

Article

Electric Load Forecasting for a Quicklime Company Using a Temporal Fusion Transformer

Jersson X. Leon-Medina ^{1,2,*} , Diego A. Tibaduiza ³ , Claudia Patricia Siachoque Celys ¹,
Bernardo Umbarila Suarez ¹ and Francesc Pozo ^{4,5,*} 

¹ Grupo de Investigación en Biochar, Suelo y Cambio Climático (Pyrosfera), Suministros Mineros e Industriales de Colombia LTDA-SUMININCO LTDA, Km1 vía Nobsa-Duitama Vereda Guaquida, Nobsa 152280, Colombia; cpsiachoque@gmail.com (C.P.S.C.); bernardoumbarila@gmail.com (B.U.S.)

² Escuela de Ingeniería Electromecánica, Facultad Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Carrera 18 con Calle 22, Duitama 150461, Colombia

³ Departamento de Ingeniería Eléctrica y Electrónica, Universidad Nacional de Colombia Sede Bogotá, Av 45 Carrera 30, Bogotá 111321, Colombia; dtibaduizab@unal.edu.co

⁴ Control, Data, and Artificial Intelligence (CoDALab), Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Campus Diagonal-Besòs (CDB), Universitat Politècnica de Catalunya (UPC), Eduard Maristany 16, 08019 Barcelona, Spain

⁵ Institute of Mathematics (IMTech), Universitat Politècnica de Catalunya (UPC), Pau Gargallo 14, 08028 Barcelona, Spain

* Correspondence: jersson.leon@uptc.edu.co (J.X.L.-M.); francesc.pozo@upc.edu (F.P.)

Abstract

Accurate short-term electric load forecasting is essential for the operation and management of energy-intensive manufacturing processes such as quicklime production, for which power demand is driven by stage-based operation, fixed schedules, and abrupt load transitions. This study presents a data-driven forecasting framework based on a Temporal Fusion Transformer (TFT) model applied to real industrial measurements collected during 2024 from an operating quicklime production plant. The dataset comprises hourly average power demand records (kW) measured at a plant level, stage-dependent motor operation, and a fixed working schedule from 08:00 to 18:00 (Monday to Friday), with weekends and non-operational hours characterized by near-zero load. Coke consumption during the calcination stage is included as an additional contextual variable. The TFT model is trained for multi-horizon forecasting and provides probabilistic prediction intervals through quantile regression. Weekly evaluations demonstrate that the proposed approach accurately captures start–stop behavior, peak-load periods, and structured inactivity intervals. In addition to point-wise accuracy metrics, cumulative energy is evaluated by integrating hourly power over the forecasting horizon, allowing the assessment of energy preservation at the operational level. The resulting energy deviation reaches 4.78% for the full horizon and 5.25% when restricted to active production hours, confirming strong consistency between predicted and actual cumulative energy. A comparative analysis against LSTM, GRU, and N-BEATS models shows that recurrent architectures achieve lower MAE and RMSE values, while the TFT model delivers superior cumulative energy consistency, highlighting a trade-off between instantaneous accuracy and operational energy fidelity. Overall, the results demonstrate that the proposed TFT-based framework provides a robust and practically relevant solution for short-term industrial electric load forecasting and decision support in stage-driven manufacturing systems under real operating conditions.

Keywords: energy consumption; quicklime production; hourly load modeling; industrial motor scheduling; Temporal Fusion Transformer (TFT); time series forecasting



Academic Editor: Frank Werner

Received: 19 January 2026

Revised: 5 March 2026

Accepted: 5 March 2026

Published: 10 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Accurate energy demand forecasting has become a crucial component in optimizing energy-intensive industrial processes [1]. In the context of manufacturing sectors such as quicklime production, effective forecasting not only supports operational efficiency but also contributes to strategic planning and sustainability initiatives [2]. Quicklime manufacturing involves multiple stages, such as crushing, screening, calcination, and grinding, each associated with distinct energy consumption patterns influenced by process dynamics and scheduling [3]. As industries strive to reduce costs and carbon emissions, the ability to anticipate and manage energy demand with precision becomes a competitive advantage.

Traditionally, forecasting approaches such as ARIMA and Exponential Smoothing have served as standard tools for time series prediction [4]. While these classical models perform well under linear and stationary conditions, they often struggle to capture the nonlinearities, complex interdependencies, and external influences prevalent in industrial environments [5]. The proliferation of sensor-based data in modern settings has made richer datasets available, but it also introduces modeling challenges that surpass the capabilities of traditional statistical techniques [6]. Comprehensive reviews have highlighted a broad spectrum of methodologies—including multiple regression, adaptive load forecasting, stochastic time series, ARMAX models with genetic algorithms, fuzzy logic, neural networks, and expert systems—as preferred alternatives for addressing these complexities [7]. This diversity reflects the multifaceted nature of energy forecasting in industrial contexts, where tailored solutions are often necessary.

Recent advances in machine learning, particularly the development of neural networks such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs), have demonstrated significant potential in predicting the power profiles of manufacturing processes before production begins. Models oriented towards batch production, for example, have achieved notable accuracy, with absolute prediction errors as low as 5% [8]. These approaches showcase the advantages of leveraging deep learning to capture temporal dependencies and complex relationships in industrial energy consumption data.

In terms of short-term load forecasting (STLF), recent work has shifted toward hybrid deep learning architectures, uncertainty quantification through probabilistic modeling, and the integration of diverse IoT and environmental data.

Some studies have concentrated on developments in the design of hybrid deep learning architectures; these are focused on combining different neural network layers (such as CNN for spatial/local features and LSTM for temporal dependencies) to improve accuracy. Some examples make use of the combination of CNN and LSTM with techniques such as Empirical Mode Decomposition (EMD) to simplify complex energy demand profiles [9], or K-means for improving classification tasks [10]. The use of attention mechanisms is combined with with LSTM [11] and a convolutional bidirectional long short-term memory autoencoder (CBLSTM-AE) model, and Beluga Whale Optimization (BWO) for hyperparameter tuning [12]. In terms of probabilistic and interval forecasting, some works have tackled uncertainty by predicting ranges or probability densities. Some examples include Quantile Regression (QR) [10,13–16] and Bayesian Approaches [17].

Optimization and Automated Machine Learning (AutoML) have also been studied by emphasizing how the model is tuned, using meta-heuristic algorithms to find the best hyperparameters, such as genetic algorithms [13], Beluga Whale Optimization [12], Bayesian Optimization [9] and AutoML for feature selection and model training, which further improved accuracy and generalization [18].

In terms of domain-specific applications, several works have studied applications on hospitals [19], industrial plants [20], manufacturing [21], residential, microgrids,

smartgrids [9,10,15,17,22] and day-ahead interval forecasting method for photovoltaic (PV) power [14].

Based on these developments, deep learning has emerged as a powerful alternative to time series forecasting in recent years, offering superior performance in capturing complex temporal dynamics [23]. Architectures like LSTM and GRU excel in learning long-range dependencies and nonlinear relationships [24]. Comparative studies across diverse domains—including energy systems—consistently demonstrate that deep learning models outperform statistical methods, especially for multivariate and multi-horizon forecasting tasks [25].

Among these advanced architectures, the Temporal Fusion Transformer (TFT) has attracted increasing attention due to its hybrid structure, which combines the strengths of recurrent layers for local processing with interpretable attention mechanisms for capturing long-term dependencies [26]. Originally introduced by Lim et al. in 2021 [27], TFT was designed to handle heterogeneous inputs, including static metadata, known future events, and time-varying covariates. Its use of gating layers, variable selection networks, and quantile regression makes it particularly suitable for forecasting tasks that require both accuracy and interpretability [28].

The benefits of TFT have been validated in various sectors. For instance, Gokhale et al. in 2023 [29] applied a transfer learning framework with TFT to forecast electricity consumption in home energy management systems, achieving a 15% reduction in mean absolute error and a 2% reduction in energy costs through improved predictive control. Similarly, Maragos and Refanidis in 2025 [30] conducted a comprehensive evaluation of state-of-the-art forecasting models, including TFT, on a dataset combining energy consumption and weather variables from 24 European countries, highlighting the robust performance of transformer-based models in complex multivariate scenarios.

Despite the demonstrated potential of these models, their application in the industrial manufacturing sector, particularly in quicklime production, remains underexplored. This study addresses this gap by implementing a Temporal Fusion Transformer to forecast energy demand in an operating quicklime plant using real industrial measurements collected during 2024. The dataset consists of hourly electrical consumption records obtained at the plant level, enriched with contextual information such as working schedules, production stages, and exogenous variables including coke consumption during the calcination process. By integrating multivariate time series with temporal and operational features derived from actual plant operation, the proposed approach captures realistic consumption patterns and variability inherent to industrial production environments.

Contributions

The main contributions of this study are summarized as follows:

1. **Industrial probabilistic forecasting under discontinuous load dynamics.** This work investigates short-term load forecasting in a stage-driven manufacturing environment characterized by fixed schedules, abrupt transitions, and structured inactivity periods. The study provides empirical evidence on how a transformer-based architecture performs under these non-continuous industrial conditions, which are rarely addressed in the prior forecasting literature.
2. **Integration of operational covariates in a multi-horizon TFT framework.** A forecasting framework based on the Temporal Fusion Transformer (TFT) is implemented by combining historical load, stage encoding, coke consumption, and known-future temporal variables. The model is configured for multi-horizon probabilistic prediction, enabling the analysis of heterogeneous inputs in a real industrial setting.

3. **Evaluation beyond point-wise accuracy using cumulative energy fidelity.** In addition to MAE and RMSE, the study introduces a cumulative energy deviation metric derived from hourly power integration, allowing the assessment of forecasting performance from an operational planning perspective, for which total energy preservation is critical.
4. **Probabilistic forecasting and uncertainty calibration in industrial operation.** The TFT model is evaluated using quantile regression to produce prediction intervals, and uncertainty quality is analyzed through coverage metrics, providing insight into risk-aware decision-making for energy management.
5. **Interpretability-oriented assessment of industrial forecasting variables.** The study examines the role of operational covariates within the TFT architecture, highlighting the relevance of stage-dependent dynamics and contextual variables in explaining industrial load behavior.
6. **Transparent comparative evaluation against deep learning baselines.** A consistent experimental protocol is implemented to compare TFT with LSTM, GRU, and N-BEATS models using identical data splits, covariates, and evaluation horizons, ensuring fairness and reproducibility.

The remainder of this paper is organized as follows. Section 1 introduces the research context and motivates the need for accurate short-term electric load forecasting in stage-driven industrial processes. Section 2 describes the materials and methods, including the data acquisition process from an operating quicklime plant, the construction of the hourly multivariate dataset, and the implementation of the Temporal Fusion Transformer (TFT) for multi-horizon and probabilistic load forecasting. Section 3 presents and discusses the forecasting results, covering deterministic performance, cumulative energy deviation, probabilistic prediction intervals, and residual analysis. Finally, Section 4 summarizes the main conclusions of the study and outlines potential directions for future research.

2. Materials and Methods

2.1. Dataset Description

An hourly dataset was collected from an operating quicklime production plant over a full year (2024) to characterize the electrical energy consumption of the manufacturing process under real industrial conditions. The recorded data reflect the actual operational behavior of the plant, including fixed working schedules, stage-based production logic, and abrupt load transitions associated with the activation and deactivation of process equipment.

Electrical energy consumption was measured at the main electrical distribution level, capturing the aggregated power demand of the production facility. The recorded load implicitly reflects the operation of multiple electric motors associated with different stages of the process, such as material reception and crushing, screening, post-screening handling, and kiln operation with grinding. Although individual motor measurements were not used as direct inputs to the forecasting model, the aggregated signal preserves the characteristic signatures of stage-dependent motor activity. The schematic representation of the hourly energy data acquisition and processing workflow for the quicklime production plant is shown in Figure 1.

The production workflow is organized into four main stages: reception and crushing, screening, post-screening, and kiln with grinding. These stages are executed sequentially following the plant's standard operating procedure. Production activities take place exclusively on working days (Monday to Friday) within a fixed daily schedule from 08:00 to 18:00, while nights, weekends, and non-operational periods are characterized by zero or near-zero electrical demand.

During active production hours, the measured electrical load exhibits natural variability caused by partial-load motor operation, process control actions, and transient events inherent to industrial operation. These fluctuations are directly reflected in the recorded dataset, allowing the forecasting model to learn realistic consumption patterns without the need for artificial load synthesis or stochastic simulation assumptions.

In addition to electrical measurements, an auxiliary operational variable representing hourly coke consumption during the calcination stage was included as an exogenous input. Coke constitutes the primary non-electrical energy input of the kiln, and its recorded usage provides valuable contextual information related to production intensity and process state. The inclusion of this variable enables a more comprehensive representation of the energy dynamics of the quicklime production process and supports subsequent analyses related to energy efficiency and operational planning.

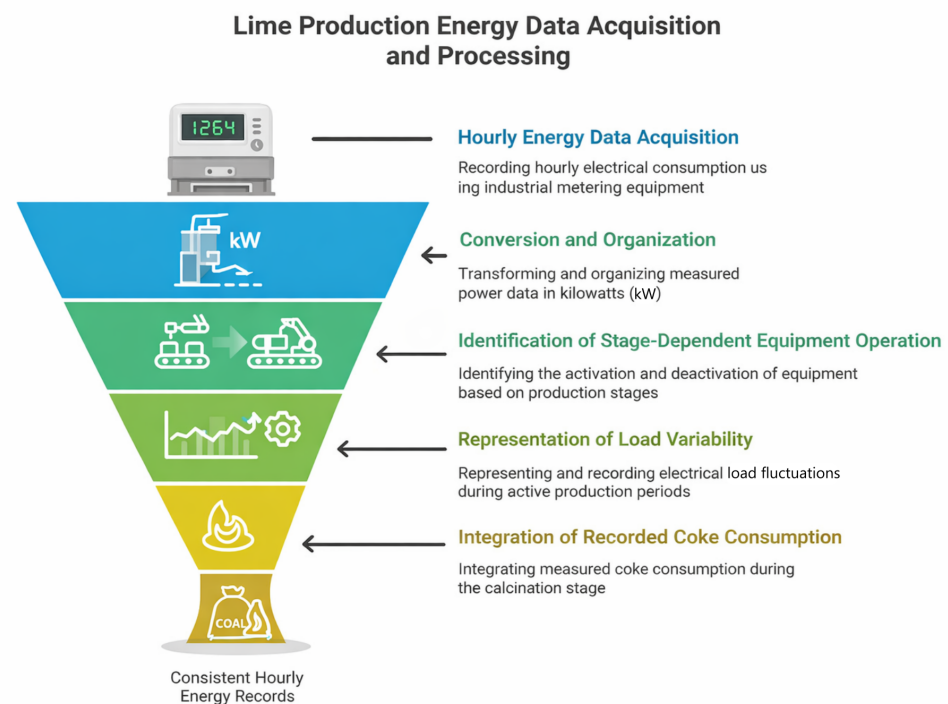


Figure 1. Schematic representation of the hourly energy data acquisition and processing workflow for the quicklime production plant.

The dataset was composed of hourly records, with each including information such as date, time, day of the week, active stage, coke consumption, and total electrical energy consumption. Below, Figure 2 details a view of one day from the dataset:

index	fecha	hora	dia_semana	etapa	coke_kg	hopper	crusher	belt1	sieve	belt2	elevator	belt3	belt4	furnace_out	mill	draft_system	consumo_total_kW	semana	time_id	planta
50	2024-01-03 02:00:00	2	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	50 SUMININCO
51	2024-01-03 03:00:00	3	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	51 SUMININCO
52	2024-01-03 04:00:00	4	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	52 SUMININCO
53	2024-01-03 05:00:00	5	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	53 SUMININCO
54	2024-01-03 06:00:00	6	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	54 SUMININCO
55	2024-01-03 07:00:00	7	descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	55 SUMININCO
56	2024-01-03 08:00:00	8	2 post_lamizaje		0.0	0.0	0.0	0.0	0.0	0.0	1.34	1.8	2.99	0.94	0.0	0.0	6.17	1	56 SUMININCO	
57	2024-01-03 09:00:00	9	2 post_lamizaje		0.0	0.0	0.0	0.0	0.0	1.2	1.79	0.9	1.16	0.0	0.0	0.0	5.05	1	57 SUMININCO	
58	2024-01-03 10:00:00	10	2 post_lamizaje		0.0	0.0	0.0	0.0	0.0	2.09	1.52	2.1	0.71	0.0	0.0	0.0	6.42	1	58 SUMININCO	
59	2024-01-03 11:00:00	11	2 post_lamizaje		0.0	0.0	0.0	0.0	0.0	1.24	1.74	2.22	1.38	0.0	0.0	0.0	6.58	1	59 SUMININCO	
60	2024-01-03 12:00:00	12	2 horno_molienda		213.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.62	4.88	16.3	21.8	1	60 SUMININCO	
61	2024-01-03 13:00:00	13	2 horno_molienda		265.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	3.92	11.58	16.4	1	61 SUMININCO	
62	2024-01-03 14:00:00	14	2 horno_molienda		208.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.95	7.72	19.49	28.16	1	62 SUMININCO	
63	2024-01-03 15:00:00	15	2 horno_molienda		222.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.96	3.92	10.0	14.88	1	63 SUMININCO	
64	2024-01-03 16:00:00	16	2 horno_molienda		393.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.91	6.02	18.08	27.01	1	64 SUMININCO	
65	2024-01-03 17:00:00	17	2 horno_molienda		308.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.19	7.67	12.55	21.41	1	65 SUMININCO	
66	2024-01-03 18:00:00	18	2 descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	66 SUMININCO
67	2024-01-03 19:00:00	19	2 descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	67 SUMININCO
68	2024-01-03 20:00:00	20	2 descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	68 SUMININCO
69	2024-01-03 21:00:00	21	2 descanso		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	69 SUMININCO

Figure 2. Portion example of the Dataset in one day.

The following Figure 3 shows the behavior of the variable “total consumption (kW)” for one week from the dataset. Data is available for the first five days (Monday to Friday), while weekends correspond to non-operational periods with a near-zero recorded load.

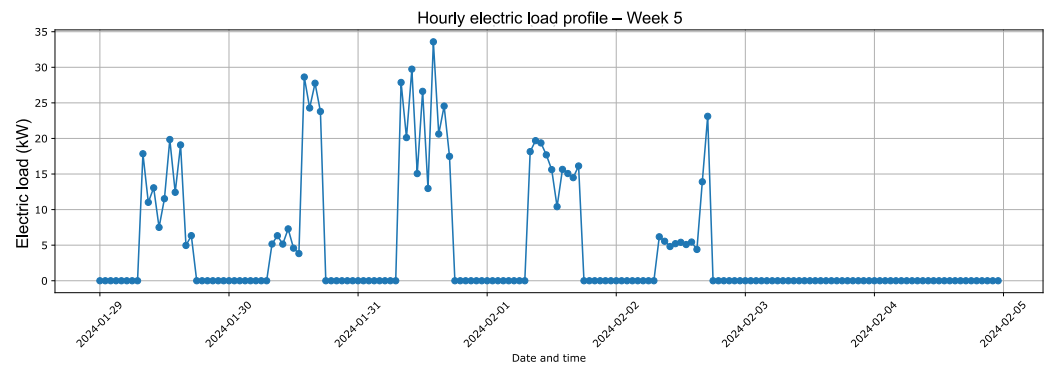


Figure 3. Hourly average electric load measured at the plant level in one week. Near-zero values correspond to non-operational periods, rather than missing data.

2.2. Data Preprocessing and Treatment of Non-Operational Periods

All measurements were first aligned to a regular hourly time index using resampling. A distinction was made between missing records and true low-demand periods.

Non-operational periods, including nights and weekends, correspond to measured near-zero power demand, rather than missing data. These values were retained as valid observations to preserve the operational structure of the plant and avoid introducing artificial bias.

When gaps in the time series were detected within operational hours, they were treated as missing values and handled through removal/interpolation, depending on the length of the gap.

This preprocessing strategy ensures that zero values represent actual plant inactivity, rather than data absence, which is essential for preserving the statistical properties of the load profile and for training the forecasting models. Algorithm 1 shows the steps for data preparation and feature construction within the developed industrial load forecasting methodology.

Algorithm 1 Data preparation and feature construction for industrial load forecasting.

Input: Raw dataset $\mathcal{D}$ with timestamp t , power demand P_t (kW), production stage S_t , coke consumption C_t

Output: Target series Y , covariates X_{past} , X_{future} , training and validation sets

- 1: Resample $\mathcal{D}$ to regular hourly time index
 - 2: Identify missing timestamps and handle according to operational log
 - 3: For non-operational hours (nights/weekends), keep measured near-zero values
 - 4: Encode production stage S_t using one-hot representation
 - 5: Construct past covariates $X_{past} = \{S_t, C_t\}$
 - 6: Construct known-future covariates $X_{future} = \{\text{hour-of-day, day-of-week}\}$
 - 7: Normalize continuous variables using scaler fitted on training data
 - 8: Define input window length $L = 168$ h
 - 9: Define forecast horizon $H = 24$ h
 - 10: Split dataset into training and validation sets chronologically
 - 11: Return prepared time series objects for model training
-

2.3. Temporal Fusion Transformer (TFT) Model

A Temporal Fusion Transformer (TFT) model was developed as a deep learning architecture designed for multivariate and multi-horizon time series forecasting. This means it can predict multiple future time steps while handling several variables simultaneously. TFT differs from traditional models by providing not only point forecasts but also prediction intervals, which help estimate uncertainty.

Model Inputs

TFT uses three types of inputs:

1. Past target values (y) within an observation window.
2. Time-dependent exogenous variables (e.g., weather, day of the week), classified as:
 - Unknown (z): Available only up to the present.
 - Known (x): Known even for future steps (e.g., holidays).
3. Static covariates (s): Metadata such as geographic location or facility type that do not change over time.

2.4. Main Components of the TFT

First, the model incorporates gating mechanisms that dynamically control the flow of information within the network. By means of gated residual blocks, the TFT can skip unnecessary transformations under certain conditions, thus adapting its depth and complexity to the characteristics of the dataset. This flexibility allows the model to behave almost linearly in simple or noisy scenarios while exploiting its nonlinear representational capacity when more complex dependencies are present.

A second innovation lies in the use of variable selection networks, which identify the most relevant inputs for prediction at each time step. This mechanism applies both to static covariates and to time-dependent variables, distinguishing between those observed in the past and those known a priori for the future. In this way, the model improves efficiency by focusing on the most informative signals, while at the same time enhancing interpretability by explicitly revealing which variables drive the forecasting process.

The third essential component is the set of static covariate encoders, designed to integrate information that does not change over time, such as the geographic location or the intrinsic nature of the entity under study. These encoders generate context vectors that condition various parts of the network, ensuring that static features appropriately influence temporal dynamics and contribute to the construction of richer internal representations.

Fourth, TFT relies on temporal processing mechanisms to capture both local and long-term dependencies. A sequence-to-sequence architecture based on recurrent networks is employed to extract short-range temporal patterns, while an interpretable multi-head self-attention block captures broader relationships across time. The integration of these two components enables the model to simultaneously identify short-term fluctuations and long-term seasonal or structural effects, thereby addressing the inherent complexity of multivariate time series.

Finally, TFT produces prediction intervals through quantile forecasting, which extends its practical applicability by providing not only point estimates but also probabilistic ranges for future values. This probabilistic perspective is particularly valuable in real-world scenarios, as it allows practitioners to quantify uncertainty and make more robust decisions in environments characterized by high variability and risk. The architecture of the Temporal Fusion Transformer (TFT) is shown in Figure 4.

2.4.1. Quantile Regression

TFT does not produce a single point estimate but generates multiple quantiles (e.g., 10%, 50%, 90%) to estimate a likely range of future values. This is trained using the pinball loss function, which is effective for uncertainty estimation.

2.4.2. Interpretability

TFT is designed with interpretability in mind, achieved through:

- A variable selection mechanism that assigns global weights.

- An attention mechanism tailored to trace key inputs. This allows the model to not only predict but also explain what information it used and why.

2.4.3. Implementation

In this work, the TFT model was implemented using the **Darts** library (version 0.30.0) in Python (version 3.10):

- **Darts:** Based on PyTorch Lightning (version 2.1.2), it offers user-friendly tools to apply TFT models in production environments.

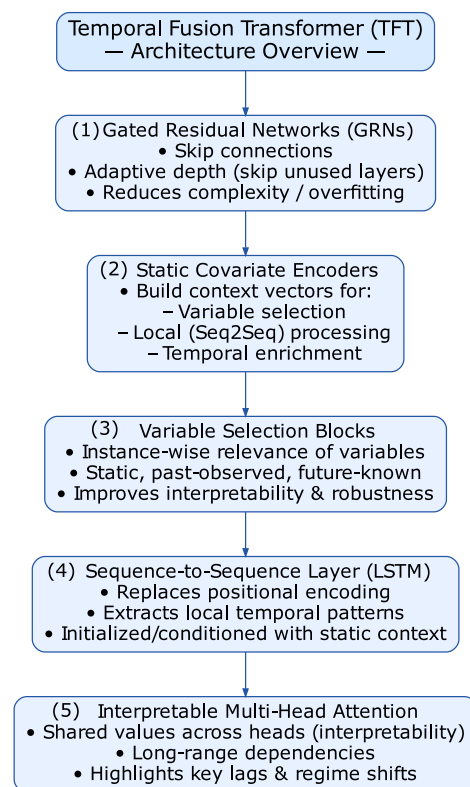


Figure 4. Architecture of the Temporal Fusion Transformer (TFT). The model is composed of five main components: (1) Gated Residual Networks (GRNs), which provide adaptive depth and reduce unnecessary complexity; (2) Static Covariate Encoders, which transform static information into context vectors for variable selection, local sequence processing, and the enrichment of temporal representations; (3) Variable Selection Blocks, which learn the relative importance of each input variable and enhance interpretability; (4) a Sequence-to-Sequence Layer, which leverages LSTM-based processing enriched with static context vectors to replace classical positional encoding; and (5) an Interpretable Multi-Head Attention mechanism, which enables the identification of key past time steps and the detection of significant temporal pattern shifts.

2.5. Energy Deviation Metric

Although the target variable corresponds to the hourly average power demand, P_t , expressed in kW, cumulative energy over a forecasting horizon is obtained by temporal integration:

$$E = \sum_{t=1}^H P_t \Delta t$$

where $\Delta t = 1$ h, yielding energy in kWh.

The predicted cumulative energy is defined as:

$$\hat{E} = \sum_{t=1}^H \hat{P}_t \Delta t$$

The energy deviation metric reported in this study is computed as:

$$EE(\%) = \frac{|\hat{E} - E|}{E} \times 100$$

This metric evaluates the ability of the forecasting model to preserve total energy over the operational horizon.

2.6. TFT Training and Forecasting Pipeline

The implementation of the forecasting framework begins with the initialization of the computational environment and the required modeling tools. Core libraries for data handling and visualization (pandas, numpy, and matplotlib) are employed, together with the main components of the Darts framework, including TimeSeries for structured time series representation, TFTModel for model construction, and QuantileRegression for probabilistic learning.

Following this initialization stage, the forecasting procedure is organized into six sequential steps covering data preparation, covariate construction, dataset partitioning, model configuration and training, multi-horizon inference, and result export. These steps define the operational workflow adopted in this study and are illustrated in Figure 5. The detailed description of each step is provided below.

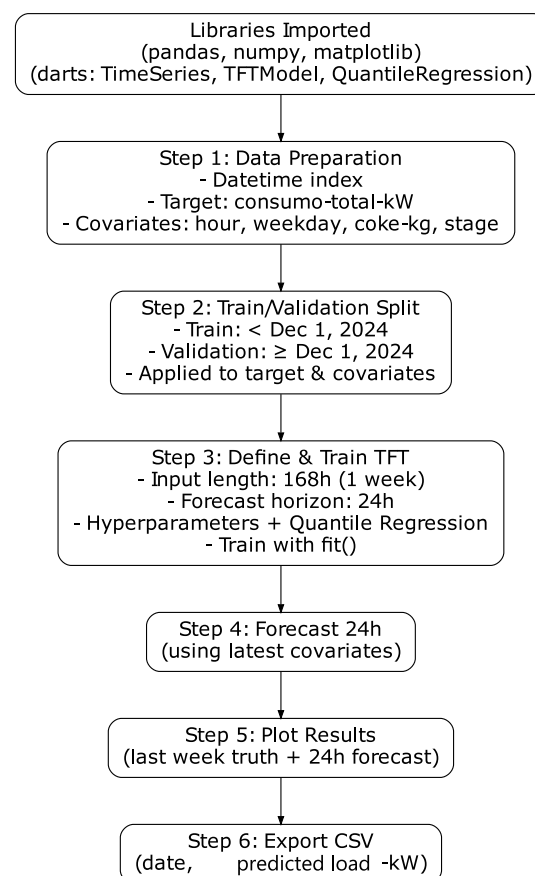


Figure 5. Simplified workflow of the Temporal Fusion Transformer (TFT) implementation. The pipeline starts with the import of required libraries (pandas, numpy, matplotlib, and darts components). Step 1 prepares the dataset by creating the target series and multivariate covariates. Step 2 splits the data into training and validation sets. Step 3 defines and trains the TFT model with appropriate hyperparameters and probabilistic quantile regression. Step 4 performs a 24 h forecast, followed by Step 5, when actual and predicted values are visualized. Step 6 exports the forecast results to a CSV file for further analysis or integration.

In Step 1, the dataset is prepared. The date column is converted to datetime format and set as the DataFrame index. The target series is built from the total consumption (kW) column. Temporal covariates are then created from columns such as hour, day of the week, coke usage (coke-kg), and production stage, which is encoded using a one-hot categorical representation to preserve stage independence and avoid ordinal bias. These covariates are transformed into a multivariate TimeSeries structure to be used as additional inputs for the model.

In Step 2, the dataset is split into training and validation sets. In this case, data prior to 1 December 2024 is used to train the model, while data after that date is reserved for validation and prediction. This split is applied to both the target series and the covariates.

In Step 3, the Temporal Fusion Transformer (TFT) model is defined. This model combines LSTM networks with Transformer-style attention mechanisms and is specially designed to capture temporal and seasonal patterns in multivariate time series. Its hyperparameters are configured, including input length (168 h, equivalent to one week), forecast horizon (24 h), number of hidden layers, batch size, and number of training epochs. Quantile Regression is also specified as the probabilistic estimation technique, and the model is trained using fit().

In Step 4, the trained model is used to forecast energy consumption for the next 24 h, using the most recent covariates (cov-val).

In Step 5, a plot is generated, showing the actual training data from the last week alongside the 24 h prediction and allowing a visualization of the model's ability to follow the consumption pattern.

Finally, in Step 6, the prediction is exported to a CSV file with columns for the date and the predicted consumption-(kW). This is useful for further analysis, reporting, or integration into energy monitoring systems. Algorithm 2 presents the methodological steps for multi-horizon TFT forecasting and evaluation.

Algorithm 2 Multi-horizon TFT forecasting and evaluation.

Input: Target series Y , covariates X_{past} , X_{future}

Output: Point forecasts $\hat{P}_t$, quantile forecasts $q_{0.1}, q_{0.5}, q_{0.9}$, evaluation metrics

- 1: Initialize TFT model with hidden size d , attention heads h
- 2: Set quantile regression loss (pinball loss) for $\tau \in \{0.1, 0.5, 0.9\}$
- 3: Train model using training set with input window L and horizon H
- 4: Apply early stopping based on validation loss
- 5: For each day in evaluation week do
- 6: Observe last L hours of Y
- 7: Predict next H hours $\hat{P}_{t:t+H}$
- 8: Store predicted median and quantiles
- 9: End for
- 10: Compute MAE and RMSE over evaluation horizon
- 11: Compute cumulative energy:

$$E = \sum P_t \Delta t, \quad \hat{E} = \sum \hat{P}_t \Delta t$$

- 12: Compute Energy Error:

$$EE = \frac{|\hat{E} - E|}{E} \times 100$$

- 13: Compute Prediction Interval Coverage Probability (PICP)
 - 14: Return forecasts and performance metrics
-

2.7. Model Configuration and Training Procedure

The Temporal Fusion Transformer (TFT) model was implemented using the Darts framework (Python, PyTorch backend). The forecasting setup employed an input window

length of 168 h (one week of historical context) and an output horizon of 24 h (direct multi-step prediction).

The main hyperparameters were configured as follows: hidden size = 32, number of LSTM layers = 1, number of attention heads = 4, dropout rate = 0.1, batch size = 64, learning rate = 10^{-3} , and number of training epochs = 50. Relative positional encoding was enabled through the `add_relative_index` option.

The model was trained using quantile regression with quantiles $\{0.1, 0.5, 0.9\}$ to obtain probabilistic forecasts. Optimization was performed using the Adam optimizer with a fixed learning rate and a deterministic random seed (42) to ensure reproducibility.

Chronological splitting was applied for training and validation. For each forecast date, the training set included all observations up to the forecast start minus the input window, while the validation horizon consisted of the subsequent 24 h. Rolling weekly backtesting was performed to evaluate generalization performance across multiple operational cycles. The main configuration parameters of the TFT model are summarized in Table 1.

The architectural framework of the Temporal Fusion Transformer (TFT) utilized in this study is illustrated in Figure 6. The model employs a multi-horizon forecasting approach that effectively integrates static covariates, known future inputs, and observed historical data. Key components include Gated Residual Networks (GRN), which enable the suppression of irrelevant inputs through sophisticated variable selection blocks, and a Temporal Self-Attention mechanism. This attention layer is critical for identifying long-term dependencies and seasonal patterns within the quicklime production energy cycles. By utilizing recurrent layers for local processing and self-attention for global dependencies, the TFT architecture ensures a robust multi-quantile probabilistic output, allowing for a comprehensive quantification of uncertainty across the forecasting horizon.

```
INFO:pytorch_lightning.utilities.rank_zero:GPU available: True (cuda), used: True
INFO:pytorch_lightning.utilities.rank_zero:TPU available: False, using: 0 TPU cores
INFO:pytorch_lightning.utilities.rank_zero:HPU available: False, using: 0 HPUS
INFO:pytorch_lightning.accelerators.cuda:LOCAL_RANK: 0 - CUDA_VISIBLE_DEVICES: [0]
INFO:pytorch_lightning.callbacks.model_summary:
```

	Name	Type	Params	Mode
0	train_metrics	MetricCollection	0	train
1	val_metrics	MetricCollection	0	train
2	input_embeddings	_MultiEmbedding	0	train
3	static_covariates_vsn	_VariableSelectionNetwork	0	train
4	encoder_vsn	_VariableSelectionNetwork	6.0 K	train
5	decoder_vsn	_VariableSelectionNetwork	896	train
6	static_context_grn	_GatedResidualNetwork	4.3 K	train
7	static_context_hidden_encoder_grn	_GatedResidualNetwork	4.3 K	train
8	static_context_cell_encoder_grn	_GatedResidualNetwork	4.3 K	train
9	static_context_enrichment	_GatedResidualNetwork	4.3 K	train
10	lstm_encoder	LSTM	8.4 K	train
11	lstm_decoder	LSTM	8.4 K	train
12	post_lstm_gan	_GateAddNorm	2.2 K	train
13	static_enrichment_grn	_GatedResidualNetwork	5.3 K	train
14	multihead_attn	_InterpretableMultiHeadAttention	2.6 K	train
15	post_attn_gan	_GateAddNorm	2.2 K	train
16	feed_forward_block	_GatedResidualNetwork	4.3 K	train
17	pre_output_gan	_GateAddNorm	2.2 K	train
18	output_layer	Linear	561	train

```
-----
60.3 K   Trainable params
0       Non-trainable params
60.3 K   Total params
0.241   Total estimated model params size (MB)
232     Modules in train mode
0       Modules in eval mode
Epoch 29: 100%
```

Figure 6. Model summary of the Temporal Fusion Transformer (TFT) implemented in PyTorch Lightning, illustrating the encoder–decoder structure, attention mechanisms, and trainable parameters of each module.

Table 1. TFT model hyperparameters.

Parameter	Value
Input window length	168 h
Forecast horizon	24 h
Hidden size	32
LSTM layers	1
Attention heads	4
Dropout	0.1
Batch size	64
Epochs	50
Learning rate	0.001
Quantiles	0.1, 0.5, 0.9
Random seed	42

2.8. Forecasting Strategy and Rolling Evaluation

The TFT model is configured to produce direct multi-horizon predictions with an output length of 24 h per inference step. Rather than recursively feeding predictions into subsequent time steps, the model generates each 24 h forecast in a single forward pass using observed historical data.

Weekly prediction curves are constructed using a rolling forecasting scheme. For each day in the evaluation period, the model receives a fixed input window of 168 h (one week of past observations) and produces a direct 24 h forecast. The input window is then shifted forward by 24 h, and the procedure is repeated sequentially across the week.

This strategy results in a series of overlapping historical windows but non-overlapping forecast horizons, ensuring independence between daily predictions while preserving temporal continuity in the evaluation.

Teacher forcing is applied during model training as part of the sequence-to-sequence learning process. However, during inference and evaluation, forecasts are generated without teacher forcing, relying exclusively on observed past inputs and known covariates.

2.9. Baseline Models and Comparative Experimental Setup

To assess the performance of the proposed TFT framework, a comparative study was conducted using three deep learning baselines: LSTM, GRU, and N-BEATS.

All models were trained and evaluated under a consistent experimental protocol. Each model received the same historical input window of 168 h and generated forecasts over a one-week horizon (168 h). Chronological splitting was applied, where training data included all observations prior to the forecast start date, and evaluation was performed on the subsequent period.

Scaling and preprocessing procedures were identical across models to ensure comparability. Training was performed for 30 epochs using the Adam optimizer, with a batch size of 64 and a learning rate of 10^{-3} . A fixed random seed (42) was used to guarantee reproducibility. Table 2 shows the hyperparameters and input configuration for baseline models.

Model-specific configurations were defined as follows:

- TFT was implemented as a probabilistic multi-horizon forecasting model using quantile regression and past covariates.
- LSTM and GRU were implemented as autoregressive sequence models using known future covariates.
- N-BEATS was implemented as a univariate deep forecasting model without exogenous covariates.

Table 2. Hyperparameters and input configuration for baseline models.

Model	Input Window	Forecast Horizon	Covariates	Epochs
TFT	168 h	168 h	Past covariates	30
LSTM	168 h	168 h	Future covariates	30
GRU	168 h	168 h	Future covariates	30
N-BEATS	168 h	168 h	None	30

2.10. Probabilistic Forecasting and Uncertainty Estimation

Prediction intervals were derived from the probabilistic predictive distribution learned through quantile regression. In practice, the Darts probabilistic prediction interface was used with $S = 200$ samples drawn from the fitted likelihood to estimate empirical prediction intervals (P10–P90) and median forecasts. This sampling procedure serves only to approximate the predictive distribution implied by the quantile-regression model.

Let $\{\hat{y}_{t+h}^{(s)}\}_{s=1}^S$ denote $S = 200$ samples drawn from the predictive distribution at horizon h . The P10–P90 interval is computed as:

$$\left[Q_{0.10}\left(\{\hat{y}_{t+h}^{(s)}\}\right), Q_{0.90}\left(\{\hat{y}_{t+h}^{(s)}\}\right) \right],$$

and the median forecast corresponds to $Q_{0.50}(\cdot)$.

This formulation provides reproducible uncertainty quantification and allows consistent comparison across forecasting models.

To further evaluate the calibration and sharpness of the probabilistic forecasts, two additional metrics were computed for the P10–P90 prediction intervals.

Prediction Interval Coverage Probability (PICP) measures the empirical proportion of observations that fall within the prediction interval:

$$PICP = \frac{1}{N} \sum_{t=1}^N \mathbb{I}(y_t \in [\hat{y}_{0.10,t}, \hat{y}_{0.90,t}])$$

Mean Prediction Interval Width (MPIW) quantifies the average width of the prediction intervals:

$$MPIW = \frac{1}{N} \sum_{t=1}^N (\hat{y}_{0.90,t} - \hat{y}_{0.10,t})$$

These metrics allow evaluating both the reliability (coverage) and sharpness (interval width) of the probabilistic forecasts.

3. Results

3.1. Weekly Energy Consumption Forecasting

Figure 7 illustrates the one-week energy consumption forecasting performance of the Temporal Fusion Transformer (TFT) model by comparing the predicted median profile with the corresponding measured data. The predicted trajectory closely follows the actual consumption pattern, accurately capturing the daily activation and shutdown of production stages, as well as the magnitude and timing of peak demand periods. Minor deviations are observed during rapid load transitions, which are primarily associated with abrupt changes in motor operation and process dynamics. Overall, the strong agreement between the predicted and observed profiles demonstrates the model's ability to learn the temporal structure and operational constraints of the quicklime production process, confirming its suitability for short-term industrial energy forecasting applications.

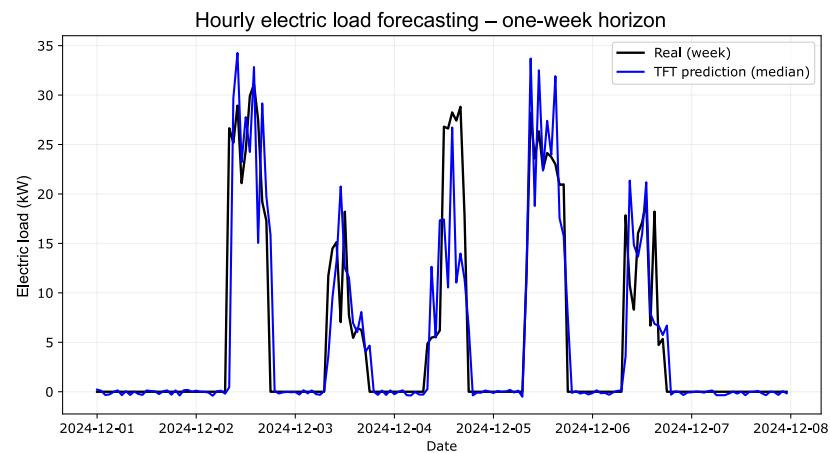


Figure 7. Weekly energy consumption predictions, where the black line shows actual data, and the blue line indicates forecasts generated by the TFT model, 30 epochs.

3.2. Forecast Error Metrics

The forecasting performance of the proposed Temporal Fusion Transformer (TFT) model was evaluated not only using point-wise accuracy metrics but also in terms of cumulative energy deviation, as summarized in Table 3. When assessing the prediction over a full one-week horizon, the model achieved a global energy error of 4.78%, indicating a very accurate preservation of the overall energy balance. This result confirms the model's ability to correctly identify non-operational periods, such as nights and weekends, while maintaining consistency in the aggregated energy estimation.

When the evaluation was restricted to active production hours only, corresponding to working days between 08:00 and 18:00, the energy error increased slightly to 5.25%. This behavior is expected due to the higher variability of the industrial process during operational stages, including partial-load motor operation and abrupt transitions between production phases. Nevertheless, the obtained value remains well below typical error margins reported in industrial energy forecasting studies, demonstrating that the proposed model reliably captures the cumulative energy demand during active operation.

The relatively small difference between global and active-hour energy errors highlights the robustness of the TFT model in balancing structural accuracy and operational variability. While short-term peak values may be mildly smoothed, the model effectively preserves the total energy consumption over the evaluation period. This characteristic is particularly advantageous for planning-oriented applications, such as energy procurement, tariff optimization, and production scheduling, for which cumulative energy accuracy is more critical than instantaneous peak matching.

Overall, the obtained results demonstrate that increasing the number of training epochs significantly enhances the model's capability to learn complex load patterns and stage-dependent consumption dynamics. With both global and active energy errors remaining below 6%, the proposed TFT framework provides a reliable and practically relevant solution for short-term energy forecasting in quicklime production processes.

It is important to note that the reported energy error metrics correspond to two different evaluation weeks. The detailed TFT performance presented in Table 3 was computed for the week starting on 1 December 2024, which was selected as the primary evaluation period for analyzing cumulative energy accuracy and uncertainty behavior. In contrast, the comparative analysis across forecasting models reported in Table 4 was conducted using a separate evaluation window starting on 15 December 2024, in order to assess model robustness and generalization under different operational conditions.

Table 3. Forecasting performance of the Temporal Fusion Transformer (TFT) model for a one-week horizon after extended training. Error metrics are reported for the full evaluation period (Global) and restricted to active production hours (Active).

Evaluation Scope	MAE (kW)	RMSE (kW)	Energy Error (%)
Global (all hours)	3.38	6.56	4.78
Active hours only	9.02	11.45	5.25

Table 4. Comparison of forecasting models for one-week energy prediction after 30 training epochs. Point-wise accuracy metrics (MAE and RMSE) are reported, together with cumulative energy error for the full horizon (Global) and active production hours only (Active, 08:00–18:00 on weekdays).

Model	MAE (kW)	RMSE (kW)	Energy Error Global (%)	Energy Error Active (%)
LSTM	1.314	2.835	2.152	2.052
GRU	1.342	2.993	1.988	2.738
N-BEATS	1.916	3.132	1.453	3.509
TFT (Quantile Regression)	2.261	4.583	1.900	3.876

3.3. Comparison of Deep Learning Models for Industrial Energy Forecasting

Table 4 summarizes the comparative performance of the proposed TFT model against three baselines (LSTM, GRU, and N-BEATS) over a one-week horizon after training all models for 30 epochs. For transparency, all models were evaluated on the same shared evaluation week to ensure a consistent comparison of forecasting performance. In terms of point-wise accuracy, the recurrent models achieved the best results, with LSTM and GRU yielding the lowest MAE (1.31–1.34 kW) and RMSE (2.84–2.99 kW), confirming the effectiveness of gated and memory-based recurrent dynamics for tracking rapid load variations and peak magnitudes during operation.

When shifting the evaluation criterion to cumulative energy deviation, all models exhibited low global energy errors (below 2.2%), indicating an overall strong preservation of the weekly energy balance. The N-BEATS model achieved the lowest global energy error (1.45%), while TFT remained highly competitive (1.90%) despite its higher point-wise errors. This result suggests that the transformer-based architecture can preserve aggregated demand even when short-term peak values are mildly smoothed.

A more stringent assessment restricted to active production hours (08:00–18:00, weekdays) reveals increased sensitivity to operational variability. In this regime, LSTM and GRU retained the best cumulative accuracy (2.05% and 2.74%, respectively), whereas TFT showed a higher active-hour energy error (3.88%). This behavior is consistent with the model’s tendency to regularize extreme values and to rely on covariate-driven structure, rather than purely local amplitude matching.

Overall, the results highlight a clear trade-off: recurrent baselines provide superior instantaneous tracking, whereas the TFT framework offers competitive energy consistency, together with probabilistic forecasting and interpretability via quantile regression and attention mechanisms. These additional capabilities are particularly relevant for decision-support scenarios such as tariff-aware planning, energy procurement, and risk-aware operational scheduling, where uncertainty quantification and transparency are critical. Evaluating the models on two non-overlapping weeks allows for the disentangling of model-specific behavior from week-dependent operational variability, strengthening the validity of the comparative conclusions.

3.4. Probabilistic Weekly Load Forecasting

Figure 8 illustrates the probabilistic weekly electric load forecasting results obtained with the TFT model for the period of 1–7 December 2024. The median prediction follows the daily operational cycles of the plant, accurately identifying start–stop patterns associated with working hours and inactivity during nights and weekends.

Prediction intervals (P10–P90), computed as described in Section 2.10, exhibit a systematic widening during the onset of production stages and peak-demand periods. This behavior reflects increased operational variability and the nonlinear dynamics of a stage-driven industrial load, indicating that the model appropriately captures uncertainty under changing production conditions.

A quantitative evaluation shows that 86% of the observed hourly load values fall within the P10–P90 interval, suggesting a good probabilistic calibration of the forecasts. While some deviations are observed around sharp peaks, the model consistently captures the timing, duration, and intensity of active production periods, demonstrating robustness in an industrial environment characterized by repetitive yet non-stationary load patterns.

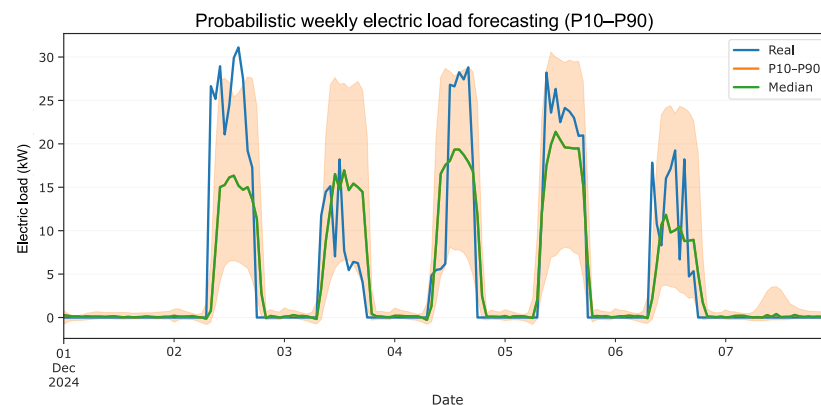


Figure 8. Probabilistic weekly electric load forecasting (P10–P90) for the period of 1–7 December 2024. The blue curve represents the observed hourly electric load, the green curve corresponds to the TFT median forecast, and the shaded orange band denotes the prediction interval derived from quantile regression, reflecting operational uncertainty during the forecasting horizon.

The second evaluation shown in Figure 9, for the subsequent week of 15–22 December 2024, validates the performance of the previously trained and stored TFT model when reloaded for inference. The results confirm that the model retains its forecasting capability without retraining, which is a critical feature for real-world industrial applications where models must be deployed efficiently. Similar to the previous week, the forecasts align with the actual load curves, capturing the cyclical nature of production activities across weekdays while maintaining near-zero predictions during weekends. The quantile-based prediction bands again highlight areas of higher uncertainty at peak demand periods, suggesting that, while the model can reproduce the general consumption structure, extreme values remain more challenging to estimate with precision. This outcome reinforces the importance of probabilistic forecasting in industrial energy management, as it not only delivers point estimates but also quantifies the reliability of predictions across operational contexts.

To further evaluate the calibration of the probabilistic forecasts, the Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) were computed for the P10–P90 interval. The obtained values are reported in Table 5. Given the nominal coverage of 80%, the empirical coverage indicates well-calibrated prediction intervals with a moderate uncertainty width, providing reliable probabilistic information for operational decision-making.

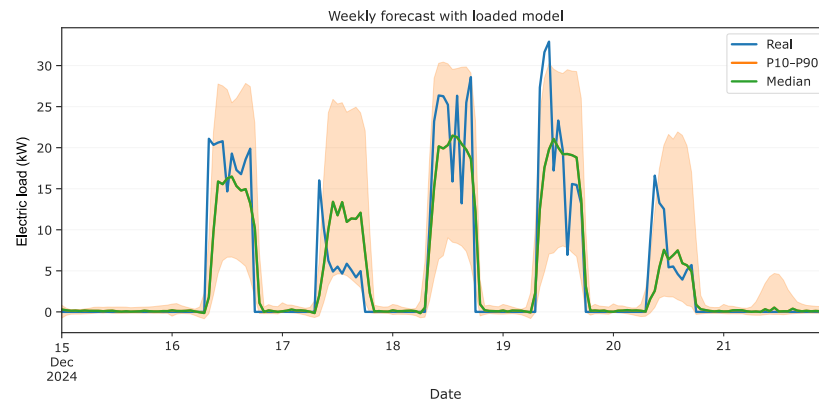


Figure 9. Probabilistic weekly electric load forecasting using the trained Temporal Fusion Transformer (TFT) model (P10–P90) for the period 15–22 December 2024. The blue curve represents the observed hourly electric load, the green curve corresponds to the TFT median forecast, and the shaded orange band denotes the prediction interval derived from quantile regression, capturing operational uncertainty across the forecasting horizon.

Table 5. Uncertainty calibration metrics for the P10–P90 prediction interval.

Metric	Value
Prediction Interval Coverage Probability (PICP)	0.86
Mean Prediction Interval Width (MPIW)	11.2 kW
Nominal Coverage (P10–P90)	0.80

3.5. Residual Analysis of the Forecasting Model

The statistical distribution of the forecasting residuals provides valuable insight into the predictive performance and robustness of the proposed model. As illustrated in Figure 10, the residuals are predominantly concentrated around zero, indicating a high level of accuracy and an overall unbiased behavior in the majority of the predictions. The narrow interquartile range reflects a low dispersion of errors under normal operating conditions, while the mean residual, slightly shifted towards positive values, suggests a marginal tendency of the model to overestimate the electrical demand. Nevertheless, the presence of isolated outliers with larger positive and negative deviations highlights the occurrence of atypical operating scenarios or transient load fluctuations that are not fully captured by the model. Despite these extreme values, the overall residual structure confirms the model's capability to reliably represent the underlying consumption dynamics, supporting its suitability for energy forecasting applications in industrial environments.

3.6. Model Explainability and Attention Analysis

The explainability analysis provided by the Temporal Fusion Transformer (TFT) reveals the internal mechanisms driving the forecasting process. Variable selection results (Figure 11) indicate that the production stage is the most influential encoder variable, followed by day-of-week and hour-of-day features, while autoregressive dependence on past load values plays a secondary role. This confirms that the electrical demand of the quicklime process is primarily stage-driven, rather than purely autoregressive.

An attention analysis further shows that the model relies on recurrent temporal patterns, rather than immediate past observations. The mean attention distribution (Figure 12) highlights strong contributions from historical windows associated with daily and weekly operational cycles. Horizon-specific attention patterns remain consistent across prediction steps, suggesting stable temporal dependencies, as indicated in Figure 13.

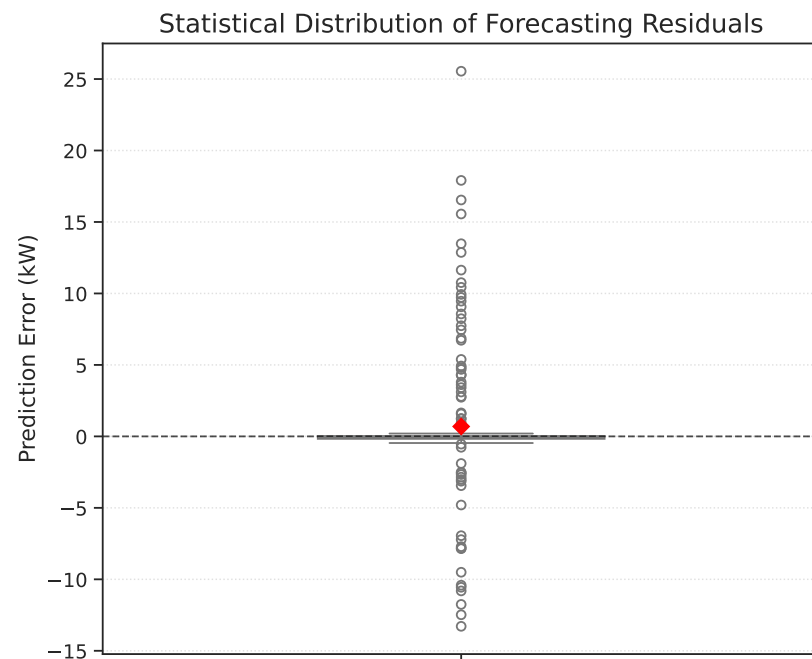


Figure 10. Statistical distribution of forecasting residuals expressed as prediction errors in kilowatts (kW). The concentration of residuals around zero indicates a high overall forecasting accuracy, while the presence of isolated outliers reflects atypical operating conditions or transient load variations not fully captured by the model.

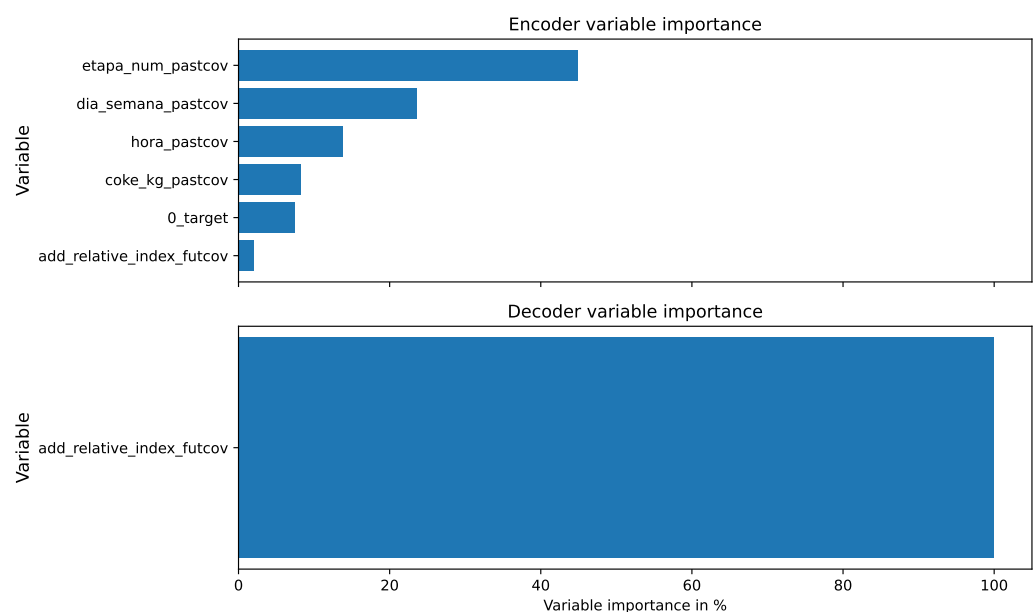


Figure 11. Variable importance derived from the Temporal Fusion Transformer (TFT) variable selection mechanism. The encoder analysis indicates that the production stage is the most influential predictor, followed by day-of-week and hour-of-day features, while coke consumption and autoregressive dependence on past load exhibit lower contributions. The decoder importance is dominated by the relative time index, reflecting the model's reliance on positional temporal information for multi-horizon forecasting.

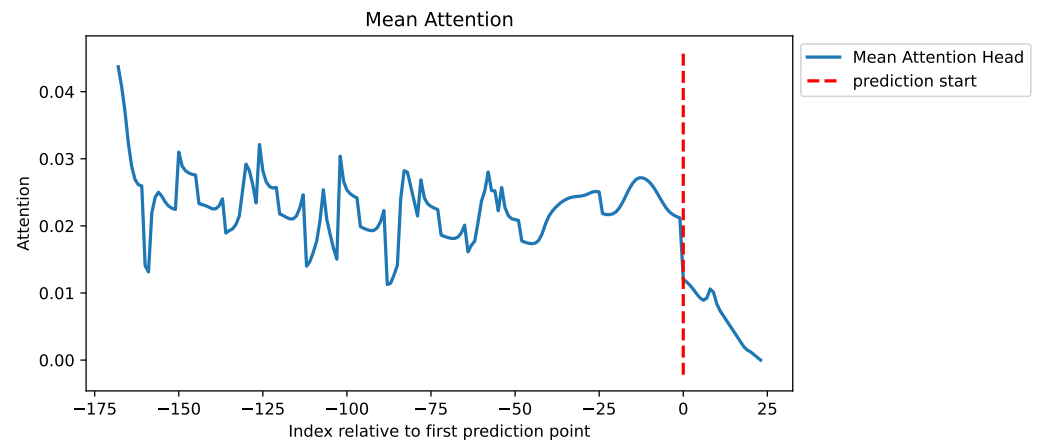


Figure 12. Mean temporal attention weights learned by the Temporal Fusion Transformer (TFT) relative to the first prediction point (red dashed line). Higher attention values are assigned to historical observations preceding the forecast horizon, indicating that the model relies on recurrent daily and weekly patterns, rather than immediate past values to generate predictions.

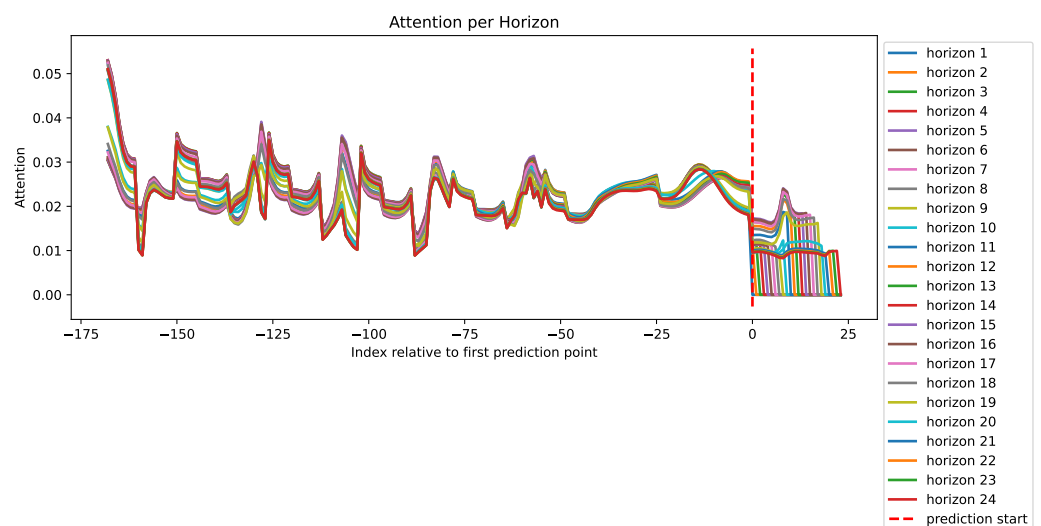


Figure 13. Attention weights for each prediction horizon generated by the Temporal Fusion Transformer (TFT). Each curve represents the temporal attention distribution associated with a specific forecast step (1–24 h ahead), relative to the first prediction point (red dashed line). The consistent patterns across horizons indicate stable temporal dependencies, with the model emphasizing recurring historical periods linked to daily operational cycles, rather than relying solely on the most recent observations.

The attention heatmap (Figure 14) reveals clear vertical structures aligned with repeated operational cycles, indicating that the model leverages analogous time periods from previous days to generate forecasts. This behavior confirms that the TFT effectively captures structured industrial dynamics and supports its interpretability in operational contexts.

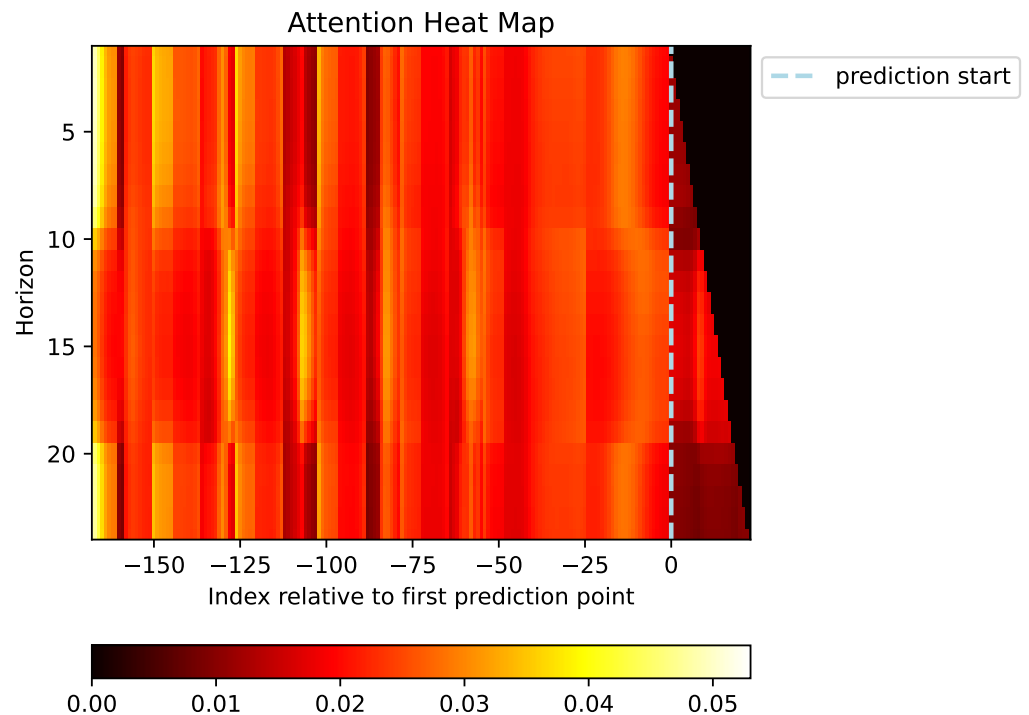


Figure 14. Temporal attention heatmap learned by the Temporal Fusion Transformer (TFT) across prediction horizons. Each row represents a forecast step (horizon), while the horizontal axis shows the historical time index relative to the first prediction point (dashed line). Warmer colors indicate higher attention weights. The vertical banding structure reveals that the model consistently focuses on analogous periods from previous operational cycles, highlighting its ability to capture repetitive stage-driven dynamics in the industrial load profile.

4. Conclusions

This study presented a short-term electric load forecasting framework for a quicklime production process based on real industrial data collected during 2024, integrating operational measurements with a Temporal Fusion Transformer (TFT) deep learning architecture. The dataset captures the actual behavior of the plant under normal operating conditions, including a stage-driven production logic, a plant-level aggregated electrical load, fixed weekday working schedules from 08:00 to 18:00, and non-operational periods during nights and weekends. As a result, the analyzed data reflect realistic industrial characteristics such as intermittent operation, abrupt load transitions, and bounded variability inherent to energy-intensive manufacturing systems.

The proposed TFT model demonstrated a strong capability to learn the temporal structure and operational constraints of the industrial load profiles. The weekly forecasting results show that the model accurately reproduces daily start–stop patterns, peak demand periods, and weekend inactivity. Beyond conventional point-wise accuracy metrics, cumulative energy deviation was explicitly evaluated to assess the preservation of the overall energy balance over planning-relevant horizons. The obtained energy errors of 4.78% over the full week and 5.25% when restricting the evaluation to active production hours confirm that the model reliably captures aggregated plant-level energy demand, even in the presence of non-linear load transitions and process variability.

The probabilistic formulation based on quantile regression further enhances the practical value of the proposed approach. Prediction intervals were derived from the quantile-based predictive distribution learned by the TFT model, with empirical P10–P90 bands estimated through sampling from the fitted likelihood. The uncertainty intervals systematically widen during the onset of production stages, where load variability and nonlinear

dynamics are more pronounced. The empirical coverage of the nominal 80% prediction interval reached 86% (PICP = 0.86), with a mean prediction interval width of 11.2 kW (MPIW), indicating well-calibrated probabilistic forecasts with moderately conservative uncertainty representation under real industrial operating conditions.

A comparative analysis against established deep learning baselines, including LSTM, GRU, and N-BEATS, revealed a meaningful trade-off between instantaneous accuracy and cumulative energy consistency. While recurrent models—particularly GRU—achieved lower MAE and RMSE values and thus better peak-tracking performance, the TFT model exhibited superior global energy consistency across the one-week forecasting horizon. This behavior highlights the suitability of TFT for planning-oriented industrial applications such as energy procurement, tariff optimization, infrastructure sizing, and production scheduling, where preserving cumulative demand is often more critical than the exact reproduction of short-term peaks.

Overall, the results demonstrate that transformer-based architectures, when applied to real industrial datasets enriched with operational covariates, offer a robust and interpretable solution for short-term energy forecasting in stage-driven manufacturing environments. Future work will focus on extending the analysis to longer operational periods, incorporating additional exogenous variables such as electricity prices and ambient conditions, evaluating and comparing probabilistic forecasting performance across multiple model families to assess uncertainty calibration more comprehensively, and coupling the forecasting framework with optimization or reinforcement learning strategies to enable automated, tariff-aware scheduling and energy-efficient operation in industrial production systems.

Author Contributions: Conceptualization, J.X.L.-M. and B.U.S.; methodology, B.U.S.; software, J.X.L.-M., F.P. and C.P.S.C.; validation, J.X.L.-M., F.P. and B.U.S.; resources, B.U.S.; data curation, D.A.T.; writing—original draft preparation, J.X.L.-M. and D.A.T.; writing—review and editing, B.U.S. and C.P.S.C.; visualization, C.P.S.C. and F.P.; supervision, B.U.S.; project administration, C.P.S.C.; funding acquisition, J.X.L.-M. and B.U.S. All authors have read and agreed to the published version of the manuscript.

Funding: Colombian ministry of science innovation and technology-Minciencias with its grant 934 “Convocatoria de estancias posdoctorales orientadas por misiones”. This research is also funding by FONDO FRANCISCO JOSÉ DE CALDAS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code of this research are available in this webpage: https://github.com/jerssonleon/Paper_TFT (accessed on 25 February 2026).

Acknowledgments: The authors thank the Colombian ministry of science innovation and technology-Minciencias with its grant 934 “Convocatoria de estancias posdoctorales orientadas por misiones” with the project entitled: “Inteligencia artificial para mejorar el uso eficiente de la energía en pequeñas empresas de Nobsa Boyacá, caso de estudio: SUMININCO LTDA”. The authors also acknowledge FONDO FRANCISCO JOSÉ DE CALDAS.

Conflicts of Interest: Authors Claudia Patricia Siachoque Celys and Bernardo Umbarila Suarez were employed by the company SUMININCO LTDA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TFT	Temporal Fusion Transformer
RMSE	Root Mean Square Error
kW	Kilowatt
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence

References

- Oguntola, O.; Boakye, K.; Simske, S. Towards leveraging artificial intelligence for sustainable cement manufacturing: A systematic review of AI applications in electrical energy consumption optimization. *Sustainability* **2024**, *16*, 4798. [\[CrossRef\]](#)
- Alcántara, V.; Cadavid, Y.; Sánchez, M.; Uribe, C.; Echeverri-Uribe, C.; Morales, J.; Obando, J.; Amell, A. A study case of energy efficiency, energy profile, and technological gap of combustion systems in the Colombian lime industry. *Appl. Therm. Eng.* **2018**, *128*, 393–401. [\[CrossRef\]](#)
- Sandström, K.; Broström, M.; Eriksson, M. Coal ash and limestone interactions in quicklime production. *Fuel* **2021**, *300*, 120989. [\[CrossRef\]](#)
- Ensafi, Y.; Amin, S.H.; Zhang, G.; Shah, B. Time-series forecasting of seasonal items sales using machine learning—A comparative analysis. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100058. [\[CrossRef\]](#)
- Kontopoulou, V.I.; Panagopoulos, A.D.; Kakkos, I.; Matsopoulos, G.K. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet* **2023**, *15*, 255. [\[CrossRef\]](#)
- Syafrudin, M.; Alfian, G.; Fitriyani, N.L.; Rhee, J. Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing. *Sensors* **2018**, *18*, 2946. [\[CrossRef\]](#)
- Alfares, H.K.; Nazeeruddin, M. Electric load forecasting: Literature survey and classification of methods. *Int. J. Syst. Sci.* **2002**, *33*, 23–34. [\[CrossRef\]](#)
- Howind, S. Load Profile Forecasting of Manufacturing Processes with a Data-Driven Model. Ph.D. Thesis, Technische Universität Wien, Vienna, Austria, 2024.
- Lotfipoor, A.; Patidar, S.; Jenkins, D.P. Deep neural network with empirical mode decomposition and Bayesian optimisation for residential load forecasting. *Expert Syst. Appl.* **2024**, *237*, 121355. [\[CrossRef\]](#)
- Zhao, Z.; Tang, J.; Liu, J.; Ge, G.; Xiong, B.; Li, Y. Short-term microgrid load probability density forecasting method based on k-means-deep learning quantile regression. *Energy Rep.* **2022**, *8*, 1386–1397. [\[CrossRef\]](#)
- Shah, S.A.H.; Ahmed, U.; Bilal, M.; Khan, A.R.; Razzaq, S.; Aziz, I.; Mahmood, A. Improved electric load forecasting using quantile long short-term memory network with dual attention mechanism. *Energy Rep.* **2025**, *13*, 2343–2353. [\[CrossRef\]](#)
- Asiri, M.M.; Aldehim, G.; Alotaibi, F.A.; Alnfai, M.M.; Assiri, M.; Mahmud, A. Short-term load forecasting in smart grids using hybrid deep learning. *IEEE Access* **2024**, *12*, 23504–23513. [\[CrossRef\]](#)
- Meng, X.; Shao, X.; Li, S. Short-Term Load Probability Prediction Based on Integrated Feature Selection and GA-LSTM Quantile Regression. *Int. J. Energy Res.* **2024**, *2024*, 5452005. [\[CrossRef\]](#)
- Wang, Z.; Wang, C.; Cheng, L.; Li, G. An approach for day-ahead interval forecasting of photovoltaic power: A novel DCGAN and LSTM based quantile regression modeling method. *Energy Rep.* **2022**, *8*, 14020–14033. [\[CrossRef\]](#)
- Masood, Z.; Gantassi, R.; Choi, Y. Enhancing short-term electric load forecasting for households using quantile LSTM and clustering-based probabilistic approach. *IEEE Access* **2024**, *12*, 77257–77268. [\[CrossRef\]](#)
- Faustine, A.; Pereira, L. FPSeq2Q: Fully parameterized sequence to quantile regression for net-load forecasting with uncertainty estimates. *IEEE Trans. Smart Grid* **2022**, *13*, 2440–2451. [\[CrossRef\]](#)
- Tziolis, G.; Lopez-Lorente, J.; Baka, M.I.; Koumis, A.; Livera, A.; Theocharides, S.; Makrides, G.; Georghiou, G.E. Direct short-term net load forecasting in renewable integrated microgrids using machine learning: A comparative assessment. *Sustain. Energy Grids Netw.* **2024**, *37*, 101256. [\[CrossRef\]](#)
- Maleki, N.; Lundström, O.; Musaddiq, A.; Jeansson, J.; Olsson, T.; Ahlgren, F. Future energy insights: Time-series and deep learning models for city load forecasting. *Appl. Energy* **2024**, *374*, 124067. [\[CrossRef\]](#)
- Qureshi, M.; Arbab, M.A.; Rehman, S.U. Deep learning-based forecasting of electricity consumption. *Sci. Rep.* **2024**, *14*, 6489. [\[CrossRef\]](#)
- Timur, O.; Üstünel, H.Y. Short-Term Electric Load Forecasting for an Industrial Plant Using Machine Learning-Based Algorithms. *Energies* **2025**, *18*, 1144. [\[CrossRef\]](#)

21. Wicaksono, H.; Trat, M.; Bashyal, A.; Boroukhian, T.; Felder, M.; Ahrens, M.; Bender, J.; Gross, S.; Steiner, D.; July, C.; et al. Artificial-intelligence-enabled dynamic demand response system for maximizing the use of renewable electricity in production processes. *Int. J. Adv. Manuf. Technol.* **2025**, *138*, 247–271. [\[CrossRef\]](#)
22. Aurangzeb, K.; Alhussein, M.; Javaid, K.; Haider, S.I. A pyramid-CNN based deep learning model for power load forecasting of similar-profile energy customers based on clustering. *IEEE Access* **2021**, *9*, 14992–15003. [\[CrossRef\]](#)
23. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [\[CrossRef\]](#)
24. Leon-Medina, J.X.; Camacho, J.; Gutierrez-Osorio, C.; Salomón, J.E.; Rueda, B.; Vargas, W.; Sofrony, J.; Restrepo-Calle, F.; Pedraza, C.; Tibaduiza, D. Temperature prediction using multivariate time series deep learning in the lining of an electric arc furnace for ferronickel production. *Sensors* **2021**, *21*, 6894. [\[CrossRef\]](#)
25. Benidis, K.; Rangapuram, S.S.; Flunkert, V.; Wang, Y.; Maddix, D.; Turkmen, C.; Gasthaus, J.; Bohlke-Schneider, M.; Salinas, D.; Stella, L.; et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Comput. Surv.* **2022**, *55*, 1–36. [\[CrossRef\]](#)
26. Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; Sahoo, D. Unified Training of Universal Time Series Forecasting Transformers. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; Volume 235, pp. 53140–53164.
27. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
28. Dai, T.Y.; Niyogi, D.; Nagy, Z. CityTFT: A temporal fusion transformer-based surrogate model for urban building energy modeling. *Appl. Energy* **2025**, *389*, 125712. [\[CrossRef\]](#)
29. Gokhale, G.; Van Gompel, J.; Claessens, B.; Develder, C. Transfer learning in transformer-based demand forecasting for home energy management system. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 458–462.
30. Maragkos, N.; Refanidis, I. A Comparative Evaluation of Time-Series Forecasting Models for Energy Datasets. *Computers* **2025**, *14*, 246. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.