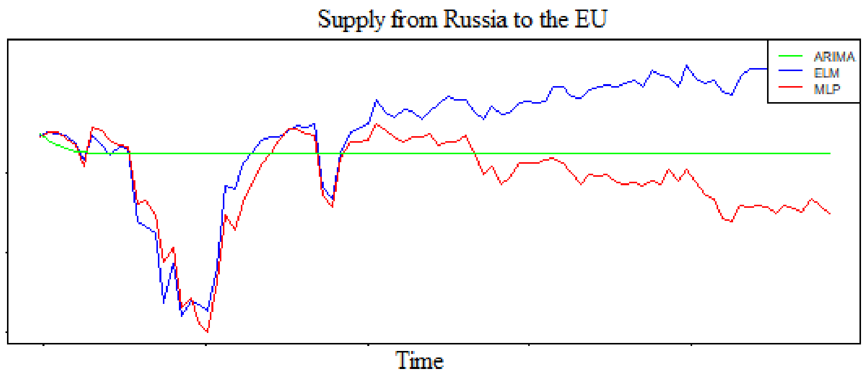

Figure 3: Time series plot of 90-day forecasts for natural gas supply from Belarus to the EU

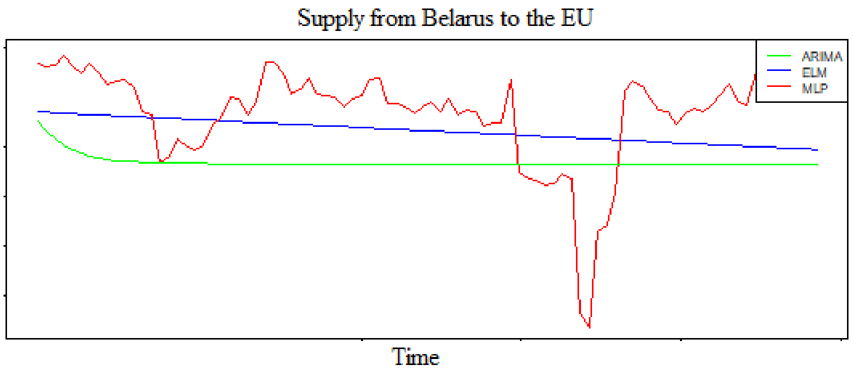

Figure 4: Time series plot of 90-day forecasts for natural gas supply from Russia and Belarus to Ukraine

When the 90-day forecasting in Figures 2, 3 and 4 are examined, it is seen that the ARIMA method found similar forecasting in three datasets. This is expected as the ARIMA method shows the lowest model performance. MLP reached the best results between ARIMA and ELM model. The 20-day forecasting obtained by applying the MLP method to the Gazprom datasets are as in Table 3.

Table 3: *Forecast data for MLP method*

Days	Russia and Belarus to Ukraine	Russia to the EU	Belarus to the EU
	Forecasting values		
1	213.51	88.78	28.40
2	212.75	90.00	28.04
3	208.74	90.09	28.22
4	209.01	88.55	29.18
5	209.68	86.54	28.07
6	208.59	81.61	27.49
7	211.34	91.16	28.42
8	210.68	90.32	27.54
9	210.56	87.97	26.42
10	209.78	86.91	26.59
11	209.00	86.23	26.89
12	208.80	71.97	25.95
13	213.48	72.69	23.67
14	212.21	69.22	23.28
15	211.31	57.50	18.35
16	208.79	60.91	19.13
17	207.33	45.96	20.79
18	209.33	48.38	19.99
19	207.99	41.70	19.72
20	206.80	39.95	20.24

In Table 3, forecasting natural gas supplied for EU in Gazprom for the period from February 7 to February 27 are given.

Conclusions

Natural gas is crucial to the global energy market and the development of many countries, with European countries relying on Russia for more than a third of their supply. This study compared forecasting methods for Gazprom's natural gas supply using ARIMA, ELM and MLP models. The MLP model achieved the best results with the lowest error metrics due to its ability to model complex, non-linear relationships in the data. This ability allows MLP to adapt to market changes more effectively than traditional statistical methods, making it particularly suitable for forecasting in volatile conditions.The results have significant practical implications for energy planners and policy makers. The improved accuracy of the MLP model enables better prediction of supply fluctuations, which is crucial for optimising storage, managing supply contracts and preparing for potential disruptions. By incorporating these forecasts into their strategies, planners can improve inventory management and develop strong contingency plans to maintain stable supply and mitigate the risks associated with shortages and price volatility.Future research could further enhance these findings by exploring additional datasets, such as historical data from other gas suppliers or broader market indicators, to refine the forecasting models.

In addition, exploring advanced deep learning approaches or hybrid models that combine different machine learning techniques could provide new insights and improve forecasting flexibility. Moreover, the current situation between Russia and Ukraine, which began in 2022, has had a significant impact on energy trade between Europe and Russia. The European Union has imposed sanctions on Russia, while Russia has restricted gas supplies. Despite these challenges, the study's focus on model comparisons makes the Gazprom case particularly valuable for understanding forecast performance in complex scenarios.

Declarations

1.4 Study Limitations

None

1.5 Funding source

None.

1.6 Competing Interests

There is no conflict of interest in this study.

1.7 Authors' Contributions

T.K., E.D and H.K. contributed equally to this work.

References

[1] Bayraç, H. N. (2018). Uluslararası doğalgaz piyasasının ekonomik yapısı ve uygulanan politikalar. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 13(3), 13-36.

[2] IEA, *Energy for All*, IEA, Paris, (2011). https://www.iea.org/reports/energy-for-all

[3] Cameron, P. D., & Brothwood, M. (2002). Competition in energy markets: law and regulation in the European Union. Oxford University Press, USA.

[4] Beyca, O. F., Ervural, B. C., Tatoglu, E., Ozuyar, P. G., & Zaim, S. (2019). Using machine learning tools for forecasting natural gas consumption in the province of Istanbul. *Energy Economics*, *80*, 937-949.

[5] Aydin, G. (2015). Forecasting natural gas production using various regression models. *Petroleum Science and Technology*, *33*(15-16), 1486-1492.

[6] Al-Fattah, S. M., & Startzman, R. A., (2001). In SPE hydrocarbon economics and evaluation symposium, OnePetro.

[7] Nguyen-Le, V., Kim, M., Shin, H., & Little, E. (2021). Multivariate approach to the gas production forecast using early production data for Barnett shale reservoir. Journal of Natural Gas Science and Engineering, 87, 103776.

[8] Zheng, C., Wu, W. Z., Xie, W., & Li, Q. (2021). A MFO-based conformable fractional nonhomogeneous grey Bernoulli model for natural gas production and consumption forecasting. Applied Soft Computing, 99, 106891.

[9] Sen, D., Günay, M. E., & Tunç, K. M. (2019). Forecasting annual natural gas consumption using socio-economic indicators for making future policies. Energy, 173, 1106-1118.

[10] Zhang, W., & Yang, J. (2015). Forecasting natural gas consumption in China by Bayesian model averaging. Energy Reports, 1, 216-220.

[11] Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., & Lei, G. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression. Journal of Petroleum Science and Engineering, 196, 107801.

[12] Manigandan, P., Alam, M. S., Alharthi, M., Khan, U., Alagirisamy, K., Pachiyappan, D., & Rehman, A. (2021). Forecasting natural gas production and consumption in United States-evidence from SARIMA and SARIMAX models. Energies, 14(19), 6021.

[13] Li, N., Wang, J., Wu, L., & Bentley, Y. (2021). Predicting monthly natural gas production in China using a novel grey seasonal model with particle swarm optimization. Energy, 215, 119118.

[14] Liu, C., Lao, T., Wu, W. Z., Xie, W., & Zhu, H. (2022). An optimized nonlinear grey Bernoulli prediction model and its application in natural gas production. Expert Systems with Applications, 194, 116448.

[15] Ma, D., Wu, R., Li, Z., Cen, K., Gao, J., & Zhang, Z. (2022). A new method to forecast multi-time scale load of natural gas based on augmentation data-machine learning model. Chinese Journal of Chemical Engineering, 48, 166-175.

[16] Bassey, M., Akpabio, M. G., & Agwu, O. E. (2024). Enhancing Natural Gas Production Prediction Using Machine Learning Techniques: A Study with Random Forest and Artificial Neural Network Models. In *SPE Nigeria Annual International Conference and Exhibition* (p. D021S010R004). SPE.

[17] Mao, S., Chen, B., Malki, M., Chen, F., Morales, M., Ma, Z., & Mehana, M. (2024). Efficient prediction of hydrogen storage performance in depleted gas reservoirs using machine learning. *Applied Energy*, *361*, 122914.

[18] Singh, S., Bansal, P., Hosen, M., & Bansal, S. K. (2023). Forecasting annual natural gas consumption in USA: Application of machine learning techniques-ANN and SVM. *Resources Policy*, *80*, 103159.

[19] Ямкин, М. А., Сафиуллина, Е. У., & Ямкин, А. В. (2024). Machine learning methods for selecting candidate wells for bottomhole formation zone treatment. Bulletin of the Tomsk Polytechnic University Geo Assets Engineering, 335(5), 7-16.

[20] Anani, A., Adewuyi, S. O., Risso, N., & Nyaaba, W. (2024). Advancements in machine learning techniques for coal and gas outburst prediction in underground mines. *International Journal of Coal Geology*, 104471.

[21] Schlüter, S., Pappert, S., & Neumann, M. (2024). Interval Forecasts for Gas Prices in the Face of Structural Breaks--Statistical Models vs. Neural Networks. *arXiv preprint arXiv:2407.16723*.

[22] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

[23] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: theory and applications. Neurocomputing, 70(1-3), 489-501.

[24] Ertuğrul, Ö. F., & Kaya, Y. (2014). A detailed analysis on extreme learning machine and novel approaches based on ELM. American Journal of computer science and engineering, 1(5), 43-50.

[25] Akın, P. (2021). Comparing extreme learning machine and multilayer perception: tourism data as an example. Research & Reviews in Science and Mathematics, 65.

[26] Torres, R. A., & Hu, Y. H. (2013). Prediction of NBA games based on Machine Learning Methods. University of Wisconsin, Madison.