Chemical and Process Engineering: New Frontiers, 2026, 47(2), e136
DOI: 10.24425/cpe.2026.158134

MLP artificial neural network model for predicting natural gas composition variability

Jolanta Szoplik, Paulina Muchel

West Pomeranian University of Technology in Szczecin, Faculty of Chemical Technology and Engineering, Piastów 42, 71-065 Szczecin, Poland

Abstract

This paper proposes an artificial neural network model, MLP 17-37-5, to predict the share of five natural gas components (methane, ethane, propane, nitrogen, and carbon dioxide) in the natural gas mixture in a pipeline network, depending on selected calendar and weather factors. Natural gas composition variability in the pipeline network results from supplying it with gas of varying composition and from different suppliers. Using statistical analysis of 35,064 actual measurement data sets, factors (model input data) significantly affecting gas composition variability were selected. Models differing in structure (the MLP 17-37-5 model or five MLP 17-37-1 models) and the number of neurons in the hidden layer (from 20 to 230 neurons) were trained using sets ranging from 8,760 to 35,064 actual data points obtained using the chromatographic method. The quality of the models was assessed based on the correlation coefficient, while the quality of the forecasts was assessed based on the nRMSE error of the forecasts obtained for a new data set of 8,760 data points. The MLP 17-37-5 model was shown to predict natural gas composition with an average error of nRMSE = 0.321. The article proposes a ready-made tool, which has no equivalent in the literature, allowing for a significant reduction in the number of chromatographic tests performed to determine the composition of natural gas. This is a completely new approach to studying changes in the composition of natural gas over time, which in the proposed forecasting model depend on selected factors (not analyzed in chromatographic studies).

*** Corresponding author, e-mail:**
jolanta.szoplik@zut.edu.pl

Article info:
Received: 11 April 2026
Revised: 25 May 2026
Accepted: 29 May 2026

Keywords
statistical data analysis, natural gas composition variability, MLP artificial neural network model, natural gas composition forecasting

1. INTRODUCTION AND STATE OF THE ART

Natural gas, a mixture of hydrocarbons, is transported from its point of extraction to end users (individual consumers or industry) via a suitably designed pipeline network. In reality, the pipeline network transports natural gas energy, the amount of which depends on the composition of natural gas and can be expressed as the heat of combustion value. The composition of natural gas depends, among other things, on the gas extraction location and the degree of purification of the gas fuel. Additionally, changes in the composition of natural gas in network pipelines can be influenced by supplying pipelines with gas from different suppliers, regasified LNG, or the introduction of alternative fuels (hydrogen or biomethane), mixing gases with different chemical compositions, and variability in the load on individual network pipelines resulting from uneven demand for gas from customers over time (Szoplik, 2015; Szoplik, 2016; Szoplik, 2017; Szoplik and Muchel, 2023a; Szoplik and Muchel, 2023b). In such a situation, the only way to obtain information on the composition of natural gas at a given point in the gas network is to monitor it in real time using gas chromatography method or to track changes in gas composition based on transient simulations of gas flow in the network or to forecast the composition of natural gas using artificial neural networks.

Direct measurements using gas chromatography method are conducted on an ongoing basis by gas network operators at gas stations located at multiple points in the gas network and provide information on the current composition of natural gas at a given measurement point. These measurements can be used to analyze the flow of gas of a given composition in individual gas network pipelines. Kumor and Porada (2021) analyzed the impact of the operation of the LNG terminal in the Polish gas transportation system on the quality of gaseous fuel at 16 selected points in the high-pressure gas network over a 5-year period (covering the days before and after the LNG terminal was launched). Selected results of chromatographic measurements of changes in natural gas composition recorded at selected points in the gas network of the Polish natural gas transportation system, supplied by several suppliers and the LNG terminal, were also analyzed in (Muchel, 2024; Szoplik and Muchel, 2023a).

Another method of obtaining data on natural gas composition is to track changes in natural gas composition in network pipelines based on gas flow simulation results. This method is based on solving selected model equations describing the processes occurring within pipelines and the topology of the pipeline network for known flows supplying the network and the gas flow rates received by end users. Variations in demand for gaseous fuel by customers do not affect changes in the gas composition in a pipeline if it is supplied by a single supplier with gas of constant composition over time. However, if streams of other gaseous fuels (hydrogen or biomethane), for example, from alternative sources, are occasionally introduced into the gas network, the uneven load on network pipelines is an additional factor influencing changes in gas composition over time. Simulation models of gas flow in the network enable tracking changes in the composition of gas mixtures with different chemical compositions both under variable gas demand conditions and under conditions of hydrogen or biomethane injection into the pipeline (Alves and Fontes, 2022; Bermúdez and Shabani, 2022; Chaczykowski and Zarodkiewicz, 2017; Chaczykowski et al., 2018; Fan et al., 2021; Osiadacz and Chaczykowski, 2020; Saedi et al., 2021; Zhang et al., 2022). Variations in gas quality over time, expressed for example by the gas's heat combustion value, caused by changes in its composition, adversely affect the maximum load of the gas transport system and the determination of its capacity. Tracking changes in gas composition or the gas's heat combustion value is essential for planning the introduction of alternative gaseous fuels into the gas network as occasional supplies and for monitoring the presence of contaminants in the gas that are corrosive or susceptible to hydrate formation during transport.

The results of simulations of the flow and mixing of gas streams extracted from onshore or offshore deposits and liquefied LNG or compressed CNG were analyzed in (Alves and Fontes, 2022). Based on the shares of the most important components of the resulting gas mixture, the heat of combustion, which is one of the parameters necessary for system balancing, was determined. The research results obtained from gas flow simulations in the Brazilian gas network allowed for a reduction in gas transportation costs by up to 22.6%.

In a paper (Bermúdez and Shabani, 2022), a multi-species model consisting of first-order partial differential equations for each pipeline and gas component was proposed to track changes in gas composition at any location and time. A characteristic feature of this model is the separate solution of momentum and mass exchange equations using two independent computational grids. Numerical calculation results, verified with real-world data, demonstrated good agreement between simulation results and actual data. According to the authors, in times of energy transformation, there is considerable interest in the introduction and transport of hydrogen or biomethane into natural gas networks, which can be analyzed, for example, using the proposed methodology.

Chaczykowski and Zarodkiewicz (2017), Chaczykowski et al. (2018) and Osiadacz and Chaczykowski (2020), taking into account the direction of changes in the natural gas transportation system and the introduction of alternative gaseous fuels, point to new challenges facing the gas industry in a multi-energy system environment. These changes are expected to cause significant fluctuations and uncertainty in the distributed supply of liquefied natural gas (LNG) or gases from renewable sources (hydrogen). Assuming non-isothermal flow conditions, the authors proposed models for tracking gas quality, assuming energy boundary conditions (energy flow rate instead of volumetric flow), for an onshore gas pipeline in the Polish transmission system and for a model of an offshore gas pipeline in the Norwegian transmission system. Differences between measured and modeled compositions and transport times were examined and it was found that both methods accurately reproduced the measured compositions and transport times. According to the authors, only a well-chosen model will provide good simulation results and facilitate the integration of the power system with the gas system.

In another paper (Fan et al., 2021), a control volume method for hydraulic simulation and a one-dimensional, non-steady-state heat transfer model for thermal simulation were proposed for tracking gas composition in real time and, based on this, calculating gas flow parameters in the network. The pipeline system in the network was represented using a special grid connecting all adjacent pipelines. The proposed model is intended to be a tool for tracking gas composition in the event of injection of new types of gas (LNG, biogas, or hydrogen), which can cause short-term and rapid fluctuations in natural gas composition. Based on a comparison of the calculation results with measurement data, the authors indicate good agreement between the gas composition simulation results and the actual data.

Saedi et al. (2021) proposed a model for tracking gas composition in an integrated power and gas system with green hydrogen injection. The model also allows for the identification of gas flow directions entering and leaving a given network node, which is important in the case of a distributed hydrogen injection system and constraints related to the size of the hydrogen flow entering the network or the share of hydrogen in the gas flow at the customer. The model was validated using two gas networks of different sizes and demonstrated that it can accurately capture gas flows entering and leaving a network node and indirectly allows for the estimation of the amount of hydrogen that can be injected into the network depending on the season, time of day, or location in the power and gas grid.

The model proposed by Zhang et al. (2022) enables tracking of gas composition, and in particular the hydrogen content of the gas stream, in a real gas network with a branched structure and multiple non-pipe elements, supplied with natural gas at multiple points, and with multiple intermittent hydrogen injection sources. The model takes into account temporal and spatial variability of gas physical properties (gas compressibility coefficient). The model was verified using a selected Belgian gas network as an example, and the effectiveness of this method in solving natural gas composition tracking problems was demonstrated. Simulation results also confirmed the impact and significance of introducing a variable gas compressibility coefficient, tracking hydrogen fraction, and varying gas demand.

Simulation methods for tracking changes in the composition of mixtures of natural gas, regasified LNG, hydrogen, or biomethane require the use of equations describing the exchange of mass, momentum, and energy within gas network pipelines. Studies of gas composition tracking have most often used a straight pipeline section, less frequently a branched network. However, in reality, the pipeline network has a more complex topology (it is branched and contains loop elements), and depending on the network pipeline load, a change in the gas flow direction is observed in some pipelines. A change in the gas flow direction in the pipeline may be caused, for example, by greater or lesser gas demand from consumers during periods of lower ambient temperatures (apartment heating) or higher ambient temperatures (apartment cooling). The variability of gas demand from consumers depending on selected weather, calendar, economic, or demographic factors was analyzed in (Bianco et al., 2014; Durmayaz et al., 2000; Geem and Roper, 2009; Kenisarin and Kenisarina, 2007; Parikh et al., 2007; Sarak and Satman, 2003; Satman and Yalcinkaya, 1999; Szoplik, 2015; Szoplik and Muchel, 2023a).

Another approach to forecasting the natural gas composition at a given point in the pipeline network is to use an artificial neural network model. In this approach, one must first identify factors that directly or indirectly influence changes in natural gas composition (using statistical analysis), and then train the artificial neural network model using a set of real data describing the natural gas composition (chromatographic methods). In studies of natural gas transportation systems, artificial neural network models have been used, among others, to forecast gas demand on a global scale (Hafezi et al., 2019), for a given country (Merkel et al., 2018), or city (Szoplik, 2015), to predict the pipeline's resistance to cracking (Wang et al., 2022) or the corrosion rate of steel pipelines (Wen et al., 2019).

However, for designers and operators of natural gas transportation systems, it is essential to know the properties of natural gas, whose changes depending on the natural gas composition, pressure, and temperature can also be predicted using artificial neural network models. Chu et al. (2021) predicted the compressibility coefficient, sound velocity, and viscosity of natural gas, while Farzaneh-Gord et al. (2018) predicted the compressibility coefficient and sound velocity. The prediction of density (Al-Quraishi and Shokir, 2011; Wood and Choubineh, 2020) and viscosity of natural gas (Al-Quraishi and Shokir, 2011; Lay et al., 2012) using an artificial neural network model was based on the dependence of these parameters on pressure, temperature, and natural gas composition. The authors considered different numbers of gas components in the forecasting model. To predict the compressibility coefficient, sound velocity, and gas viscosity, the model included 10 gas components (Chu et al., 2021), while in (Lay et al., 2012), the model for predicting gas viscosity included 12 natural gas components and 21 gas components for predicting the compressibility coefficient and gas viscosity (Farzaneh-Gord et al., 2018).

Artificial neural network models have been successfully used to predict the heat of combustion or composition of natural gas mixtures produced in gas network pipelines supplied by various suppliers (Szoplik and Muchel, 2023a; Szoplik and Muchel, 2023b), the composition of liquid-gas mixtures produced in a distillation column (Singh et al., 2007), the composition of mixtures produced in the gasification process of various types of biomass (George et al., 2018), plant biomass components (Postawa et al., 2022), and reforming gas components in diesel engine exhaust (Park et al., 2020). In the analyzed cases, an MLP model with one or two hidden layers and the number of neurons in the output layer equal to the number of predicted components was used. The quality of the trained models was assessed based on the correlation or regression coefficient, while the quality of the forecasts was assessed based on the MAPE or RMSE errors. Analysis of the literature data has revealed numerous cases of artificial neural network models being used to predict the properties and composition of process mixtures. However, only Szoplik and Muchel (2023a) attempted to predict the share of five natural gas components at a selected point in a pipeline network supplied with gas of varying quality at multiple points. Szoplik and Muchel (2023a) proposed the use of the MLP 18-65-5 model trained on a set of 8,760 actual natural gas composition measurements obtained using gas chromatography method. The quality of the forecasts was assessed based on the average component-weighted MAPE error.

The aim of the research presented in this paper is to develop an MLP model for forecasting the shares of five natural gas mixture components at a selected point in the Polish gas transportation system, taking into account the minimum number of weather or calendar factors influencing the variability of the natural gas mixture composition. The selection of factors influencing the gas composition will be based on a statistical analysis of the relationships between these factors and the forecasted components. The determined number of factors influencing the natural gas composition will be the basis for determining the number of neurons in the MLP model's input layer, while the number of neurons in the model's output layer is five, equal to the number of forecasted components. The number of neurons in the hidden layer will be selected experimentally based on the results of the correlation coefficient of the model, the quality of the representation of the relationships between the forecasted component shares, and the mean $nRMSE_{aver}$ forecast errors. Training of MLP models with different structures to select the best model quality was conducted on the basis of varying sizes of real data sets presenting the change in natural gas composition in individual hours between 2017 and 2020.

2. STATISTICAL ANALYSIS OF THE INPUT AND OUTPUT DATA OF THE MLP MODEL

Developing a high-quality artificial neural network forecasting model requires statistical analysis of the correlation between factors that influence the predicted value, as well as analysis of the correlation between factors influencing the predicted value and the predicted values. It was assumed that there should be no correlation between factors selected as model input parameters and that there should be a correlation between the model input parameters and the values predicted by the model. Statistical analysis was performed based on a dataset of $N_4 = 35,064$ datasets representing the shares of five selected natural gas components (forecasted values, model output data) in consecutive hours over four calendar years (2017–2020) at one of the natural gas composition measurement points in the Polish gas transportation system. The following factors influencing the natural gas mixture composition (model input data) were selected: month (M), day of the month (d/M), day of the week (d/W), hour of the day (h), ambient temperature (T), and gas flow rate (Q). Correlation analysis between input data, and then between input and output data, was performed by determining the Pearson correlation coefficient at a significance level of $\alpha = 5\%$. It was assumed that the correlation between variables is statistically significant when the p-test statistic is less than the assumed significance level ($p \leq \alpha$). However, if the p-test statistic is greater than the α-significance level, it indicates a lack of significant relationship between the two features.

The selection of natural gas mixture components, whose share in natural gas depends indirectly on gas consumption by consumers and may change over time, was analyzed in Szoplik and Muchel (2023a). Based on the significance of differences between the mean shares and standard deviation, it was demonstrated that the relationships between the five natural gas components selected for forecasting (CH_4, C_2H_6, C_3H_8, N_2, and CO_2) were statistically significant. The strongest positive correlation was observed between the shares of methane and carbon dioxide, while the strongest negative correlation was found between ethane and carbon dioxide. This is also confirmed by the results of Pearson correlation coefficient calculations determined for all pairs of the five selected natural gas components, which are summarized in Table 1. Analysis of the co-occurrence of selected gas component shares in the natural gas mixture transported through the pipeline network will be one of the criteria for assessing the quality of forecasts of these natural gas components obtained using the trained MLP model.

The selection of factors significantly affecting the composition of the natural gas mixture at a given measurement point on the high-pressure pipeline was based on an analysis of correlations between six factors that could affect the composition of natural gas, followed by a correlation between these six factors and the proportion of each of the five natural gas components. The analysis resulted in the selection of factors that lack significant correlation (model input data) but that significantly influence the predicted quantities. This ensures that the MLP neural network model will only contain input variables that significantly influence the predicted proportions of natural gas composition.

The results of correlation analysis among the model input data (month, day of the month, day of the week, hour of the day, ambient temperature, and gas flow rate) are summarized in Table 2. These variables can be divided into two groups: quantitative variables expressed numerically (day of the month, hour of the day, temperature, and volumetric flow rate) and qualitative variables expressed descriptively (month and day of the week). Temperature and flow rate were shown to be moderately correlated (-0.4495), which allows for the elimination of one of these variables. Weak correlations were also observed between temperature and month (0.1992) and temperature and hour of the day (0.1728), which may be due to temperature changes depending on the season (month) and time of day (hour of the day). Taking the above into account, the model can be simplified by eliminating temperature (T) or flow rate (Q) and month (M) or hour of the day (h) from the input data.

Table 1. Pearson correlation coefficients (top row) and p-test statistics (bottom row) for the model output variables ($\alpha = 0.05$).

Component	CH_4	C_2H_6	C_3H_8	N_2	CO_2
CH_4	1.000	–0.5464	–0.2722	–0.2245	0.7570
	$p = —$	$p = 0$	$p = 0$	$p = 0$	$p = 0$
C_2H_6	–0.5464	1.000	–0.5735	–0.6639	–0.6075
	$p = 0$	$p = —$	$p = 0$	$p = 0$	$p = 0$
C_3H_8	–0.2722	–0.5735	1.000	–0.7791	–0.0048
	$p = 0$	$p = 0$	$p = —$	$p = 0$	$p = 0.365$
N_2	–0.2245	–0.6639	0.7791	1.000	–0.0009
	$p = 0$	$p = 0$	$p = 0$	$p = —$	$p = 0.0281$
CO_2	0.7570	–0.6075	–0.0048	0.0281	1.000
	$p = 0$	$p = 0$	$p = 0.365$	$p = 0$	$p = —$

The results of the correlation coefficient analysis between the model's input and output variables are summarized in Table 3, which clearly indicate a lack of a statistically significant relationship between the time of day and the shares of natural gas components. Assuming that a correlation should be present between the input and output variables of the forecasting model, the model was simplified by eliminating the hour of day (h) parameter from the input data. Based on the correlation coefficient analysis, it was assumed that it would be reasonable to exclude the gas flow rate (Q) and hour of day (h) from the input data of the MLP forecasting model, as these parameters have too strong a correlation with ambient temperature, and additionally, there is no correlation between the hour of day and natural gas components. Based on the results of statistical analysis of the structure of the set of 35,064 results of the share of methane, ethane, propane, nitrogen and carbon dioxide in natural gas depending on calendar and weather factors and gas flow, the following input parameters were selected for the forecasting model: month M, day of the month d/M, day of the week d/W and ambient temperature T, which, among the analyzed factors, have a statistically significant impact on the shares of natural gas components at a given composition measurement point in the high-pressure pipeline.

Table 2. Pearson correlation coefficients (top row) and p-test statistics (bottom row) for model input variables ($\alpha = 0.05$).

Variable	d/W	d/M	M	h	T	Q
d/W	1.000	0.0293	–0.0257	–0.0001	0.0317	0.0756
	$p =$ —	$p = 0$	$p = 0$	$p = 0.986$	$p = 0$	$p = 0$
d/M	0.0293	1.000	0.0105	0.000	0.0204	0.0006
	$p = 0$	$p =$ —	$p = 0.049$	$p = 1.000$	$p = 0$	$p = 0.918$
M	–0.0257	0.0105	1.000	0.000	0.1992	0.0004
	$p = 0$	$p = 0.049$	$p =$ —	$p = 1.000$	$p = 0$	$p = 0.940$
h	–0.0001	0.000	0.000	1.000	0.1728	0.0691
	$p = 0.986$	$p = 1.000$	$p = 1.000$	$p =$ —	$p = 0$	$p = 0$
T	0.0317	0.0204	0.1992	0.1728	1.000	–0.4495
	$p = 0$	$p = 0$	$p = 0$	$p = 0$	$p =$ —	$p = 0$
Q	0.0756	0.0006	0.0004	0.0691	–0.4495	1.000
	$p = 0$	$p = 0.918$	$p = 0.940$	$p = 0$	$p = 0$	$p =$ —

Table 3. Pearson correlation coefficients (top row) and p-test statistics (bottom row) for model input and output variables ($\alpha = 0.05$).

Variable	CH_4	C_2H_6	C_3H_8	N_2	CO_2
d/W	0.0566	–0.0144	–0.0258	–0.0370	0.0556
	$p = 0$	$p = 0.007$	$p = 0$	$p = 0$	$p = 0$
d/M	0.0574	–0.0154	–0.0684	–0.0107	0.0466
	$p = 0$	$p = 0.004$	$p = 0$	$p = 0.046$	$p = 0$
M	0.0364	0.0141	–0.0654	–0.0374	–0.0238
	$p = 0$	$p = 0.008$	$p = 0$	$p = 0$	$p = 0$
h	0.0008	–0.0001	–0.0004	–0.0003	–0.0009
	$p = 0.882$	$p = 0.989$	$p = 0.944$	$p = 0.954$	$p = 0.866$
T	–0.0424	0.1925	–0.1514	–0.1905	–0.1352
	$p = 0$	$p = 0$	$p = 0$	$p = 0$	$p = 0$
Q	0.0182	–0.1494	0.1367	0.1554	0.0183
	$p = 0.001$	$p = 0$	$p = 0$	$p = 0$	$p = 0.001$

3. RESEARCH METHODOLOGY

To predict the composition of the natural gas mixture, an MLP artificial neural network model with a single hidden layer was selected, in which signals flow in one direction from the input layer X to the output layer Z (Fig. 1). Development of the MLP artificial neural network model involves determining the number of neurons in each of the model layers (input, hidden and output). The number of neurons in the input layer was selected based on a statistical analysis of factors influencing the change in the natural gas mixture composition. The developed model uses a total of 17 neurons in the input layer of the MLP model, containing 15 qualitative variables representing 3 types of day of the week (working D_1, holiday D_2 and weekend D_3 days) and 12 months of the year ($M_1, \ldots, M_{12}$) and 2 quantitative variables corresponding to the day of the month (d/M) and the ambient temperature (T). In the output layer of the MLP model, 5 neurons were used, corresponding to the shares of methane y_{CH_4}, ethane $y_{C_2H_6}$, propane $y_{C_3H_8}$, nitrogen y_{N_2} and carbon dioxide y_{CO_2} in natural gas.

The number of neurons in the hidden layer of the MLP model was selected experimentally based on the analysis of the correlation coefficient results and the mean errors of $nRMSE_{aver}$ forecasts performed using models trained on data sets of different sizes (8,760 – 35,064) and the analysis of the predicted co-occurrence of selected components in natural gas.

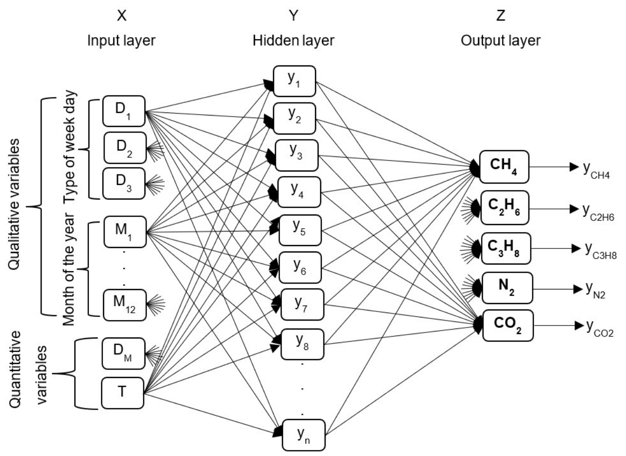


Figure 1. Schematic diagram of the MLP 17-y_n-5 model with any number of neurons in one hidden layer.

The quality of natural gas composition forecasts using MLP models was assessed based on the average $nRMSE_{\text{aver}}$ forecast error determined from Equations (1)–(3), based on the predicted and actual shares of selected natural gas components (i: y_{CH_4}, $y_{C_2H_6}$, $y_{C_3H_8}$, y_{N_2}, y_{CO_2}). The mean normalized $nRMSE_{\text{aver}}$ forecast error was used instead of the popular $MAPE_{\text{aver}}$ error due to the different values of the predicted shares of natural gas components (from 0 to 0.97).

$$nRMSE_{\text{aver}} = \frac{\sum\limits_{i=1}^{i=5} nRMSE_i}{5} \tag{1}$$

$$nRMSE_i = \frac{RMSE_i}{y_{i,\text{exp,max}} - y_{i,\text{exp,min}}} \tag{2}$$

$$RMSE_i = \sqrt{\frac{1}{N}\sum \left(y_{i,\text{exp}} - y_{i,\text{MLP}}\right)^2} \tag{3}$$

The model was trained in Statistica using datasets of various sizes ($N_1 \div N_4$), including the shares of the five natural gas components (model output) and previously selected weather and calendar factors (model input). The data set sizes used in the model training process are summarized in Table 4.

Neural network training included: a network model training process involving modification of the weights of the hidden and output layer neurons; a validation process for ongoing monitoring of the training process; and a testing process used to assess the network's quality after training. During the neural network training process, each dataset was randomly divided into three subsets: a training subset (L) containing 70% of the data, a validation subset (V) containing 10% of the data and a test subset (T) containing 20% of the data from the entire N dataset. The training process involved selecting network neuron parameters in subsequent iterations (epochs) using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) variable metric method, testing various activation functions for the hidden and output layers (linear, logistic, or hyperbolic tangent). For each dataset, differing in the number of cases, 200 MLP 17-y_n-5 models were trained for different numbers of neurons in the hidden layer in the range $y_n \in (30 \div 230)$. Models with the highest values of correlation coefficients R were selected for further analyses and comparisons (R_L, R_V, R_T).

4. RESEARCH RESULTS

The development of the MLP 17-y_n-5 model for forecasting the shares of five selected natural gas components was based on experimental selection of the number of neurons in the hidden layer of the model, trained on various size datasets. Additionally, the forecast results for the shares of the five natural gas components, obtained using the best-quality MLP 17-y_n-5 model, would be compared with the forecast results for the natural gas components obtained individually using five MLP 17-y_n-1 models. The best-quality model would be selected based on forecast results assessed based on the nRMSE forecast error. Model training and verification, as well as the forecasting of the gas component shares, were based on real data representing the natural gas composition at successive hours of the day over a five-year period, provided by the GAZ-SYSTEM S.A. – Gas Transmission Operator of the Polish natural gas network under a mutual cooperation and information exchange agreement.

4.1. Selecting the size of the training dataset and the number of neurons in the hidden layer of the MLP model

To select the number of neurons in the model's hidden layer and the size of the training dataset, a total of 16,000 MLP models were trained, differing in the number of neurons in the hidden layer ($y_n = 30 \div 230$ neurons) and the size of the training dataset (from $N_1 = 8,760$ to $N_4 = 35,064$ data). The results, in the form of correlation coefficients R_T and the prediction error nRMSE as a function of the number of neurons in the model's hidden layer, are presented in Fig. 2 and Fig. 3. Based on the analysis of the data, it was found that the size of the dataset used to train the MLP models has a greater impact on the model quality (expressed by the R coefficient) and the forecast quality (determined by the nRMSE error) than the number of neurons in the model's hidden layer.

The highest values of correlation coefficients for the training, test, and validation sets (R_L, R_T and R_V) are characteristic of the smallest data set $N_1 = 8,760$, while the lowest nRMSE forecast errors correspond to models trained on the largest data set $N_4 = 35,064$. The reduction in the correlation coefficients for models trained with the largest data set can be explained by the presence of data characterizing the COVID-19 pandemic period, which, due to unusual

Table 4. The sizes of data sets used in the MLP model training process.

Period of time	N_1 (2018)	N_2 (2017÷2018)	N_3 (2017÷2019)	N_4 (2017÷2020)	N_5 (2021)
Dataset sizes	8,760	17,520	26,280	35,064	8,760
Purpose	training	training	training	training	verification

changes in gas pipeline load and changes in flow direction in network pipelines (these values have an indirect impact on the composition of natural gas), was not representative of the pre-pandemic period.

Assuming that the best quality model should be characterized by a high correlation coefficient and low forecast error, such a model cannot be clearly identified based on the results presented in Fig. 2 and Fig. 3. Furthermore, there was no clear effect of the number of neurons in the hidden layer on the quality of the model or forecast, therefore, models with the lowest number of neurons in the hidden layer, trained on the N_1 or N_4 set, were selected to analyze the quality of the shares of gas component forecasts. Detailed results in the form of forecast errors for the shares of individual gas components, the average $nRMSE_{aver}$ error, and the correlation coefficient for the test set are summarized in Table 5.

Therefore, the selection of the data set size used to train the MLP model was based on an analysis of the interdependencies between the forecasted natural gas components. Figures 4 and 5 present the results of forecasts of the methane and ethane shares in natural gas, which were made for new data $N_5 = 8,760$ using the MLP 17-40-5 model (trained on N_1 data) and the MLP 17-37-5 model (trained on N_4 data). Analysis of the results in Figures 4 and 5 indicates that the relationships between the two forecasted natural gas components were best modeled using the MLP 17-40-5 model (N_1). The nature of the variability of the individual natural gas components in this case is similar to that observed in the real

Table 5. The summary of forecast error $nRMSE_i$ and $nRMSE_{aver}$ results.

MLP model dataset	MLP 17-40-5 ($N_1 = 8,760$)	MLP 17-37-5 ($N_4 = 35,064$)
component	$nRMSE_i$	$nRMSE_i$
CH_4	0.3135	0.2556
C_2H_6	0.4086	0.2784
C_3H_8	0.3315	0.3155
N_2	0.2986	0.2684
CO_2	0.2575	0.1805
$nRMSE_{aver}$	0.3219	0.2597
R_T	0.9700	0.5700

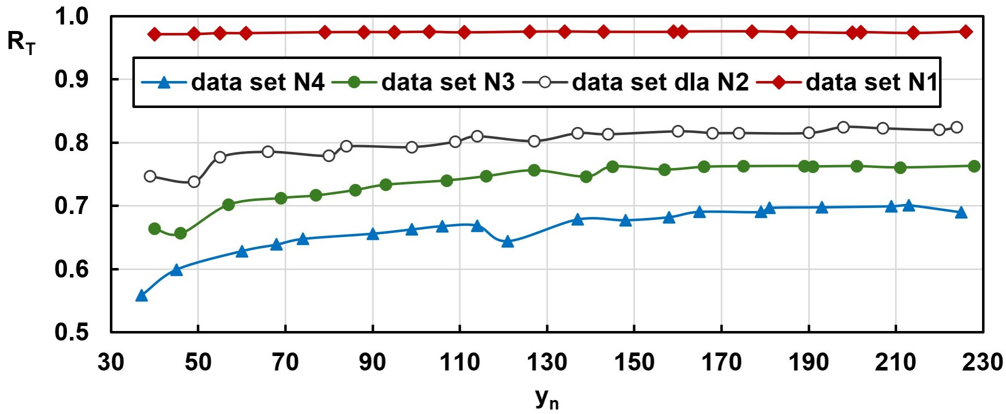


Figure 2. Values of R_T coefficient of MLP 17-y_n-5 models trained in different numbers of data ($N_1 = 8,760$, $N_2 = 17,520$, $N_3 = 26,280$, $N_4 = 35,064$).

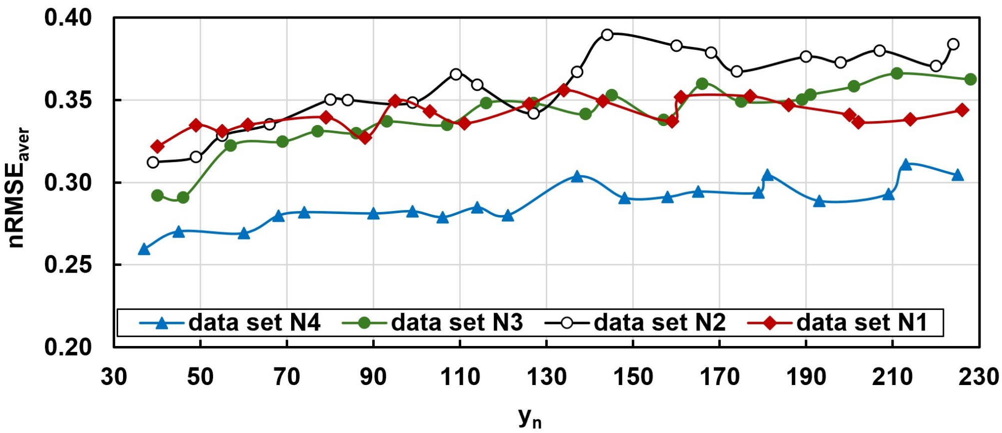


Figure 3. Values of $nRMSE_{aver}$ prediction errors of MLP 17-y_n-5 models trained on sets with different numbers of data (N_1, N_2, N_3, N_4).

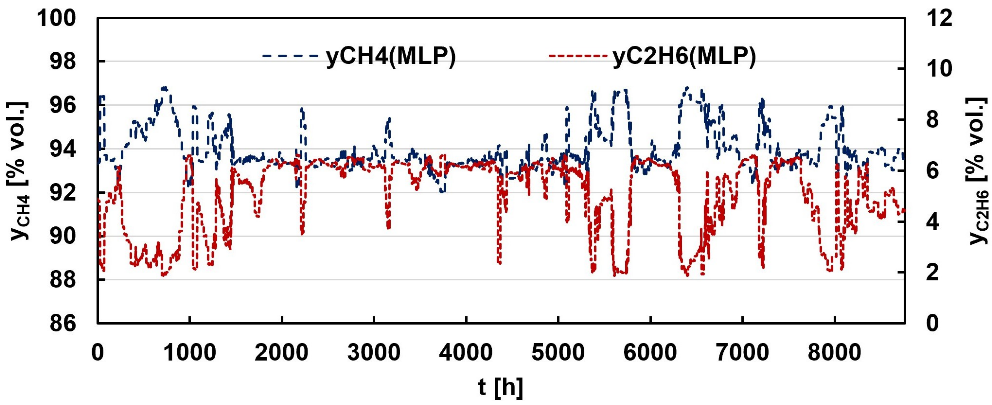


Figure 4. Variability of the predicted methane and ethane shares using the MLP 17-40-5 (N_1) model; forecast results for data set N_5.

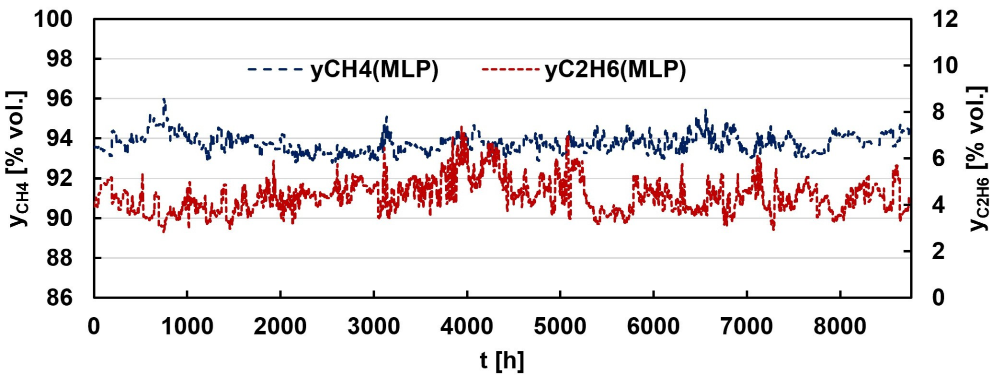


Figure 5. Variability of the predicted methane and ethane shares using the MLP 17-37-5 (N_4) model; forecast results for data set N_5.

data (an increase of methane content in natural gas causes a decrease in ethane content, and a decrease in methane content to a level of approximately 93%vol. is associated with a reduction in the carbon dioxide share to 0%).

However, increasing the data set size (N_4) for training the MLP model resulted in the relationships between the forecasted components becoming less clear and less accurate, and the forecasted values exhibiting greater variability over shorter time periods. This may be because models trained on a larger dataset recognize additional relationships or trends that were not apparent during statistical data analysis and that can complicate forecasting of natural gas components without introducing an additional input variable that significantly affects the composition of the natural gas mixture. Based on the results of the MLP model training quality, forecast errors, and the quality of the representation of relationships between the forecasted natural gas components, it was decided to train the MLP model for natural gas composition forecasting on a data of $N_1 = 8,760$.

The number of neurons in the hidden layer of the 17-y_n-5 MLP model was experimentally selected from a new set of 2,200 MLP models trained only on the $N_1 = 8,760$ dataset. The BFGS learning algorithm was applied and 200 MLP models were trained for 11 different numbers of neurons in the hidden layer of the network, narrowing the number of neurons in the hidden layer to $y_n \in \langle 35 \div 45 \rangle$. The best quality models for a given number of neurons y_n (ranging from 35 to 45) were then used to predict the natural gas composition for the new N_5 dataset.

Based on the model quality results and the natural gas composition forecast results obtained for the N_5 new data set, it was found that the MLP 17-37-5 model (logistic activation functions of neurons in the hidden and output layers) is characterized by the highest correlation coefficient $R_T = 0.9711$ and a low mean forecast error $nRMSE_{\text{aver}} = 0.321$.

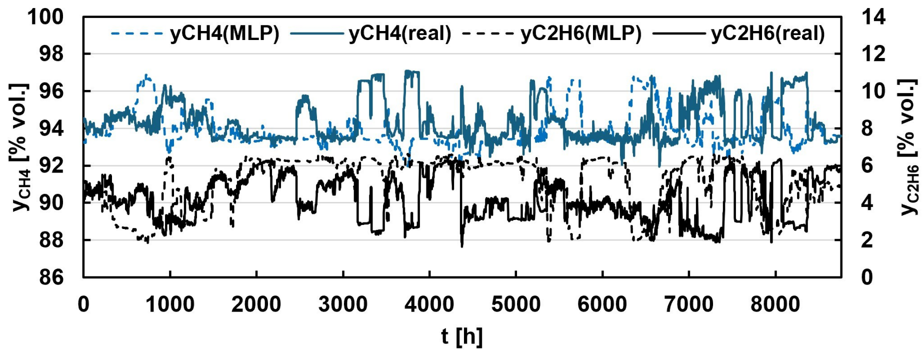
4.2. Results of forecasting the shares of natural gas components using the MLP 17-37-5 or MLP 17-37-1 models

Detailed forecast results for the five natural gas components (methane, ethane, propane, nitrogen, and carbon dioxide) obtained for a new data set ($N_5 = 8,760$), which were not used in the training process of the MLP 17-37-5 model (model schematically presented in Fig. 1), compared with actual data, are presented in Fig. 6. Analysis of these results showed that the differences between the predicted and real shares of a given natural gas component can reach up to 4%, and most often, overestimating the share of one component leads to underestimating the share of another component, and vice versa.

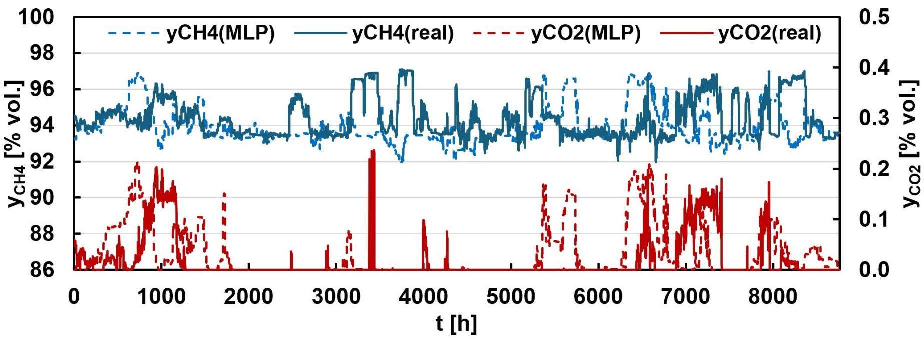
However, the relationships between the shares of the main natural gas components are well represented (e.g., a higher share of methane reduces the share of ethane). Based on the forecast results and data on the actual shares of the five selected natural gas components, the nRMSE forecast error values for each component and the average forecast error made by the MLP 17-37-5 model were determined (results in Fig. 9). Taking into account the size of the average error nRMSE, it was found that a model with such a structure can be successfully used to simultaneously forecast five natural gas components.

To verify whether it is possible to obtain gas component forecasts with an average $nRMSE_{\text{aver}}$ error of less than 0.3210 (for the MLP 17-37-5 model), five MLP 17-37-1 models were trained to forecast each natural gas component individually. Similarly to the training of the MLP 17-37-5 model, logistic activation functions of the hidden and output layer neurons and the BFGS learning algorithm were used. The diagrams of the five MLP 17-37-1 models are presented in Fig. 7. The MLP 17-37-1 models were trained using a data set $N_1 = 8,760$ (the same set used to train the MLP 17-37-5 model).

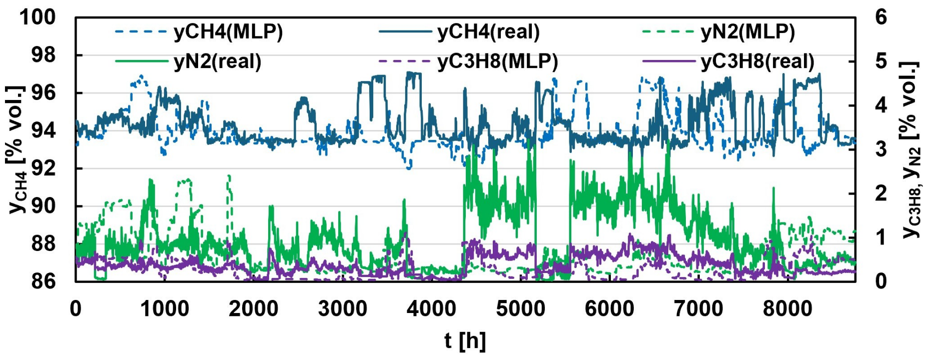
The results of individual forecasts of the shares of five natural gas components made using the MLP 17-37-1 models for the new $N_5 = 8,760$ data set and the actual share data of these components in natural gas are presented in Fig. 8. Also in this case, regardless of the forecasted gas component, the forecast results are observed to be



(a)



(b)



(c)

Figure 6. Variability of the actual and predicted shares of the five natural gas components using the MLP 17-37-5 model trained on a data set of $N_1 = 8,760$; a) methane and ethane, b) methane and carbon dioxide, c) methane, propane, and nitrogen; forecast results for data set $N_5 = 8,760$.

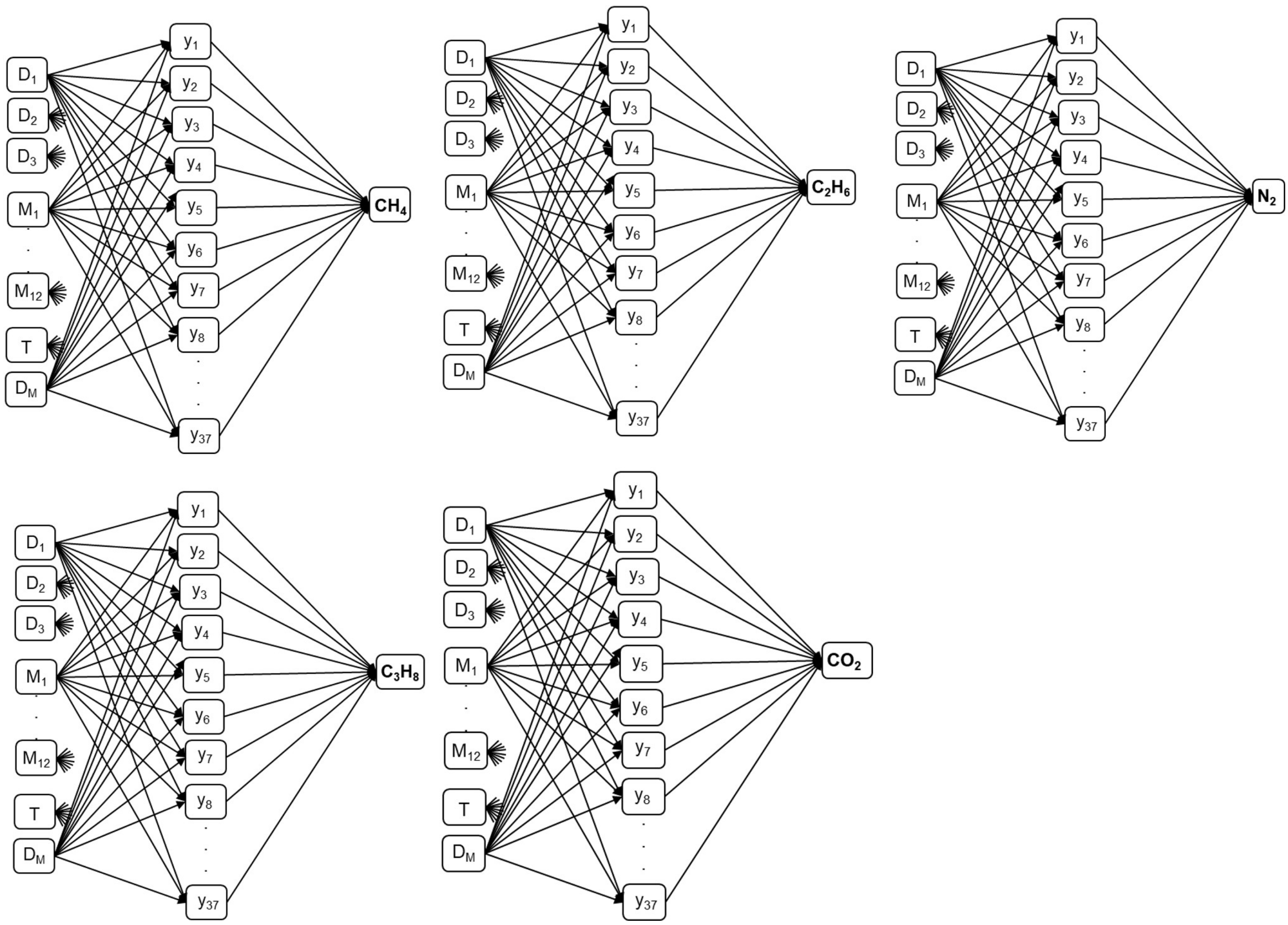


Figure 7. Schematic diagram of the MLP 17-37-1 for predicting the content of individual natural gas components: CH_4, C_2H_6, C_3H_8, N_2, CO_2.

higher or lower than the actual data by approximately 4%. Comparing the results presented in Fig. 6 and Fig. 8, it was observed that for a given natural gas component, better quality forecasts can be obtained using the MLP 17-37-5 model compared to the MLP 17-37-1 model.

Table 6 summarizes the correlation coefficients of the MLP 17-37-1 models and the MLP 17-37-5 model trained on the $N_1 = 8,760$ data set. The training results of the MLP 17-37-1 models indicate that the R_T correlation coefficient values of the models forecasting each natural gas component individually differ slightly, and the average R_T value for all MLP 17-37-1 models is only 0.37% higher compared to the R_T value obtained for the 17-37-5 model. However, based on the results of the real and forecasted (for the $N_5 = 8,760$ data set) shares of the five natural gas components, the nRMSE forecast errors were determined for each component and the averaged value for each MLP model.

Detailed results of the calculations of the average nRMSE forecast errors for each component forecasted by the MLP 17-37-5 model or the five MLP 17-37-1 models are presented in Fig. 9. Considering the results presented in Fig. 9, it can be concluded

Table 6. Parameters R_T and nRMSE of MLP 17-37-1 or MLP 17-37-5 models.

Model	Component	R_T	Algorithm	$nRMSE_{\text{aver}}$
	CH_4	0.9663	BFGS 671	0.3383
MLP 17-37-1	C_2H_6	0.9777	BFGS 750	
	C_3H_8	0.9838	BFGS 588	
	N_2	0.9805	BFGS 764	
	CO_2	0.9651	BFGS 430	
MLP 17-37-5	CH_4, C_2H_6, C_3H_8, N_2, CO_2	0.9711	BFGS 820	0.3210

that the better quality MLP model for forecasting of the five natural gas components shares was trained using five neurons in the output layer (lower forecast error $nRMSE_{\text{aver}}$). The MLP 17-37-5 model, although characterized by a slightly lower correlation coefficient compared to the mean R_T obtained for the MLP 17-37-1 models, allows for the preparation of better quality forecasts of the shares of the natural gas components.

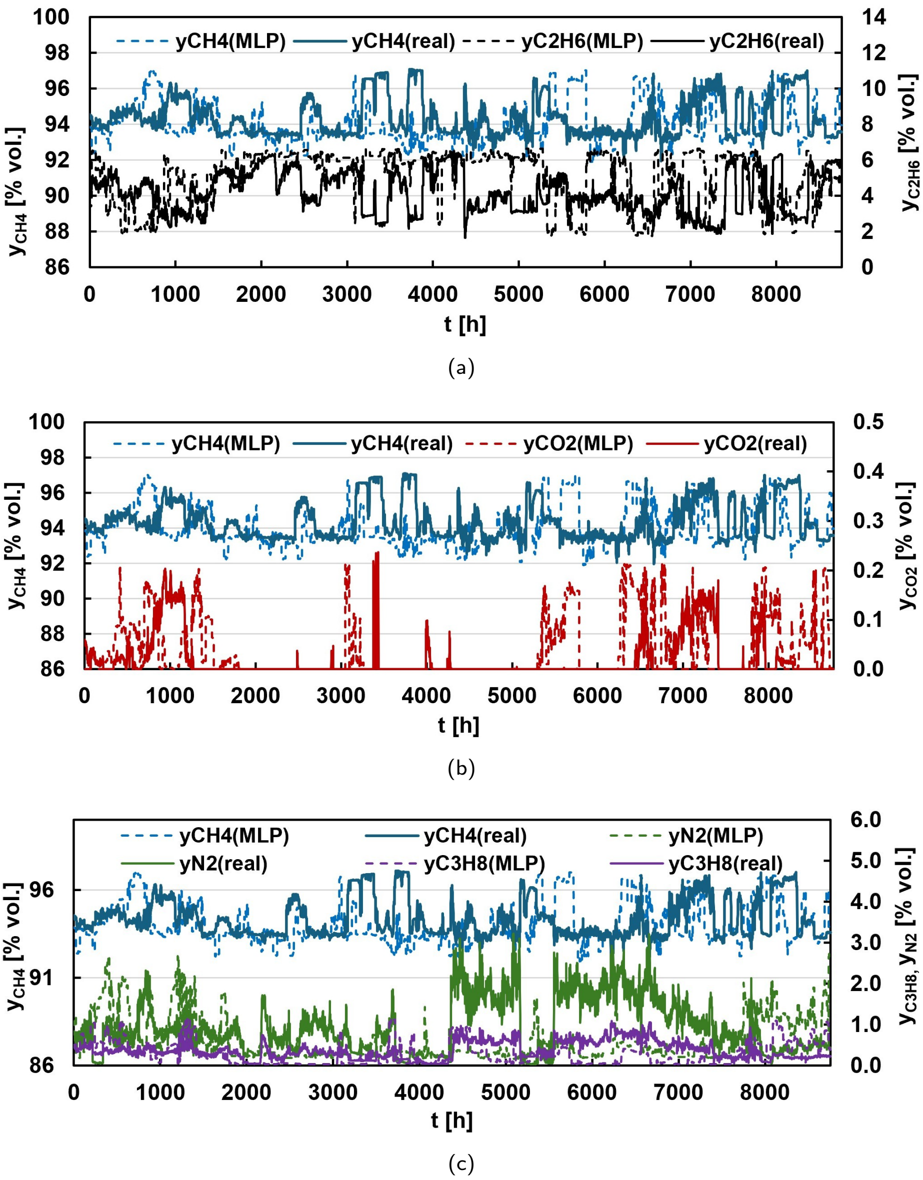


Figure 8. Variability of the actual and predicted shares of the five natural gas components using the MLP 17-37-1 models trained on a data set of $N_1 = 8,760$; a) methane and ethane, b) methane and carbon dioxide, c) methane, propane, and nitrogen; forecast results for data set $N_5 = 8,760$.

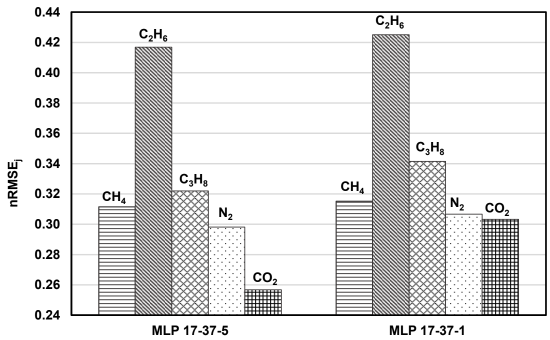


Figure 9. Forecast errors $nRMSE_j$ and $nRMSE_{\text{aver}}$ obtained with the MLP 17-37-1 and MLP 17-37-5 models; results for dataset $N_5 = 8,760$.

5. CONCLUSIONS

Based on the data set $N_1 = 8,760$ or $N_4 = 35,064$, the MLP 17-37-5 model was trained to forecast five natural gas components (methane, ethane, propane, nitrogen, and carbon dioxide) with an average prediction error nRMSE = 0.321. The input data for the MLP model were selected based on the results of statistical analysis of the correlations between input variables and between input variables and forecasted components. This allowed to reduce the number of input variables of the forecasting model to 17 neurons.

Compared to the results of the research presented in Szoplik and Muchel (2023a), the number of neurons in the model's input layer was reduced from 18 to 17, which also resulted in a reduction of the number of neurons in the hidden layer from 65 to 37. However, such model simplification allowed to maintain good quality forecasts of the five natural gas component shares.

Based on the results of MLP forecasting models with datasets of different sizes ($N_1 = 8,760$ or $N_4 = 35,064$), it was also demonstrated that other unusual factors can indirectly influence changes in gas composition. One such unusual factor was the presence of the COVID-19 pandemic, which indirectly influenced the change in gas flow direction in some gas network pipelines due to changes in the load on the gas pipelines. In this case, an additional factor describing this unusual situation should be included among the MLP model input data.

The proposed MLP 17-37-5 model allows to forecast natural gas composition over a one-year period with a weighted average error of MAPE $\approx 3\%$, while the mean relative error of the heat of combustion calculated based on the predicted gas composition and the heat of combustion calculated based on the actual natural gas composition is 1.57%. Furthermore, no better quality forecasts of natural gas component shares were found by using five MLP 17-37-1 models to individually forecast the share of each component.

ACKNOWLEDGMENTS

The calculations presented in the paper were made on the basis of real data provided by the GAZ-SYSTEM — Gas Transmission Operator of the Polish natural gas network. The authors would like to thank the company for cooperation.

REFERENCES

Al-Quraishi A.A., Shokir E.M., 2011. Artificial neural networks modeling for hydrocarbon gas viscosity and density estimation. *J. King Saud Univ. Eng. Sci.*, 23, 123–129. DOI: 10.1016/j.jksues.2011.03.004.

Alves Jr. O., Fontes C., 2022. Modeling and optimization of natural gas distribution networks for new supplier projects. *Energy Convers. Manage., X*, 15, 100240. DOI: 10.1016/j.ecmx.2022.100240.

Bermúdez A., Shabani M., 2022. Numerical simulation of gas composition tracking in a gas transportation network. *Energy*, 247, 123459. DOI: 10.1016/j.energy.2022.123459.

Bianco V., Scarpa F., Tagliafico L.A., 2014. Scenario analysis of nonresidential natural gas consumption in Italy. *Appl. Energy*, 113, 392–403. DOI: 10.1016/j.apenergy.2013.07.054.

Chaczykowski M., Sund F., Zarodkiewicz P., Hope S.M., 2018. Gas composition tracking in transient pipeline flow. *J. Nat. Gas Sci. Eng.*, 55, 321–330. DOI: 10.1016/j.jngse.2018.03.014.

Chaczykowski M., Zarodkiewicz P., 2017. Simulation of natural gas quality distribution for pipeline systems. *Energy*, 134, 681–698. DOI: 10.1016/j.energy.2017.06.020.

Chu J., Liu X., Zhang Z., Zhang Y., He M., 2021. A novel method overcomeing overfitting of an artificial neural network for accurate prediction: application on thermophysical property of natural gas. *Case Stud. Therm. Eng.*, 28, 101406. DOI: 10.1016/j.csite.2021.101406.

Durmayaz A., Kadioğlu M., Şen Z., 2000. An application of the degree-hours method to estimate the residential heating energy requirement and fuel consumption in Istanbul. *Energy*, 25, 1245–1256. DOI: 10.1016/S0360-5442(00)00040-2.

Fan D., Gong J., Zhang S., Shi G., Kang Q., Xiao Y., Wu C., 2021. A transient composition tracking method for natural gas pipe networks. *Energy*, 215, 119131. DOI: 10.1016/j.energy.2020.119131.

Farzaneh-Gord M., Mohseni-Gharyehsafa B., Ebrahimi-Moghadam A., Jabari-Moghadam A., Toikka A., Zvereva I., 2018. Precise calculation of natural gas sound speed using neural networks: an application in flow meter calibration. *Flow Meas. Instrum.*, 64, 90–103. DOI: 10.1016/j.flowmeasinst.2018.10.013.

Geem Z.W., Roper W.E., 2009. Energy demand estimation of South Korea using artificial neural network. *Energy Policy*, 37, 4049–4054. DOI: 10.1016/j.enpol.2009.04.049.

George J., Arun P., Muraleedharan C., 2018. Assessment of producer gas composition in air gasification of biomass using artificial neural network model. *Int. J. Hydrogen Energy*, 43, 9558–9568. DOI: 10.1016/j.ijhydene.2018.04.007.

Hafezi R., Akhavan A.N., Zamani M., Pakseresht S., Shamshirband S., 2019. Developing a data mining based model to extract predictor factors in energy systems: application of global natural gas demand. *Energies*, 12, 4124. DOI: 10.3390/en12214124.

Kenisarin M., Kenisarina K., 2007. Energy saving potential in the residential sector of Uzbekistan. *Energy*, 32, 1319–1325. DOI: 10.1016/j.energy.2006.09.014.

Kumor M., Porada S., 2021. Polish LNG terminal influence on natural gas quality available in the Polish gas transmission network. *IOP Conf. Ser.: Earth Environ. Sci.*, 642, 012003. DOI: 10.1088/1755-1315/642/1/012003.

Lay E.N., Peymani M., Sanjari E., 2012. Prediction of natural gas viscosity using artificial neural network approach. *Int. J. Chem. Mol. Eng.*, 6, 577–583. DOI: 10.5281/zenodo.1333845.

Merkel G.D., Povinelli R.J., Brown R.H., 2018. Short-term load forecasting of natural gas with deep neural network regression. *Energies*, 11, 2008. DOI: 10.3390/en11082008.

Muchel P., 2024. *Modeling the variability of the composition of natural gas using the Artificial Neural Networks Method*. PhD Thesis. West Pomeranian University of Technology in Szczecin, Poland.

Osiadacz A.J., Chaczykowski M., 2020. Modeling and simulation of gas distribution networks in a multienergy system environment. Proc. IEEE, 108, 1580–1595. DOI: 10.1109/JPROC.2020.2989114.

Parikh J., Purohit P., Maitra P., 2007. Demand projections of petroleum products and natural gas in India. *Energy*, 32, 1825–1837. DOI: 10.1016/j.energy.2007.02.008.

Park J., Cho J., Choi H., Park J., 2020. Prediction of reformed gas composition for diesel engines with a reformed EGR system using an artificial neural network. *Energies*, 13, 5886. DOI: 10.3390/en13225886.

Postawa K., Fałtynowicz H., Pstrowska K., Szczygieł J., Kułażyński M., 2022. Artificial neural networks to differentiate the composition and pyrolysis kinetics of fresh and long-stored maize. *Bioresour. Technol.*, 364, 128137. DOI: 10.1016/j.biortech.2022.128137.

Saedi I., Mhanna S., Mancarella P., 2021. Integrated electricity and gas system modelling with hydrogen injections and gas composition tracking. *Appl. Energy*, 303, 117598. DOI: 10.1016/j.apenergy.2021.117598.

Sarak H., Satman A., 2003. The degree-day method to estimate the residential heating natural gas consumption in Turkey: a case study. *Energy*, 28, 929–939. DOI: 10.1016/S0360-5442(03)00035-5.

Satman A., Yalcinkaya N., 1999. Heating and cooling degree-hours for Turkey. *Energy*, 24, 833–840. DOI: 10.1016/S0360-5442(99)00035-3.

Singh V., Gupta I., Gupta H.O., 2007. ANN-based estimator for distillation using Levenberg–Marquardt approach. *Eng. Appl. Artif. Intell.*, 20, 249–259. DOI: 10.1016/j.engappai.2006.06.017.

Szoplik J., 2015. Forecasting of natural gas consumption with artificial neural networks. *Energy*, 85, 208–220. DOI: 10.1016/j.energy.2015.03.084.

Szoplik J., 2016. Improving the natural gas transporting based on the steady state simulation results. *Energy*, 109, 105–116. DOI: 10.1016/j.energy.2016.04.104.

Szoplik J., 2017. Changes in gas flow in the pipeline depending on the network foundation in the area. *J. Nat. Gas Sci. Eng.*, 43, 1–12. DOI: 10.1016/j.jngse.2017.03.020.

Szoplik J., Muchel P., 2023a. Using an artificial neural network model for natural gas heat combustion forecasting. *Chem. Process Eng.*, 44, e19. DOI: 10.24425/cpe.2023.146721.

Szoplik J., Muchel P., 2023b. Using an artificial neural network model for natural gas compositions forecasting. *Energy*, 263, 126001. DOI: 10.1016/j.energy.2022.126001.

Wang B., Guo Y., Wang D., Zhang Y., He R., Chen J., 2022. Prediction model of natural gas pipeline crack evolution based on optimized DCNN-LSTM. *Mech. Syst. Signal Process.*, 181, 109557. DOI: 10.1016/j.ymssp.2022.109557.

Wen K., He L., Liu J., Gong J., 2019. An optimization of artificial neural network modeling methodology for the reliability assessment of corroding natural gas pipelines. *J. Loss Prev. Process Ind.*, 60, 1–8. DOI: 10.1016/j.jlp.2019.03.010.

Wood D.A., Choubineh A., 2020. Transparent machine learning provides insightful estimates of natural gas density based on pressure, temperature and compositional variables. *J. Nat. Gas Geosci.*, 5, 33–43. DOI: 10.1016/j.jnggs.2019.12.003.

Zhang Z., Saedi I., Mhanna S., Wu K., Mancarella P., 2022. Modelling of gas network transient flows with multiple hydrogen injections and gas composition tracking. *Int. J. Hydrogen Energy*, 47, 2220–2233. DOI: 10.1016/j.ijhydene.2021.10.165.