Preprints.org

Article

Robust Gas Demand Prediction Using Deep Neural Networks: A Data-Driven Approach to Forecasting Under Regulatory Constraints

Kostiantyn Pavlov [*] , Olena Pavlova , Tomasz Wołowiec , Svitlana Slobodian , Andriy Tymchyshak , Tetiana Vlasenko

Article

Robust Gas Demand Prediction Using Deep Neural Networks: A Data-Driven Approach to Forecasting Under Regulatory Constraints

**Kostiantyn Pavlov [1], Olena Pavlova [1,2], Tomasz Wołowiec [3], Svitlana Slobodian [4],
Andriy Tymchyshak [1] and Tetiana Vlasenko [5,6]**

[1] Faculty of Economics and Management, Lesya Ukrainka Volyn National University, Voli Ave, 13, 43-025 Lutsk, Volyn Oblast, Ukraine

[2] Faculty of Management, AGH University of Krakow, A. Mickiewicza Ave. 30, 30-059 Kraków, Poland

[3] Vice-Rector for Science and International Cooperation of the Lublin Academy of WSEI, Projektowa 4 20-209 Lublin, Poland

[4] Faculty of Mathematical and Computer Science, Vasyl Stefanyk Precarpathian National University, 57 Shevchenka str., 76018 Ivano-Frankivsk, Ukraine

[5] Department of Management, *Academy of Silesia,* ul. Rolna 43, Katowice, Poland

[6] Organization Department, Simon Kuznets Kharkiv National University of Economics, Nauky Ave., 9-A, 61166, Kharkiv, Ukraine

* Correspondence: pavlovkostya@gmail.com; Tel.: +380 97 283 08 08

Abstract: Accurate gas consumption forecasting is critical for modern energy systems due to complex consumer behaviour and regulatory requirements. Deep neural networks (DNNs), such as Seq2Seq with attention, TiDE, and Temporal Fusion Transformers, are promising for modelling complex temporal relationships and nonlinear dependencies. This study compares state-of-the-art architectures using real-world data from over 100,000 consumers to determine their practical viability for forecasting gas consumption under operational and regulatory conditions. Particular attention is paid to the impact of data quality, feature attribution, and model reliability on performance. The main use cases for natural gas consumption forecasting are tariff setting by regulators and system balancing for suppliers and operators. The study used monthly natural gas consumption data from 105,527 households in the Volyn region of Ukraine from January 2019 to April 2023 and meteorological data on average monthly air temperature. Missing values were replaced with zeros or imputed using seasonal imputation and K-nearest neighbours. The results showed that previous consumption is the dominant feature for all models, confirming their autoregressive origin and the high importance of historical data. Temperature and category were identified as supporting features. Improvised data consistently improved the performance of all models. Seq2SeqPlus showed high accuracy, TiDE was the most stable, and TFT offered flexibility and interpretability. Implementing these models requires careful integration with data management, regulatory frameworks, and operational workflows.

Keywords: natural gas consumption and demand forecasting; neural networks; Seq2Seq; TiDE; temporary fusion transformers; regulatory restrictions; energy systems; tariff policy; capacity balancing and reservation; machine learning; long short-term memory; time series analysis; SHAP processing of missing data

1. Introduction

Accurate and robust forecasting of gas consumption has become increasingly critical in the context of modern energy systems, where supply reliability, market efficiency, and environmental responsibility intersect.

With the growing complexity of consumption behaviour, influenced by factors such as urbanization, climate variability, and diversified consumer categories, traditional forecasting approaches often fall short in capturing the underlying dynamics of energy demand.

In response to these challenges, deep neural network (DNN) models have emerged as a promising alternative, offering the capacity to model complex temporal relationships and nonlinear dependencies. Architectures such as Seq2Seq with attention, TiDE, and Temporal Fusion Transformers (TFT) enable learning from rich, multi-dimensional datasets that include historical consumption, external variables like temperature, and categorical consumer attributes.

The potential of DNNs extends beyond technical innovation; it holds tangible value in meeting the regulatory and operational demands of the gas sector. Regulators increasingly require energy providers to deliver accurate, transparent, and timely forecasts that support network stability, market transparency, and environmental compliance.

From forecasting aggregated demand for capacity planning to ensuring adherence to balancing responsibilities and emission targets, DNN-based models can provide both the precision and interpretability needed for modern regulatory frameworks.

This study evaluates and compares several state-of-the-art DNN architectures using a large-scale, real-world dataset of over 100,000 consumers, aiming to determine their practical viability for gas consumption forecasting in operational and regulatory settings.

Focus is placed on how data quality, feature attribution and model robustness influence performance—providing actionable insights for real-world deployment.

The necessity of short- and long-term forecasting of natural gas consumption is of crucial importance in the current conditions of Ukraine's natural gas market as an integral part of the European Energy market. In its context accurate forecasting ensures effective interaction among market participants and helps reduce operational and transaction costs as well as mitigates the risks of negative regulatory effects [1].

In particular, the predicted consumption and distribution volumes of natural gas are used by the state regulatory authority—The National Energy and Utilities Regulatory Commission when setting tariffs for gas distribution system operators and the gas transmission system operator (reference to NERC resolution).

Currently, existing methodologies involve using historical data for tariff calculations, with periodic adjustments based on actual data from the previous periods (mostly a calendar year). This approach creates discriminatory effects on consumers and leads to cash flow gaps for DSO.

Additionally, forecasting is essential for natural gas suppliers and traders to balance and reserve capacities (reference to the GTS Codes). Furthermore, inaccurate forecasting increases suppliers' costs due to penalties and the need to purchase additional gas volumes on the spot market [2].

Recent advancements in gas consumption forecasting leverage a variety of machine learning (ML) techniques and hybrid models to enhance predictive accuracy. These approaches utilize historical consumption data, meteorological factors, and even social variables to create robust forecasting models. The following sections outline the key methodologies currently employed in the field.

Machine Learning Techniques:

Deep Learning Models: Techniques such as Long Short-Term Memory (LSTM) and Deep Neural Networks (DNN) have shown significant promise. For instance, a DNN incorporating social factors outperformed traditional models in forecasting natural gas consumption in Greece [3].

Hybrid Models: Combining statistical methods with ML, such as Facebook's Prophet and Holt-Winters Method, has improved accuracy in predicting natural gas demand [4].

Change Point Detection

Dynamic Adaptation: The integration of change point detection mechanisms allows models to adapt to shifts in consumption patterns, enhancing forecasting reliability in real-time scenarios [5].

Comparative Studies:

Model Evaluation: A comprehensive review of various forecasting strategies highlights the effectiveness of hybrid models and the importance of data decomposition methods in improving prediction accuracy [6].

While these advanced techniques significantly enhance forecasting capabilities, challenges remain, particularly in data privacy and the interpretability of complex models. Addressing these issues will be crucial for the future development of gas consumption forecasting methodologies.

2. Actual Regulatory Use Cases of Consumption Forecasting

There are several major use cases concerning a proper forecasting of natural gas consumption engaging all the participants of the gas market—consumers, system operators, suppliers and state regulatory authority. Let's consider them deeply. The first one is Forecasting for Tariff Setting. The National Energy and Utilities Regulatory Commission (hereinafter – NEURC) uses gas consumption forecasts to determine fair and economically justified tariffs for gas distribution and transportation.

Tariffs for gas distribution system operators (DSOs) and the gas transmission system operator (TSO) are set for a period of 1 to 5 years based on predicted gas consumption and transmission capacity volumes. In this case precious long-term forecasts of gas consumption help estimate expected revenues and costs for network operators, ensuring that tariffs cover operational expenses while maintaining affordability for consumers [7,8].

But there are certain challenges with the current approach, such as historical data dependency, delayed adjustments and consumer inequality. The current methodology relies on past consumption data to determine tariffs, with periodic adjustments —on actual figures. This can be problematic in times of fluctuating demand.

If actual consumption deviates from forecasts, tariff corrections may take a full calendar year, leading to financial imbalances for operators. In addition, inaccurate forecasting can result in some consumer groups overpaying or underpaying, creating a discriminatory effect on different types of consumers.

The second use case considers Forecasting for Balancing and Capacity Reservation due to the The Gas Transmission System Code requirements [9]. Suppliers and network operators use day- and month-ahead gas consumption forecasts to ensure the gas system remains balanced and to optimize capacity reservation. Natural gas supply must match demand in real-time to maintain stable system pressure and prevent shortages or excess gas buildup, meanwhile suppliers must forecast demand accurately to avoid imbalances, which could lead to financial penalties.

While Capacity Reservation is maintained, gas suppliers must reserve transmission and distribution capacities in advance based on their expected demand. Incorrect forecasts may result in overbooking (leading to unnecessary costs) or under booking (causing supply shortages and urgent, expensive spot market purchases).

Financial and operational risks may occur in case of poor forecasting, such as penalty costs, reluctant spot market purchases or liquidity issues. If suppliers do not accurately predict consumption, they may face penalties for imbalances. Unexpected demand surges require suppliers to purchase additional gas on the spot market, which often has significantly higher prices than long-term contracts. Inaccurate forecasts can create cash flow gaps, where suppliers either have excess gas, they cannot sell or face sudden high costs to meet unexpected demand.

Therefore, accurate gas consumption forecasting is critical for both regulatory tariff-setting and market-based balancing operations. For regulatory authorities and consumers, it ensures stable, fair tariffs that reflect actual market conditions. For suppliers, it minimizes costs, prevents penalties, and improves financial planning. Now let's review up-to-date forecasting solutions applying for the subject of our research.

Our current study is pursuing to find the best extant solutions to satisfy both short- and long-term forecasting demands. The key point of the possible approaches we are to consider is data, its availability, robustness and relevance. The volumes of data available to gas market participants and

the state regulatory authority are enormous and constantly growing, although only a small part of them can be used for forecasting purposes (Figure 1).
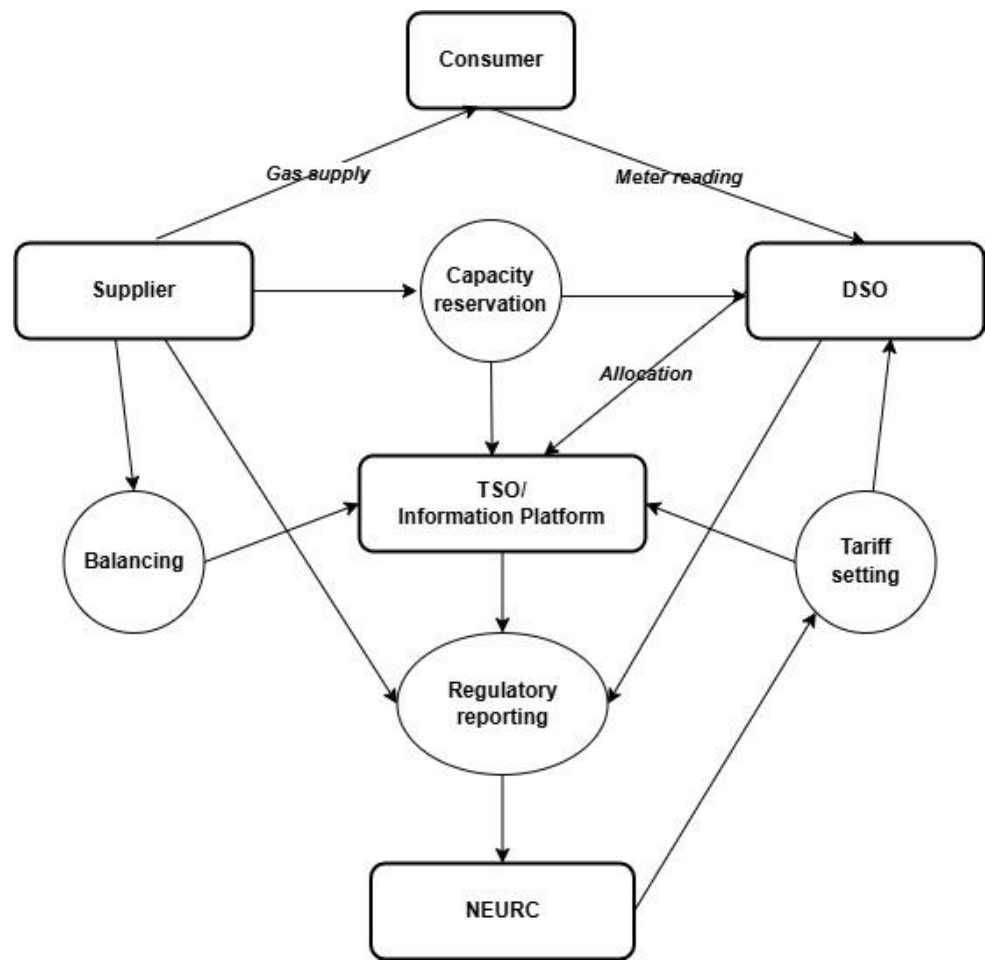

Figure 1. Gas market data flow scheme.

The TSOUA, as the authorized operator of the Information Platform, possesses large amounts of data on all the consumers of Ukraine—in particular, their territorial location, consumption volumes, affiliation with DSO and gas suppliers [10].

In turn, DSO and suppliers have more discrete data, but only regarding the consumers to whom they provide their services. At the same time, regulatory factors, such as the requirements of regulatory legal acts (for example, the methodology for calculating the volume of gas consumed in the absence of meter readings or existing technological limitations (for example, the low level of coverage of household consumers with remote reading meters and the inability of DSO controllers to cover all consumers) cause the incompleteness of this data [11].

The NERC consolidates these data through regulatory reporting submitted by market participants, forming a holistic picture of the regulated market, while making the information more general and publicly available. Meanwhile current legislative restrictions, regarding the protection of personal data and commercial concerns, may further limit their use for forecasting purposes, which is to be mitigated by amending existing regulations. That is, the corresponding models must consider this incompleteness of the data, or it is necessary to additionally apply specified data wrangling and preprocessing solutions, which can lead to a significant increase in time and computational power.

The structure of consumer types and their share in the total consumption volume differs significantly in certain regions of Ukraine—commercial consumers prevail in the East and Center, and household consumers in the West and South.

This may be another significant factor that may influence the choice of forecasting approaches. Commercial consumers demonstrate more stable patterns of consumer behavior with low seasonality, while the patterns of household and utility consumers, on the contrary, have a pronounced seasonality and dependence on environmental temperature. Moreover, within households, separate groups whose consumption patterns are extremely different can be distinguished. This depends primarily on the purpose of consumption (home heating, water heating, cooking), volumes of consumption and housing floor area.

Under the conditions of state regulation of prices and tariffs for household consumers, which is currently applied in Ukraine, the natural gas cost factor has a negligible statistical impact on their consumption level, and therefore we will not use it in this study [12].

Summarizing the above, we can outline the framework of our requirements for predictive approaches—the ability to process large volumes of raw incomplete data, detect and consider seasonal trends, as well as accept additional features—environmental temperature, category, etc.

Moreover, we need a single predictive model that considers all of more than 100 000 consumers' unique time series simultaneously attempting to capture the core patterns that govern the series, thereby mitigating the potential noise that each series might introduce. At the output, the solutions should provide short- and long-term forecasts of consumption volumes both in terms of distinct consumers, consumer categories and in general for certain regions. Since the task is quite complex, in our research we will narrow the scope and focus on the most essential part of it—household consumption forecasting.

3. Materials and Methods

Currently, time series forecasting problems can be solved using a wide range of approaches. Among them, we can mention classical statistical approaches—ARIMA, which is effective for short-term forecasting when consumption data follows a trend and seasonality and SARIMA—an extension of ARIMA that accounts for seasonal patterns in gas demand. Exponential Smoothing (Holt-Winters method) which is useful for capturing trends and seasonal variations., regression models—Multiple Linear Regression (MLR) that relates gas consumption to external variables like temperature, industrial activity, and population, as well as Generalized Additive Models (GAMs) which is more flexible than linear regression, allowing non-linear relationships between predictors—can be added to the list.

These methods have been widely used due to their interpretability and relatively low computational requirements and deliver an acceptable level of precision. A significant limitation of these methods is the inability to train global models that allow taking into account the behavior of many consumers simultaneously, and therefore they are unsuitable for achieving our goals.

In the last decade, due to the impressive development of computing power, completely new approaches have appeared, which are based on machine learning and artificial intelligence.

Support Vector Machines (SVMs) e.g., are useful for medium-term forecasting, especially when relationships between variables are complex, Random Forest and Gradient Boosting Trees (XGBoost, LightGBM) handle non-linear relationships well and can incorporate many external variables, while Feedforward Neural Networks (FNNs) are effective for learning complex patterns but require significant data and Long Short-Term Memory Networks (LSTMs) and Gated Recurrent Units (GRUs) are ideal for sequential and time-series forecasting. But the most cutting edge solutions are based on the concept of Deep Neural Network [13–15].

These methods require huge amounts of data and immense computational facilities, but in return a higher level of prediction accuracy can be granted.

Deep neural network methods in gas consumption forecasting. So let's get closer to the most up-to-date realm of time series forecasting approaches—deep neural network models, in particular Time-series Dense Encoder, Temporal Fusion Transformer and Sequence to Sequence plus Attention.

The evolution of deep learning models for time series forecasting reflects a gradual shift from sequence modeling roots in natural language processing to highly specialized architectures tailored for temporal data. Sequence-to-Sequence (Seq2Seq) models were initially developed for machine translation and later adapted to time series forecasting. By incorporating the attention mechanism, these models overcame limitations of fixed-length context vectors, allowing them to selectively focus on relevant time steps—an innovation that improved their ability to handle long-term dependencies.

Building on these foundations, the Temporal Fusion Transformer (TFT) was introduced as a purpose-built architecture for interpretable multivariate time series forecasting. It fully integrates attention mechanisms—not only temporally, but also across input features—combined with gating and variable selection layers to enhance both performance and interpretability. TFT represents a significant leap in leveraging attention for structured time-series tasks.

In contrast, TiDE (Time-series Dense Encoder) emerged more recently as a minimalist, fully feedforward alternative to attention-based models. Eschewing traditional sequence-to-sequence and attention architectures, TiDE employs learned temporal embeddings and multi-layer perceptrons (MLPs) to capture time-dependent patterns.

It is designed to be computationally efficient and particularly well-suited for long-horizon forecasting, showing that competitive accuracy can be achieved without explicit recurrence or attention.

Time-series Dense Encoder (TiDE) [16].

Deep learning model TiDE is designed to address the limitations of both linear models and Transformer-based approaches in long-term time-series forecasting [17].

While recent work demonstrated that simple linear models can outperform Transformers on certain benchmarks, linear methods fail to capture nonlinear dependencies or leverage covariates effectively. TiDE bridges this gap by introducing a dense Multi-Layer Perceptron (MLP)-based encoder-decoder framework that handles non-linear patterns while maintaining scalability. TiDE encodes the past of a time-series along with covariates using dense MLP's and then decodes the encoded time-series along with future covariates.

Key Innovations. TiDE employs channel-independent processing, where each i-th time-series is modeled separately using its past observations $y_{1:L}^{(i)}$, dynamic covariates $x_{1:L+H}^{(i)}$, and static attributes $s^{(i)}$, while sharing global weights across the dataset, i.e.,

$$f: \left(\left\{y_{1:L}^{(i)}\right\}_{i \in I}, \left\{x_{1:L+H}^{(i)}\right\}_{i \in I}, \left\{s^{(i)}\right\}_{i \in I}\right) \rightarrow \left\{\hat{y}_{L+1:L+H}^{(i)}\right\}_{i \in I}$$

where $i \in I$, I denotes the set of unique entities in a given time-series dataset, f is a forecasting model.

The architecture relies on residual MLP blocks, which consist of a ReLU-activated hidden layer, a linear skip connection, and dropout with layer normalization for stable training.

The model operates in two phases: an encoding phase that compresses historical data and covariates into a low-dimensional latent representation, and a decoding phase that maps this representation to future predictions using projected future covariates. This design ensures efficient non-linear modeling while maintaining interpretability and scalability.

TiDE Architecture. Encoding Stage. Feature Projection. Dynamic covariates $x_t^{(i)}$ of time-series i at time t are compressed via a residual block to reduced dimension $\tilde{r} \ll r$:

$$\tilde{x}_t^{(i)} = Residual\ Block\left(x_t^{(i)}\right), \tilde{x}_t^{(i)} \in R^{\tilde{r}}.$$

This avoids the high dimensionality of flattened raw covariates.

Dense Encoder. The encoder stacks projected covariates, static attributes, and past observations, processing them through n_e residual blocks:

$$e^{(i)} = Encoder\left(y_{1:L}^{(i)}; \tilde{x}_{1:L+H}^{(i)}; s^{(i)}\right).$$

Decoding Stage. Dense Decoder. The latent representation $e^{(i)}$ is transformed through n_d residual blocks into a tensor $D^{(i)} \in R^{p \times H}$, where each column $d_t^{(i)}$ corresponds to a future time step:

$$D^{(i)} = Reshape\left(Decoder\big(e^{(i)}\big)\right).$$

Temporal Decoder. For each horizon step t, a residual block combines $d_t^{(i)}$ with projected future covariates $\tilde{x}_{L+t}^{(i)}$:

$$\hat{y}_{L+t}^{(i)} = Temporal\ Decoder\big(d_t^{(i)};\ \tilde{x}_{L+t}^{(i)}\big), t \in \{1, 2, \dots, H\}.$$

This "highway" connection ensures direct covariate influence.

Global Residual Connection.

A linear projection of the look-back window $y_{1:L}^{(i)}$ is added to the predictions, ensuring the model subsumes linear baselines.

Training and Evaluation. TiDE is trained using mini-batch gradient descent with root mean square error (RMSE) loss, where each batch contains multiple time-series segments consisting of look-back windows and their corresponding forecast horizons, allowing overlapping training sequences for comprehensive learning. The model's performance is evaluated through rolling-window validation, testing all possible consecutive (look-back, horizon) pairs in the test set to thoroughly assess forecasting accuracy. This evaluation approach, consistent with established time-series forecasting practices, can also be applied to a validation set for hyperparameter optimization and model selection.

Temporal Fusion Transformer (TFT) [18]. The Temporal Fusion Transformer (TFT) is designed with specialized components to handle diverse input types (static, known, and observed) for robust time-series forecasting. It adopts quantile regression to multi-horizon forecasting. Each quantile forecast takes the form

$$\hat{y}^{(i)}(q, t, H) = f_q\big(H, y_{t-k:t}^{(i)}, z_{t-k:t}^{(i)}, x_{t-k:t+H}^{(i)}, s^{(i)}\big), i \in I,$$

where $\hat{y}^{(i)}(q, t, H)$ is predicted q-th quantile for entity i at time $t + H$, $f_q(\cdot)$ is a forecasting model, H is forecast horizon, $y_{t-k:t}^{(i)}$ are past targets, $z_{t-k:t}^{(i)}$ are past unknown observed inputs, $x_{t-k:t+H}^{(i)}$ are past and future known inputs, $s^{(i)}$ are static covariates, I is a set of unique entities in a given time-series dataset. The model outputs probabilistic forecasts through quantile predictions (10th, 50th, and 90th percentiles), providing both point estimates and uncertainty quantification.

Key Innovations. TFT hybrid architecture strategically combines long short-term memory (LSTM), which excel at capturing local temporal patterns, with Transformer components that model long-range dependencies, creating a powerful framework for complex forecasting tasks.

For enhanced interpretability, TFT employs variable selection networks to identify important features and multi-head attention to reveal meaningful temporal relationships like seasonality trends. These innovations collectively enable TFT to handle complex forecasting tasks while maintaining computational efficiency and model transparency.

TFT Architecture. TFT employs gating mechanisms to dynamically adjust network complexity, variable selection networks to prioritize relevant features at each step, and static encoders to integrate time-invariant metadata. For temporal processing, TFT combines a sequence-to-sequence layer for local patterns with an interpretable multi-head attention block for long-range dependencies.

1. *Gating mechanisms.* To dynamically assess the importance of features, TFT employs Gated Residual Networks (GRNs). The GRN operates on an input vector a and an optional context vector c through the following formulation:

$$GRN_\omega(a, c) = LayerNorm\big(a + GLU_\omega(\eta_1)\big),$$

$$\eta_1 = W_{1,\omega}\eta_2 + b_{1,\omega}, \eta_2 = ELU\big(W_{2,\omega}a + W_{3,\omega}c + b_{2,\omega}\big),$$

where ELU is the Exponential Linear Unit activation function, $\eta_1, \eta_2 \in R^{d_{model}}$ are intermediate layers, LayerNorm is standard layer normalization, and ω is an index to denote weight sharing [19,20]; and Gated Linear Units (GLU) for input $\eta, \eta \in R^{d_{model}}$, takes the form:

$$GLU_\omega(\eta) = \sigma\big(W_{4,\omega}\eta + b_{4,\omega}\big)\odot\big(W_{4,\omega}\eta + b_{4,\omega}\big),$$

where $\sigma(\cdot)$ is the sigmoid activation function, $W_{(\cdot)} \in R^{d_{model} \times d_{model}}$, $b_{(\cdot)} \in R^{d_{model}}$ are the weights and biases, $\odot$ is the element-wise Hadamard product, and d_{model} is the hidden state size.

2. *Variable selection networks* TFT determines the importance of each input variable through its variable selection networks, which analyze both static features and time-varying inputs for every

prediction. This process reveals which factors most influence forecasts and filters out irrelevant or noisy data that could reduce accuracy.

The model computes variable selection weights by processing both the flattened vector Ξ_t of all past inputs at time t and an external context vector c_s through a GRN, followed by a Softmax normalization:

$$v_{x_t} = Softmax\left(GRN_{x_t}(\Xi_t, c_s)\right),$$

where $x_t \in R^{m_x}$ represents the vector of variable importance weights, c_s is the static context vector from the static covariate encoder. For static variables, c_s is excluded since they already contain static information.

Each feature $\xi_t^{(j)}$ undergoes additional non-linear processing at each time step via its own GRN:

$$\tilde{\xi}_t^{(j)} = GRN_{\tilde{\xi}^{(j)}}\left(\xi_t^{(j)}\right).$$

The processed features are then combined using the selection weights:

$$\tilde{\xi}^{(j)} = \sum_{j=1}^{m_x} v_{x_t}^{(j)} \tilde{\xi}_t^{(j)},$$

where $v_{x_t}^{(j)}$ is the j-th element of the weight vector v_{x_t}.

3. *Static covariate (attribute feature) encoders*. TFT uses dedicated GRN encoders to transform static metadata into context vectors that guide temporal variable selection, local feature processing, and static-temporal fusion in the decoder.

4. *Multi-head attention mechanism*. Attention mechanisms scale values based on relationships between keys and queries in the following way:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q (Query) represents the current focus of the attention mechanism, K (Keys) encodes the content of all time steps, V (Values) contains actual information to aggregate, d_k is a scaling factor for stable gradients.

The multi-head attention approach improves upon standard attention by employing parallel attention heads that each focus on distinct feature representations. The outputs of different heads are then combined via concatenation [21].

5. Temporal fusion decoder uses the different layers.

6. Quantile outputs. TFT produces prediction intervals alongside point forecasts by directly outputting multiple percentiles (e.g., 10th, 50th, 90th) at each time step, computed via a linear transformation of the temporal fusion decoder's output.

Training and Evaluation. TFT is trained using a quantile loss function that jointly optimizes prediction percentiles, enabling probabilistic forecasting with uncertainty intervals. During evaluation, TFT employs rolling-window validation to assess performance across all forecast horizons, while attention weights and variable selection networks provide interpretable insights into feature importance and temporal patterns.

Compared to TiDE's RMSE-based training, TFT's quantile loss offers richer uncertainty quantification but requires more computation due to its LSTM and attention components. The model's performance is measured through horizon-specific metrics including quantile coverage, mean absolute error (MAE) for median predictions, and analysis of attention patterns for temporal dependencies.

Sequence-to-Sequence Plus Attention (Seq2SeqPlus) [22,23].

Key Innovations. Seq2SeqPlus introduces critical enhancements over the traditional Seq2Seq model, primarily through the attention mechanism, which allows dynamic focus on relevant input tokens during decoding, improving handling of long sequences [24].

The Transformer architecture replaces RNNs (Recurrent Neural Network) with self-attention, enabling parallel processing and capturing long-range dependencies more effectively.

Seq2SeqPlus Architecture. Seq2SeqPlus improves upon Seq2Seq by integrating attention, Transformer blocks, and hybrid mechanisms. It consists of two main components—Encoder Block and Decoder Block.

Encoder Block. The encoder processes sequential input data to generate contextual annotations. We use a bidirectional RNN (BiRNN) to capture both past and future temporal dependencies within a fixed window. Forward RNN processes the time-series x_j chronologically (from x_1 to x_{T_x}), producing hidden states $\vec{h}_j$ that encode historical trends:

$$\vec{h}_j = RNN_{forward}(x_j, \vec{h}_{j-1}),$$

backward RNN processes the time-series x_j in reverse (from x_{T_x} to x_1), generating hidden states $\overleftarrow{h}_j$ to incorporate future context within the window:

$$\overleftarrow{h}_j = RNN_{backward}(x_j, \overleftarrow{h}_{j+1}).$$

These states are concatenated into an annotation vector $h_j = [\vec{h}_j; \overleftarrow{h}_j]$, which summarizes both past and future context around x_j. The annotations $\{h_j\}_{j=1}^{T_x}$ are later used by the decoder to compute dynamic attention weights.

Decoder Block. The decoder generates future predictions step-by-step, conditioned on both past decoder states and relevant historical patterns identified by the attention mechanism. At each step i, the decoder computes the conditional probability of the next value y_i as:

$$p(y_1, \dots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i),$$

where s_i is the decoder's hidden state, y_{i-1} is the previous prediction, and c_i is a time-dependent context vector encoding attended historical observations.

The decoder state s_i updates recursively using:

$$s_i = f(s_{i-1}, y_{i-1}, c_i),$$

where f is a recurrent unit (LSTM). The context vector c_i is a weighted sum of encoder annotations h_j:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j,$$

with weights α_{ij} computed via a soft attention mechanism:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})},$$

here, e_{ij} scores how well the context window around h_j aligns with the current forecast step i, parametrized by a feedforward network α.

4. Results

The primary raw data we used for research were contained in two original datasets. The first one (Figure 1) consisted of 55 columns containing billing data of 105 527 households in Volyn region of Ukraine on monthly natural gas consumption from January 2019 to April 2023 (52 periods totally) in cubic meters. The dataset columns were as follows:

Column 1 "ID"—household ID;

Column 2 "Category"—household categories prescribed due to Gas Distribution System Code depending on the purpose of consumption (home heating, water heating, cooking), volumes of consumption and housing floor area [25].

Column 3 "GDS"—Household affiliation with a gas distribution station (39 GDS totally);

Columns 4-55 "Consumption"—Monthly volumes of consumption from January 2019 to April 2023 in cubic meters.

Exploratory data analysis we had performed showcased the presence of seasonality and autocorrelation in consumption volumes, as well as a statistically significant level of correlation between the volume of natural gas consumed and environmental temperature in certain categories of household consumers (categories 4,5,6,7,10,11,12,14,15,16,17). (Figures 2 and 3; Table 1).
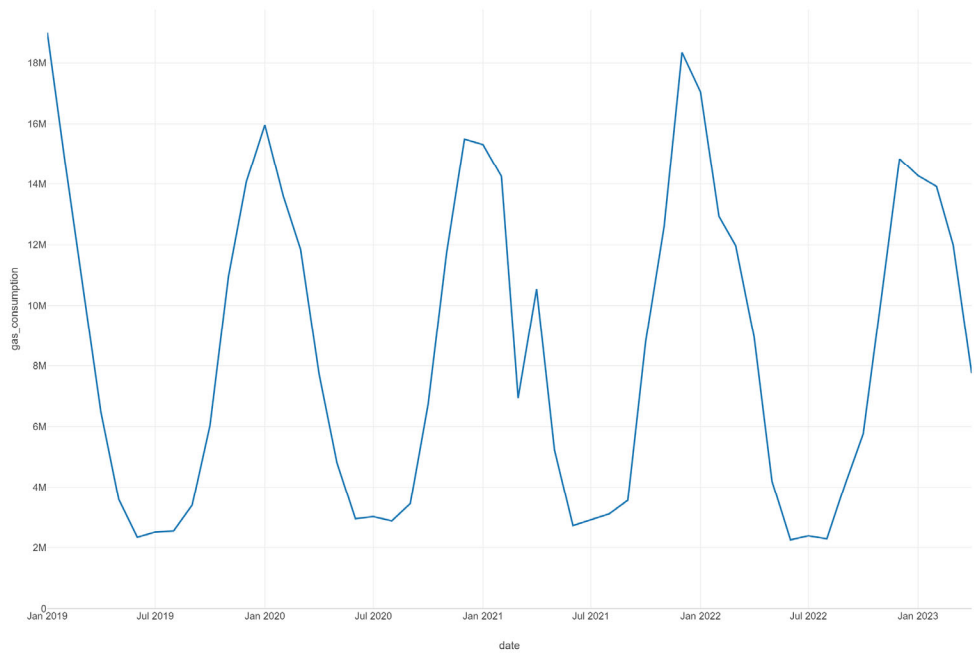
Figure 2. Seasonality of gas consumption.

A slight level of correlation and autocorrelation is observed in consumer groups that use natural gas for cooking only (categories 1,2,8) or for cooking and/or water heating (categories 3,9).

The second dataset consisted of 39 rows and 53 columns containing meteorological data of average monthly air temperature per month measured on each of 39 GDS in the period of January 2019 to April 2023. The dataset columns were as follows (Figures 2 and 3):

Column 1—"GDS"—name of a gas distribution station

Columns 2—53 "Temperature"—average monthly temperature measured on a gas distribution station.
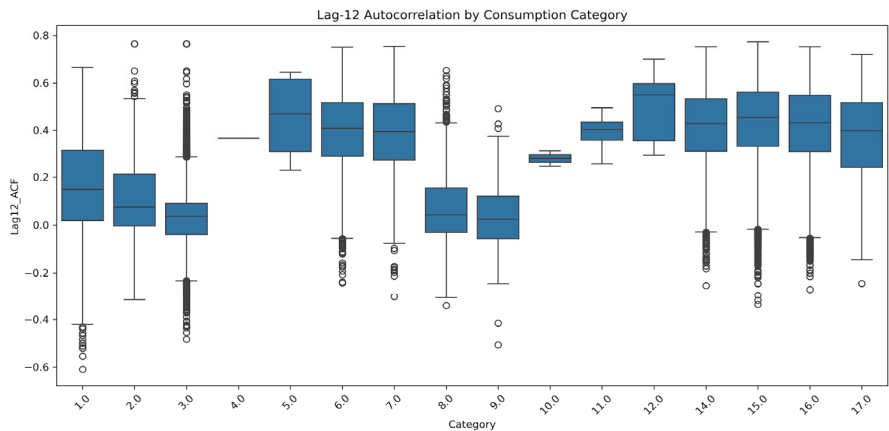

Figure 3. Autocorrelation of gas consumption.

The both primary datasets had been transformed into a long dataset consisting of 5 381 877 rows and 5 columns aiming for further model training processes.

The authenticity of our data origin caused missing values in certain months for some consumers totalling 113014 rows, so we had to fix it for further development.

Due to our uncertainty regarding the reasons of NaN values origin—whether because there were no gas consumption in particular months or the meter readings were not taken—we are to apply two approaches resulting in two different datasets the models to be trained on.

Table 1. Correlation between consumption and environmental temperature.

Category	Correlation	P Value	t Value
1	-0.1077196787164641	0	-143.14615241829136
2	-0.1029918495127673	1.6738761249561719e-195	-29.90818397839599
3	-0.0898838682933454	0	-71.65630977055373
4	-0.6111798276700345	1.901215406512052e-6	-5.405310539643045
5	-0.5621617874528735	2.1546658143801656e-18	-9.660890728796526
6	-0.5763568492042973	0	-410.6086271959947
7	-0.5517534525364725	0	-233.5648189798738
8	-0.1335519359322037	0	-39.527176227944956
9	-0.13760491992730364	4.822132033816754e-27	-10.82111701876933
10	-0.6684467451396608	1.636614697340057e-14	-8.987398331518012
11	-0.5852469722518012	3.836439139880669e-20	-10.258210616748146
12	-0.706282363260936	7.886147076541504e-40	-15.868948788773727
14	-0.5898449444053748	0	-391.57682330968424
15	-0.6226004374063476	0	-995.114255839583
16	-0.5921695995263664	0	-514.3867148669385
17	-0.508729170725958	0	-85.0251952906726

A simpler approach involves replacing the "NaN" values with "0" values. Given the small number of such values (about 2%), we can assume that this will not have a critical impact on the performance of the models.

More sophisticated approach involves replacing NaN values using Household-Specific Seasonal Imputation and K-Nearest Neighbors (KNN) techniques [26].

The Household-Specific Seasonal Imputation technique was chosen because it considers seasonality and fills missing values using the average gas consumption for the same month in other years for each consumer individually.

However, there was an issue—a certain part of a particular household values for given months in different years were missing in the original dataset. The solution was to apply a more complicated technique—the KNN, which uses patterns from similar households to estimate missing values and needs a large dataset with strong correlations.

Once imputation had been done, we compared the original and imputed data to check if the imputed values follow the original distribution. Mean & Variance Check resulted in a Table 2 confirming the imputation had been succeeded.

Table 2. Statistics of Replaced by 0 and Imputed datasets.

Dataset	Replaced by 0	Totally Imputed
count	5.381877e+06	5.381877e+06
mean	7.925611e+01	8.081569e+01
std	1.251031e+02	1.254847e+02
min	0.000000e+00	1.000000e-02
25%	8.950000e+00	9.780000e+00

50%	2.231000e+01	2.360000e+01
75%	9.947000e+01	1.006500e+02
max	5.506240e+03	5.506240e+03
var	15650.793002383958	15746.508345587017

As a result of data wrangling and preprocessing we obtained two final time series datasets (NaN values replaced by 0 and imputed within special techniques respectively) on consumption data with pronounced seasonality, statistically significant correlation to temperature data and featuring consumption category data as a supporting attribute (Figure 4). The final datasets columns were as follows:

Column 1 "ID"—household ID (Series Identifier);

Column 2 "Date"—month and year (Timestamp);

Column 3 "Temperature"—average monthly temperature measured on a gas distribution station the household is affiliated with (Covariate Feature);

Column 4 "Category"—household category (Attribute Feature);

Column 5 "Consumption"—Monthly volumes of consumption from January 2019 to April 2023 in cubic meters (Target Variable).

Now they are ready to be performed in a model training process.

Three cutting age Deep Neural Network models—Seq2seqPlus, TiDE and TFT had been trained on the final datasets with following parameters:

Chronological data splitting was applied—the earliest 80% of rows were assigned to training, the next 10% to validation, and the most recent 10% to test.

Forecast horizon was set to 12 months pursuing to maintain a year-ahead forecasting and the context window that defines the input lags to the model for each time series was set to 24.

Data granularity—monthly

The RMSE metric was set as an optimization objective.

The following metrics were applied for models evaluation:

MAE: The mean absolute error is the average absolute difference between the target values and the predicted values. This metric ranges from zero to infinity; a lower value indicates a higher quality model.

RMSE: The root-mean-squared error is the square root of the average squared difference between the target and predicted values. RMSE is more sensitive to outliers than MAE,so if you're concerned about large errors, then RMSE can be a more useful metric to evaluate. Similar to MAE, a smaller value indicates a higher quality model (0 represents a perfect predictor).

RMSLE: The root-mean-squared logarithmic error metric is similar to RMSE, except that it uses the natural logarithm of the predicted and actual values plus 1. RMSLE penalizes under-prediction more heavily than over-prediction. It can also be a good metric when you don't want to penalize differences for large prediction values more heavily than for small prediction values. This metric ranges from zero to infinity; a lower value indicates a higher quality model. The RMSLE evaluation metric is returned only if all labels and predicted values are non-negative.

r^2: r squared (r^2) is the square of the Pearson correlation coefficient between the labels and predicted values. This metric ranges between zero and one. A higher value indicates a closer fit to the regression line.

WAPE: Weighted absolute percentage error is the overall difference between the value predicted by a model and the values observed over the values observed. Compared to RMSE, WAPE is weighted towards the overall differences rather than individual differences, which can be highly influenced by low or intermittent values. A lower value indicates a higher quality model.

After the training was performed and models were tested, we obtained the following results (Table 3).

Table 3. Performance metrics of trained models.

Model	Dataset	MAE	RMSE	RMSLE	R^2	WAPE
Seq2seq + attention	Inputed	28.721	61.134	0.485	0.804	25.194
	NaN	59.075	101.258	0.697	0.772	52.293
TiDE	Inputed	35.356	71.154	0.435	0.801	31.013
	NaN	36.62	72,37	0.489	0,789	31.605
TFT	Inputed	30.657	64.31	0.502	0.791	26.891
	NaN	33.511	66.202	0.576	0.776	28.921

Model feature attribution is expressed using the SHAP methodology [27].

SHAP (SHapley Additive exPlanations) is a powerful method for interpreting machine learning models, including deep neural networks, by quantifying the contribution of each feature to the model's predictions. SHAP is based on Shapley values from cooperative game theory, which fairly distribute the prediction outcome among the input features. It explains how much each feature contributes (positively or negatively) to a given prediction.

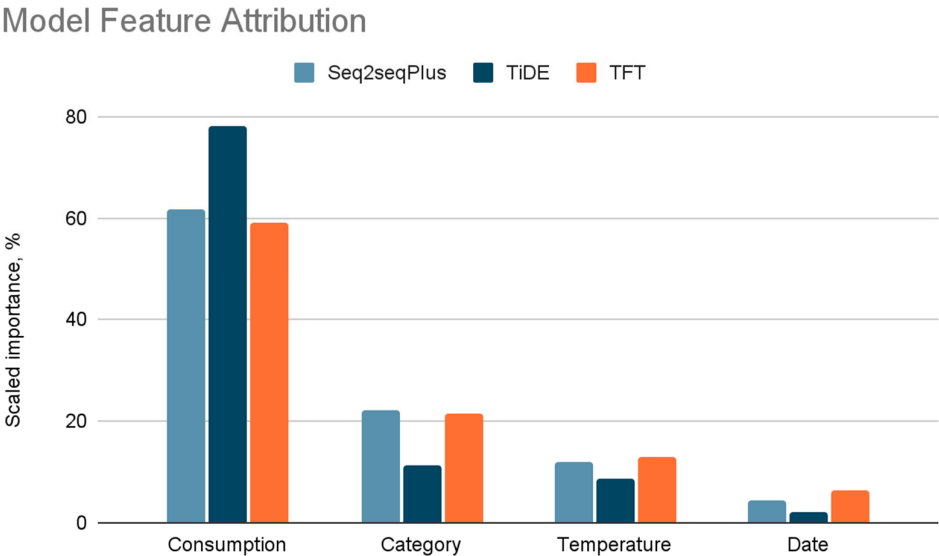The feature attribution of trained models is presented in Figure 4.

Figure 4. Model feature attribution of trained models.

Model feature attribution tells us how important each feature is when making a prediction. Attribution values are expressed as a percentage; the higher the percentage, the more strongly that feature impacts a prediction on average.

5. Discussion

Considering the results of our research the following conclusions can be drawn.

Past consumption is a dominant feature of all trained models, which may confirm their autoregressive origins and high importance of historical consumption data in our context. As well these demonstrate strong seasonal and temporal patterns of household consumption.

Meanwhile temperature and category features are to be considered as supportive due to their moderate influence level—they provide useful fine-tuning rather than the primary signal.

This confirms the fact that our models are specifically designed to handle both autoregressive dependencies (past gas consumption) and exogenous factors (temperature, category, etc.), as well as their ability to learn temperature effects indirectly through past gas consumption.

Anyway, the importance of temperature feature was shown previously in our recent paper using Granger Causality test, which means our models should benefit from including it explicitly [28].

The consumption category feature is valuable as well because it helps differentiate between groups with strong and weak seasonal behavior. It prevents seasonality bias, ensuring that groups with non-seasonal behavior are still predicted accurately and interacts with temperature, helping the model decide when temperature matters more or less.

Without this feature, our models would treat all consumers the same, potentially leading to overgeneralization when the model may only learn the dominant pattern (e.g., strong seasonality) and ignore less seasonal groups, or loss of detail when consumers with flat or irregular consumption patterns might not be predicted well.

There are several general trends we spectact regarding models trained on "Totally Imputed" vs. "Zero filled" datasets due to performance metrics obtained:

Totally Imputed data consistently improves performance across all models compared to Zero filled data.

R^2 drops slightly with Zero filled, meaning the models explain less variance.

Seq2Seq suffers the most from missing values, while TiDE is the most stable.

All models show approximately the same levels of accuracy, especially on processed and imputed data, which may indicate the acceptability of all described DNN techniques for achieving our goals using real life data.

When finally choosing a model for use at the deployment stage, one should consider such forecasting parameters as completeness and robustness of the data and forecasting horizon. Of course, further periodic monitoring of accuracy levels and retraining on new data is recommended to maintain the functional parameters of the models.

Model Insights and Suitability. Seq2SeqPlus model with attention mechanism demonstrated strong performance in terms of MAE and RMSE, particularly on the imputed dataset. It effectively captured temporal dependencies and historical consumption trends, leading to reliable short-term forecasts.

However, its MAPE was comparatively high, suggesting that the model was less effective at capturing low-consumption fluctuations, a common issue when prioritizing absolute over percentage-based errors.

While *Seq2SeqPlus* may require additional feature engineering to generalize across consumption categories with non-seasonal behaviors, it remains a strong candidate for short-term aggregate forecasting, particularly when the input data is complete and well-processed.

TiDE emerged as the most balanced model across all evaluation metrics. It showed robust performance on both datasets, with minimal degradation in the presence of missing data, which points to its stability under real-world conditions. Although it did not outperform the other models in any single metric, its consistency and relatively low variance make it a reliable choice for general-purpose forecasting.

TiDE's simplicity also offers faster training times and easier hyperparameter tuning, which are advantageous in operational environments that require frequent model retraining.

TFT model offered the greatest flexibility and interpretability, excelling in handling diverse feature types (static, known, and observed time-varying variables).

While slightly underperforming Seq2SeqPlus on some absolute metrics (MAE, RMSE), its ability to model complex temporal relationships and provide explanatory insight via attention mechanisms and SHAP analysis makes it valuable for long-term forecasting and strategic planning applications.

However, TFT's complexity and sensitivity to imputation strategies make it more challenging to deploy in volatile or sparse data environments without additional regularization and tuning.

Based on the comparative analysis and the specific forecasting objectives the models can be recommended for different use cases (Table 4):

Table 4. Predictive models that can be used.

Use Case	Recommended Model	Rationale
Short-term forecasting	Seq2SeqPlus	High accuracy in total volume, strong short-horizon memory
General-purpose deployment	TiDE	Stable, low-variance predictions; robust to data imperfections
Long-term forecasting + explainability	TFT	Captures long-term patterns; allows interpretability via attention

For real-world deployment, it is advisable to consider MAE or WAPE as primary optimization objectives, as these metrics align more closely with minimizing total consumption error—critical for operational planning. Furthermore, an ensemble approach combining the strengths of Seq2SeqPlus and TiDE may offer a balance between responsiveness and robustness.

The influence of data preprocessing—particularly the treatment of missing values—proved to be a critical factor in model performance. All three models consistently showed better accuracy on the imputed dataset compared to the version where missing values were simply replaced with zeros. The difference was most apparent in the MAE and RMSE values, which substantially degraded on the NaN-filled data.

This suggests that imputation strategies should be treated as an integral part of model development, not just a preprocessing step.

Given the high seasonality observed in some consumer categories, imputing based on historical seasonal trends and similar profile behavior appears to preserve important patterns that the models can effectively learn from.

For large-scale utility forecasting applications, it is essential to adopt context-aware imputation methods, such as seasonal medians or clustering-based smoothing, to maintain model accuracy while minimizing information loss.

6. Conclusions

The application of deep neural network (DNN) models to gas consumption forecasting represents a significant advancement in aligning predictive analytics with the operational and regulatory needs of modern energy systems. Unlike traditional statistical approaches, DNN-based models—such as Seq2SeqPlus, TiDE, and TFT—offer the flexibility to incorporate complex temporal patterns, consumer heterogeneity, and external variables such as weather or calendar effects.

In regulated energy markets, where forecasting accuracy directly impacts supply planning, tariff setting, and network balancing obligations, the ability of these models to deliver reliable, high-resolution forecasts at scale is particularly valuable.

Moreover, explainable architectures like TFT further enhance transparency, supporting compliance with regulatory requirements for model interpretability and auditability.

Equally important, these models can be tuned to prioritize total consumption accuracy, which aligns well with regulatory metrics such as aggregate demand forecasting accuracy, under-/over-supply penalties, and carbon accounting targets.

As regulators increasingly demand real-time responsiveness, robust handling of incomplete data, and forward-looking risk modeling, DNN-based forecasting systems are poised to become a cornerstone of compliant, data-driven energy operations.

Ultimately, the successful deployment of these models requires more than technical precision—it demands careful integration with data governance, regulatory frameworks, and operational workflows. When implemented thoughtfully, DNN forecasting tools can significantly enhance both regulatory compliance and strategic energy planning in the gas sector.

Data Availability Statement: All data used in this study are publicly available, all datasets are in the public domain and were available as of 01 May 2025. The results of the author's calculations are the author's own and are reliable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sala, D., Pavlov, K., Pavlova, O., Demchuk, A., Matiichuk, L., & Cichoń, D. (2023). Determining of the Bankrupt Contingency as the Level Estimation Method of Western Ukraine Gas Distribution Enterprises' Competence Capacity. Energies, 16(4), 1642. https://doi.org/10.3390/en16041642

2. Pavlov K., Pavlova O., Korotia M. et al. (2020) Determination and Management of Gas Distribution Companies' Competitive Positions. In: Mrugalska B., Trzcielinski S., Karwowski W., Di Nicolantonio M., Rossi E. (eds) Advances in Manufacturing, Production Management and Process Control. AHFE 2020. Advances in Intelligent Systems and Computing, vol 1216. Springer, Cham. DOI https://doi.org/10.1007/978-3-030-51981-0_38

3. Anagnostis, A., Papageorgiou, E., & Bochtis, D. (2020). Application of Artificial Neural Networks for Natural Gas Consumption Forecasting. Sustainability, 12(16), 6409. https://doi.org/10.3390/su12166409

4. Almazrouee, A. I., Almeshal, A. M., Almutairi, A. S., Alenezi, M. R., & Alhajeri, S. N. (2020). Long-Term Forecasting of Electrical Loads in Kuwait Using Prophet and Holt–Winters Models. Applied Sciences, 10(16), 5627–5627. https://doi.org/10.3390/app10165627.

5. Svoboda, R., Sebastián Basterrech, Kozal, J., Platoš, J., & Michał Woźniak. (2024). A natural gas consumption forecasting system for continual learning scenarios based on Hoeffding trees with change point detection mechanism. Knowledge-Based Systems, 112482–112482. https://doi.org/10.1016/j.knosys.2024.112482.

6. Tian, N., Shao, B., Bian, G., Zeng, H., Li, X., & Zhao, W. (2024). Application of forecasting strategies and techniques to natural gas consumption: A comprehensive review and comparative study. Engineering Applications of Artificial Intelligence, 129, 107644. https://doi.org/10.1016/j.engappai.2023.107644.

7. On Approval of the Methodology for Determining and Calculating the Tariff for Natural Gas Distribution Services (2016). Official web portal of the Parliament of Ukraine. https://zakon.rada.gov.ua/laws/show/z1434-16#Text.

8. On Approval of the Methodology for Determining and Calculating Tariffs for Natural Gas Transmission Services for Entry and Exit Points Based on Multi-Year Incentive Regulation (2015). Official web portal of the Parliament of Ukraine. https://zakon.rada.gov.ua/laws/show/z1388-15#Text

9. On Approval of the Gas Transmission System Code (2015). Official web portal of the Parliament of Ukraine.https://zakon.rada.gov.ua/laws/show/z1378-15#Text

10. Gas Transmission System Operator of Ukraine. (2025). Informational Platform (IP). Retrieved June 18, 2025, from https://tsoua.com/en/business-services/how-to-become-a-customer/information-platform/

11. National Commission for State Regulation of Energy and Public Utilities (n.d.). Charges in the presence of an individual meter. Retrieved June 18, 2025, from https://www.nerc.gov.ua/sferi-diyalnosti/prirodnij-gaz/pobutovi-spozhivachi/narahuvannya-pri-nayavnosti-individualnogo-lichilnika

12. Sala, D.; Pavlov, K.; Bashynska, I.; Pavlova, O.; Tymchyshak, A.; Slobodian, S. Analyzing Regulatory Impacts on Household Natural Gas Consumption: The Case of the Western Region of Ukraine. Appl. Sci. **2024**, 14, 6728. https://doi.org/10.3390/app14156728

13. Zhang, Z. (2022). Prediction of Economic Operation Index Based on Support Vector Machine. Mobile Information Systems, 2022, 1–11. https://doi.org/10.1155/2022/3232271.

14. Dabrowski, J.J., Zhang, Y., Rahman, A. (2020). ForecastNet: A Time-Variant Deep Feed-Forward Neural Network Architecture for Multi-step-Ahead Time-Series Forecasting. In: Yang, H., Pasupa, K., Leung, A.CS., Kwok, J.T., Chan, J.H., King, I. (eds) Neural Information Processing. ICONIP 2020. Lecture Notes in Computer Science(), vol 12534. Springer, Cham. https://doi.org/10.1007/978-3-030-63836-8_48

15. Dabrowski, J. J., Zhang, Y., & Rahman, A. (2020). ForecastNet: A Time-Variant Deep Feed-Forward Neural Network Architecture for Multi-step-Ahead Time-Series Forecasting. Lecture Notes in Computer Science, 579–591. https://doi.org/10.1007/978-3-030-63836-8_48

16. Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., & Yu, R. (2023). Long-term Forecasting with TiDE: Time-series Dense Encoder. ArXiv.org. https://arxiv.org/abs/2304.08424#

17. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022). Are Transformers Effective for Time Series Forecasting? ArXiv.org. https://arxiv.org/abs/2205.13504

18. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

19. Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). ArXiv.org. https://arxiv.org/abs/1511.07289

20. Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. ArXiv.org. https://arxiv.org/abs/1607.06450

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

22. Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv.org. https://arxiv.org/abs/1409.0473

23. Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. ArXiv.org. https://arxiv.org/abs/1508.04025

24. Sutskever, I., Vinyals, O., & Le, Q. V. (2025). Sequence to Sequence Learning with Neural Networks. ArXiv.org. https://arxiv.org/abs/1409.3215

25. Household consumer groups | Energy Map. (2015). Energy Map. https://map.ua-energy.org/en/datasets/f7882596-db70-4565-bb57-03c778a7334b

26. 1.6. Nearest Neighbors. (2025). Scikit-Learn. https://scikit-learn.org/stable/modules/neighbors.html

27. Divyani Sen. (2025). Explainable Deep Learning for Time Series Analysis: Integrating SHAP and LIME in LSTM-Based Models. Journal of Information Systems Engineering and Management, 10(16s), 412–423. https://doi.org/10.52783/jisem.v10i16s.2627

28. Sala, D.; Pavlov, K.; Bashynska, I.; Pavlova, O.; Tymchyshak, A.; Slobodian, S. Analyzing Regulatory Impacts on Household Natural Gas Consumption: The Case of the Western Region of Ukraine. Appl. Sci. 2024, 14, 6728. https://doi.org/10.3390/app14156728