ORIGINAL RESEARCH **Open access**

A Two-Stage Neural Network Framework for Multivariate Energy Time Series Imputation

Mikko Rantanen[1*], Elina Virtanen[1], Jussi Mäkelä[2]

Abstract

Incomplete multivariate time series frequently emerge across various domains, particularly within energy datasets. Conventional imputation techniques frequently struggle to capture both temporal patterns and cross-dimensional dependencies simultaneously. This study introduces a dual-stage approach utilizing an attention mechanism to reconstruct missing observations in multivariate energy time series. Initially, the inherent distribution of the omitted data is modeled. Subsequently, this learned distribution facilitates the training of an attention-based imputation network. Ascertaining the distribution before the actual imputation allows the architecture to adapt dynamically to the unique traits of the source data. The proposed framework is evaluated using European energy records sourced from the European Network of Transmission System Operators for Electricity. Through comprehensive experiments employing various performance metrics and baseline models, the findings indicate that the introduced approach outperforms comparative methods and reliably estimates missing records.

Keywords Energy data, Neural networks, Multivariate time series, Missing value estimation, Attention model

*Correspondence:
Mikko Rantanen
mikko.rantanen@aalto.fi
[1] Department of Intelligent Power Systems and AI, School of Electrical Engineering, Aalto University, Espoo, Finland
[2] Department of Energy Data Science and Analytics, Faculty of Science, University of Helsinki, Helsinki, Finland

Introduction

Multivariate time series are prevalent in numerous fields, including meteorology, finance, and the energy sector. In energy research, these series often represent power supply and demand, stock market valuations, storage capacities and other variables. Data gaps frequently arise in such datasets primarily because of measurement anomalies or communication failures, which can significantly disrupt subsequent applications in energy system analyses.

Numerous techniques have been formulated to address missing values within multivariate time series. Nevertheless, traditional imputation strategies typically concentrate exclusively on either cross-dimensional relationships or temporal dynamics, but seldom address both concurrently. Conversely, the application of deep learning algorithms for time series analysis, particularly for imputation, has grown considerably. Although these architectures fundamentally rely on sequence-to-sequence networks [1-3], alternative paradigms such as adversarial networks [4], autoencoders [5], and attention mechanisms [6] have also been adapted for data reconstruction.

This research presents a dual-phase imputation framework anchored by an attention-based neural network. The methodology comprises an initial model designed to learn the multivariate distribution of the data gaps, followed by a secondary model that predicts the precise missing values. Because the missing value distribution is determined before the imputation phase, it can be leveraged to produce synthetic data gaps to train the imputation architecture, thereby

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

allowing it to capture the dataset's specific properties. For ease of reading, this proposed framework is designated as MVTSI (multivariate time series imputation) in this manuscript. Although the formulated model is suitable for any form of multivariate time series, this study focuses specifically on its application to energy statistics retrieved from the ENTSO-E transparency platform.

The primary contributions of this paper are as follows:

· We construct an LSTM-driven architecture configured to learn the distribution of absent data within multivariate time series. After training, this model can produce artificial data gaps that mirror the original data's distribution. This process imposes no assumptions or limitations concerning the temporal dynamics or the fundamental nature of the data.

· We introduce an attention-based estimation framework that incorporates the LSTM network's forecasted distribution into its training regimen. The synergy between these two components ensures versatile adaptability to the input dataset. Consequently, the imputation model simultaneously captures cross-dimensional and temporal correlations, calculating all absent records within a specified temporal window concurrently.

· We tailor and rigorously assess our framework explicitly for reconstructing incomplete data within the ENTSO-E dataset. We conduct comprehensive evaluations comparing our methodology against multiple baseline techniques across various data subsets. Ultimately, the proposed model demonstrates advantageous performance relative to the benchmark approaches.

Related Work

This segment reviews pertinent literature fundamental to the present study. Offers a concise summary of existing research regarding the significance and consequences of data gaps in energy datasets. Subsequently, examines diverse methodologies used in time series analysis.

Importance of missing values in energy data

Although incomplete data occurs across virtually all scientific disciplines, its resolution is critically important in the energy sector because most energy system models require gap-free datasets. Such models govern a broad spectrum of applications, including the design of future energy infrastructures, the formulation of CO2 reduction strategies, and the projection of capital investments. Ruggles *et al*. [7] demonstrate that even when solely imputing gaps within electricity demand time series, the outputs of a power system model can fluctuate by 5% depending on the selection between two advanced imputation techniques. This variance occurs even within elementary analyses constrained to a single geographical area.

Because energy system models cannot process missing records, researchers are compelled to implement some form of imputation. Furthermore, the aforementioned study highlights that outputs derived from rudimentary imputation procedures are typically inadequate in quality. This inadequacy is particularly pronounced in models tasked with long-term forecasting of power generation capacity and output. Because a consistent electricity supply must be guaranteed and demand continuously satisfied, variations in the reconstructed data can lead to divergent strategic planning decisions.

Nonetheless, data gaps present significant challenges for short-term forecasting frameworks as well. For instance, in wind energy operations, short-term scheduling is vital for maintaining the equilibrium between power generation and consumption [8]. Because electrical supply and demand require perpetual balance, erroneous approximations and assumptions can induce grid instability, thereby necessitating supplementary stabilization interventions. Under these circumstances, managers of renewable energy facilities are often subjected to financial penalties resulting from inaccurate forecasting.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

To generate these predictions, statistical forecasting techniques are typically favored [9]. Nevertheless, these approaches demand up-to-date measurements of wind speed and power output, often sourced from multiple wind installations. Such observational data is highly susceptible to loss stemming from maintenance activities, transmission delays, or communication failures. For instance, Akçay and Filik [10] utilized a dataset containing wind speed readings from five distinct Turkish monitoring stations, documenting that an average of 2.17% of the data was absent.

The PJM dataset [11] in the United States and the ENTSO-E transparency platform [12] in Europe serve as two widely utilized data repositories for energy system analysis. These databases supply time series detailing national electricity pricing, aggregate demand, and power generation metrics across all deployed technologies. This manuscript centers its analysis on the European platform. The aggregation of measurements from an expansive network of power generation facilities—encompassing decentralized units like residential solar arrays—is inherently complex and yields time series afflicted with substantial data gaps. For example, within the ENTSO-E database, the production mix records for the years 2015 and 2016 exhibited an average deficit of 1000 values per week [13]. As previously established, this level of incompleteness can severely disrupt energy models reliant on comprehensive datasets. Consequently, resolving this issue requires robust data imputation techniques.

Models for time series prediction and imputation

An extensive body of research addresses the challenge of reconstructing omitted observations within time-series datasets. Fundamental approaches include strategies such as mean substitution, linear interpolation, or carrying the preceding value forward. Unfortunately, such techniques fall short of capturing intricate temporal interrelations. Autoregressive frameworks—notably ARIMA or GARCH configurations—extract the underlying trend from the sequence and calculate correlations among successive intervals. The utility of ARIMA is further broadened through state space architectures utilizing Kalman filters [14]. Techniques anchored in k-nearest-neighbor logic [15] attempt to fill gaps by drawing values from the most closely related data clusters. These methodologies have demonstrated considerable efficacy in filling time-series gaps and exhibit broad cross-disciplinary applicability [16, 17].

Nonetheless, recent breakthroughs in deep and machine learning have catalyzed a surge in the use of these advanced algorithms for data recovery. Deploying sequence-based architectures, particularly recurrent neural networks (such as RNNs or LSTMs), has become a prevalent strategy for resolving incomplete time sequences [1-3]. The methodology delineated by Zhang *et al*. [1] puts forward a deep learning framework grounded in bi-directional LSTMs to reconstruct absent sensor measurements. Implementing this framework on Great Barrier Reef water quality records yielded performance improvements of up to 70% across evaluation criteria when compared with selected baselines. Cao *et al.* [18] introduced BRITS, a self-supervised paradigm constructed upon a bidirectional LSTM network. This technique proves effective for completing univariate sequences as well as managing multivariate datasets characterized by interrelated variables. Their investigation evaluates the algorithm against diverse benchmarks—encompassing alternative RNN methodologies—and documents heightened precision across three distinct, autonomous data sets. Additional strategies leverage other cutting-edge neural architectures. Generative Adversarial Networks (GANs), for instance, exhibit profound efficiency in modeling the statistical distribution of multivariate sequences [4]. This learned distribution is subsequently applied to deduce the absent entries within the dataset. The GRUI protocol, outlined by Luo *et al*. [4], demonstrates superior fidelity compared to designated baselines across a variety of time-series collections. Fortuin *et al*. [5] hybridize variational autoencoders with Gaussian processes to approximate the probabilistic distribution of missing elements, specifically targeting medical informatics. Their research cross-examines the proposed architecture across multiple corpora, incorporating clinical time-series records. Outcomes across diverse operations (such as reconstruction and subsequent classification) substantially exceed those of conventional benchmarks, achieving parity with frontier techniques like BRITS [18].

More recent paradigms emphasize attention-driven configurations to discern complex dependencies within multivariate arrays. The self-attention protocol pioneered by Vaswani *et al*. [19] now defines the frontier in various machine learning sectors, most prominently within natural language processing. Nonetheless, this foundational concept translates

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

effectively to sequential tracking and computer vision tasks. Zerveas *et al.* [20] deployed an attention-centric schema specifically for the classification and regression of temporal data. Their design features a neural encoder powered by attention mechanisms, meticulously optimized to handle multivariate temporal inputs. The findings indicate that this architecture outperforms comparative baselines in both classification and regression contexts. Crucially, the authors note the framework's inherent adaptability for future forecasting and data imputation. Ma *et al*. [6] developed CDSA, a cross-dimensional self-attention structure engineered for multivariate sequences tagged with geographic coordinates. They advocate fractionating the model's attention matrix to embed inter-dimensional correlations without inflating computational expenditures. Consequently, this blueprint yields reduced reconstruction errors relative to selected benchmarks across multiple geographically annotated datasets. **Table 1** presents a synthesized overview of the reviewed literature concerning time-series reconstruction. Although many strategies exist, we are not aware of any deep learning frameworks tailored exclusively to energy-sector time sequences.

Table 1. An overview of relevant models and literature.

Method	Study	Approach	Application
KNN	[15]	K-nearest neighbor method	Power dataset
SSIM	[1]	LSTM-based	Sensor data
BRITS	[18]	LSTM-based	Medical data, sensor data
GRUI	[4]	GAN-based	Medical data, sensor data
GP-VAE	[5]	Autoencoder-based	Medical data
CDSA	[6]	Attention-based	Sensor data
MVTSI	This study	Attention-based	Energy data

Data

The subsequent segment provides an in-depth characterization of the dataset driving our experimental evaluations. The primary information source is the ENTSO-E transparency platform.1 Operating as the European Network of Transmission System Operators for Electricity, ENTSO-E functions as a unified consortium fostering collaboration among various transmission system operators (TSOs) across the continent. This coalition encompasses 39 distinct TSOs spanning 35 sovereign nations. Specific details about ENTSO-E, such as its core objectives and participant roster, are openly available on the initiative's official website.2 Through a dedicated API, the ENTSO-E transparency platform facilitates streamlined retrieval of records about the broader European power network. While a comprehensive audit of this platform exceeds the current article's parameters—such a macroscopic review is documented by Hirth *et al*. [13]—a concise appraisal of data integrity is provided below, with a strong emphasis on the probabilistic spread of missing entries. To contextualize this, the following discussion briefly touches upon the findings of Hirth *et al*. [13], who conducted a rigorous diagnostic of ENTSO-E data fidelity. Yet, given their analysis was confined to the 2015–2016 operational window, we append a supplementary statistical evaluation to reflect a broader timeline.

Hirth *et al*. [13] deliver a profound examination of the ENTSO-E transparency portal, evaluating its overarching goals, technical traits, operational methodologies, and overall data viability. Because our proposed framework prioritizes data cleanliness, tangential operational details regarding the platform are omitted from this discourse. The dataset utilized by Hirth *et al*. [13] was extracted in April 2017. Consequently, the updated archival data deployed in our current investigation may exhibit discrepancies owing to subsequent retroactive corrections. Evidence indicates that the prevalence of missing

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

inputs fluctuates radically based on the specific metric, the focal year, and the geographical region. As a case in point, Denmark boasts 100% data continuity across all parameters, excluding “Hydro Run-of-river and poundage,” which suffers a 32.4% deficit. Conversely, Italy demonstrates a mere 50% completeness rate, a shortfall primarily driven by a full missing calendar year across the majority of recorded dimensions. The Completeness of “Aggregated Generation per Type” table originally constructed by Hirth *et al.* [13] has been regenerated using the expanded 2015–2020 dataset. Moving forward, however, our analytical focus pivots away from the baseline integrity of the raw ENTSO-E archives. It concentrates entirely on the pre-processed state required for our algorithmic pipeline.

The finalized experimental corpus encompasses 29 regional territories tracked by ENTSO-E. This collection captures the EU27 block—minus Cyprus and Malta—while absorbing the United Kingdom, Norway, Serbia, and Switzerland. Crucially, the records for Germany and Luxembourg are mathematically merged, reflecting their status as a unified bidding zone within the energy exchange. This informational array spans the entire 2015 to 2020 interval (inclusive) and was compiled via the public API on 09.06.2021. In addition, a multitude of distinct variables underwent strategic consolidation. This choice was driven by both the need to validate the model on a streamlined data subset and the realization that numerous distinct parameters lack significant predictive value. To illustrate, the “Fossil Peat” metric was completely excised because it was only logged by three of the 29 constituent nations. The surviving columns were condensed according to this schema: “Hydro Water Reservoir” and “Hydro Run-of-river and poundage” were conflated into a singular “Hydro” umbrella. Concurrently, categories encompassing “Geothermal,” “Fossil Coal-derived gas,” “Fossil Oil,” “Waste,” “Other renewable,” and “Fossil Peat” were absorbed into the overarching “Other” designation. Omissions were rigorously propagated throughout this fusion stage; thus, if any isolated sub-feature was missing, the resulting macroscopic feature was immediately categorized as null, reflecting the impossibility of calculating the true aggregate sum. This systematic distillation left a final feature space of n = 13. **Figure 1** visualizes a sample from this structured dataset.

The subsequent paragraphs deliver a granular examination of the dispersion and inherent traits of the data vacuums. **Table 2** presents descriptive statistics for these absences across varying aggregation scales. For instance, a breakdown of missing measurements per localized 24-hour cycle reveals an average deficit of 10.33 records. Despite this, **Table 2** also verifies that the data matrix remains completely intact for over 50% of the recorded days. Because the ultimate intent is to replace these gaps algorithmically, prioritizing the daily void count specifically on days experiencing data loss provides vastly greater diagnostic utility. After all, a flawlessly recorded 24-hour period necessitates zero restorative action. Corresponding statistical summaries for this filtered subset are similarly tabulated in **Table 2**. **Figure 2** displays a frequency distribution illustrating the daily volumetric data loss per nation, isolating the flawed daily cycles exclusively. A striking pattern emerges: when a given day registers data gaps, the predominant frequency of those missing values spikes at precisely 24. This phenomenon almost certainly correlates to an entire metric being offline for the duration of that specific calendar day. We additionally investigate the severity of sequential recording failures (i.e., gap length). Although the mathematical mean hovers around 21 successive absent steps, this metric is heavily skewed by severe outliers where certain metrics vanished for almost an entire year. **Table 2** clarifies that the 75%-quantile for continuous gap length is 6, meaning that exactly 25% of all systemic data breaks consist of 6 or more consecutive failed measurements. Our diagnostics confirm that when fully intact intervals are removed from the equation, the density of missing data across given time boundaries (day, timestamp) remains alarmingly severe. The propensity for multi-dimensional, cascaded recording failures inherently invalidates the use of rudimentary statistical or univariate algorithms for recovery, thereby justifying the advanced methodology introduced within this manuscript.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

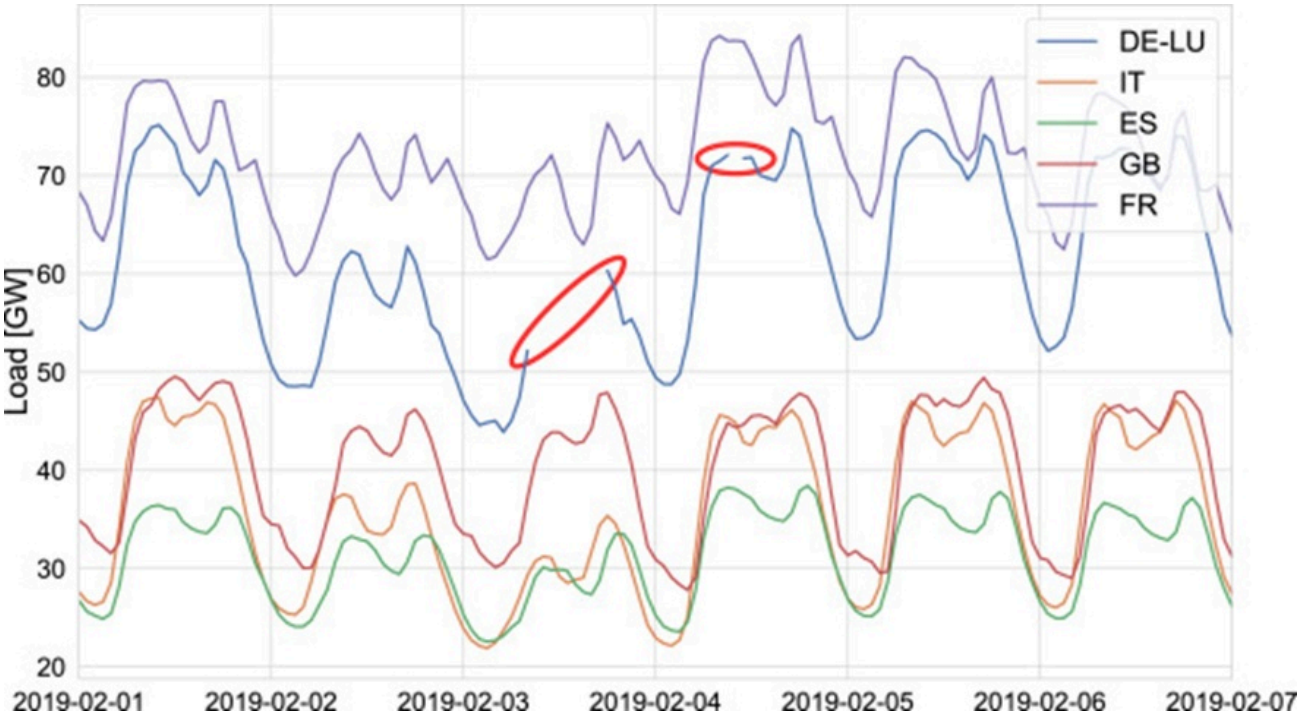


Figure 1. Weekly load profile for five different countries. During the weeks, there are two gaps of different lengths with missing values for the bidding zone DE-LU (Germany & Luxembourg), indicated by the red ellipses (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 2. Summary statistics of different aggregation schemes for the missing values.

Aggregation method	Max	75%	50%	25%	Min	Mean
Per day and country[a]	242	13	0	0	0	10.33
Per day and country[b]	242	27	24	13	1	31.51
Per day and feature[a]	150	45	16	0	0	21.58
Per day and feature[b]	150	49	28	24	1	34.23
Per timestamp[a]	72	15	10	7	3	11.69
Per timestamp[b]	72	15	10	7	3	11.69
Gap length[b]	46222	6	3	1	1	21.36

[a]: Refers to complete data.

[b]: Only includes data with missing values in the specific aggregation.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

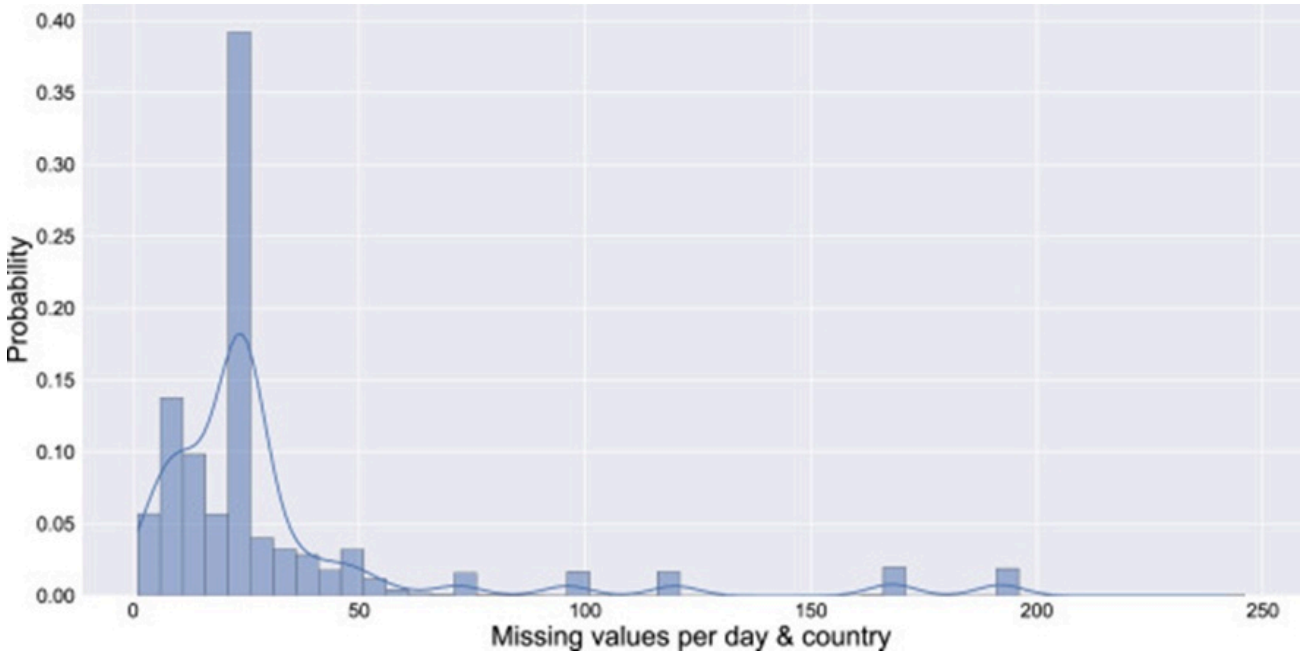


Figure 2. Histogram of the number of missing values per day and country, including only incomplete days.

At this juncture, it is critical to state our assumption that the dataset exhibits either completely random omissions or data missing at random. The former hypothesis postulates that recording lapses occur exclusively as a result of stochastic, unpredictable disruptions. Within our experimental framing, this appears overly bold, given the highly chronological nature of the gaps—such as the recurring phenomenon of exactly 24 hours missing. Conversely, a missing at random assumption posits that the likelihood of a gap could be probabilistically linked to other observed attributes within the matrix. To conceptualize this, the vulnerability to data loss in one specific country could be statistically higher than in a neighboring jurisdiction. Even if declaring the gaps as missing completely at random is overly extreme, proceeding under the missing at random paradigm is defensible for our architecture, as the variables that can dictate failure probabilities are actively fed into the network. Nonetheless, we must concede the possibility that the omissions are strictly missing not at random. Such a reality would imply the presence of invisible, unrecorded external factors fundamentally driving the data outages. Under those specific conditions, the predictive outputs generated by our model would carry the risk of systemic bias.

Materials and Methods

This section outlines the proposed framework comprehensively, detailing the underlying mechanics of the dual-stage model. The structural configuration of the pre-processed input data is illustrated in **Figure 3a**. Assume $C = \{C_1, \ldots, C_m\}$ represents a collection of m countries (such as “France” or “Germany”) and $S = \{S_1, \ldots, S_n\}$ denotes a group of n distinct time series variables (for instance, “electricity production from solar” or “electricity prices”). Additionally, let $Y = \{Y_1, \ldots, Y_l\}$ signify the designated years, while $T = \{T_1, \ldots, T_{8760}\}$ serves as an hourly temporal index for a given year, capturing details regarding the corresponding hour, day, week, and month.

For any given country, the associated features are represented by a multivariate time series $\mathbf{S} = \{s_1. \ldots s_T\}$ comprising T total observations. A single observation at timestamp t encompasses N specific features, formulated as $\mathbf{s}_t = \{ s_t^1, \ldots, s_t^N\} \in \mathbb{R}^N$. To effectively represent missing values, a masking vector xt is introduced, which encodes whether a specific feature is present or absent. This vector is mathematically established as follows:

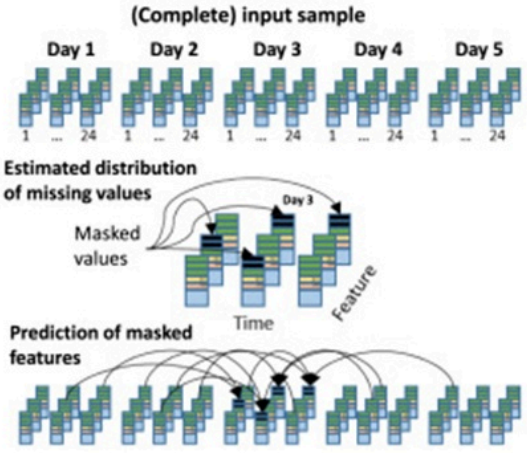
© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

$$x_t^n = \begin{cases} 1 & \text{if } s_t^n \text{ is missing} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

As established previously, this construct functions as a multivariate time series $\mathbf{X} = \{\mathbf{x}_1. \ldots \mathbf{x}_T\}$. To accurately capture the internal correlations within the time series, the input data undergoes formatting transformations. Specifically, a sliding-window technique with a 24-hour step size is applied to extract a multivariate sequence for each respective country. Consequently, each generated sample encompasses a five-day window spanning all features. As depicted in **Figure 3b**, the MVTSI methodology is designed to impute absent data points located within the central day of this sequence. A key benefit of this configuration is its capacity to explicitly leverage data chronologically adjacent to the gaps, while broader relational patterns are inferred automatically from the dataset. By integrating the masking vector directly into its inputs, the system maintains robust flexibility when encountering further missing data points across varying features and days. Theoretically, different window dimensions could be selected. Nevertheless, because attention mechanisms exhibit quadratic memory complexity, expanding this window might render the process computationally unviable depending on hardware constraints. While the window span is a parameter that can be optimized, empirical findings indicate highly favorable outcomes without rigorous tuning; thus, it is kept constant at five days.

Year	Country	Timestamp	S_1	S_2	...	S_n
Y_1	C_1	T_1	.	.	...	.
		T_2	.	.	...	.
		...	...	...	...	...
		T_{8760}	.	.	...	.
	C_2	T_1	.	.	...	.
		T_2	.	.	...	.
		...	...	...	...	...
		T_{8760}	.	.	...	.
	...	...	...	...	...	...

(a) Demonstration of the input data.



(b) Example of the masking setup. Missing values are imputed for the middle day of five consecutive days.

Figure 3. The left figure shows the raw multivariate time series, as obtained from the ENTSO-E data. Additional features, like the masking vector, are added before the data is fed into the model. The right figure shows an imputation example for a complete input sample. However, additional missing values across the time domain are not a problem for the imputation model.

The suggested pipeline for data imputation involves several phases illustrated in **Figure 4.** Following data structuring via the sliding-window strategy, the initial stage deploys an LSTM-centric neural network tasked with modeling the distribution patterns of absent values within the dataset. Subsequently, this trained network generates synthetic data gaps that closely mimic the natural distribution. Because assessing imputation accuracy on truly missing instances is impossible (due to the lack of ground truth), this generation phase is critical for computing training loss and optimizing the secondary imputation model. Guided by these artificially masked inputs, the second model trains to deduce values exactly as gaps naturally manifest in the raw set. During the inference phase, only this attention-based mechanism is required to estimate the missing numbers. While the methodology relies on these two distinct architectures, their specific framework is technically interchangeable. For instance, alternative neural network designs could be substituted into either the

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

distribution-learning phase or the imputation phase. These two core components are elaborated upon in the subsequent sections.

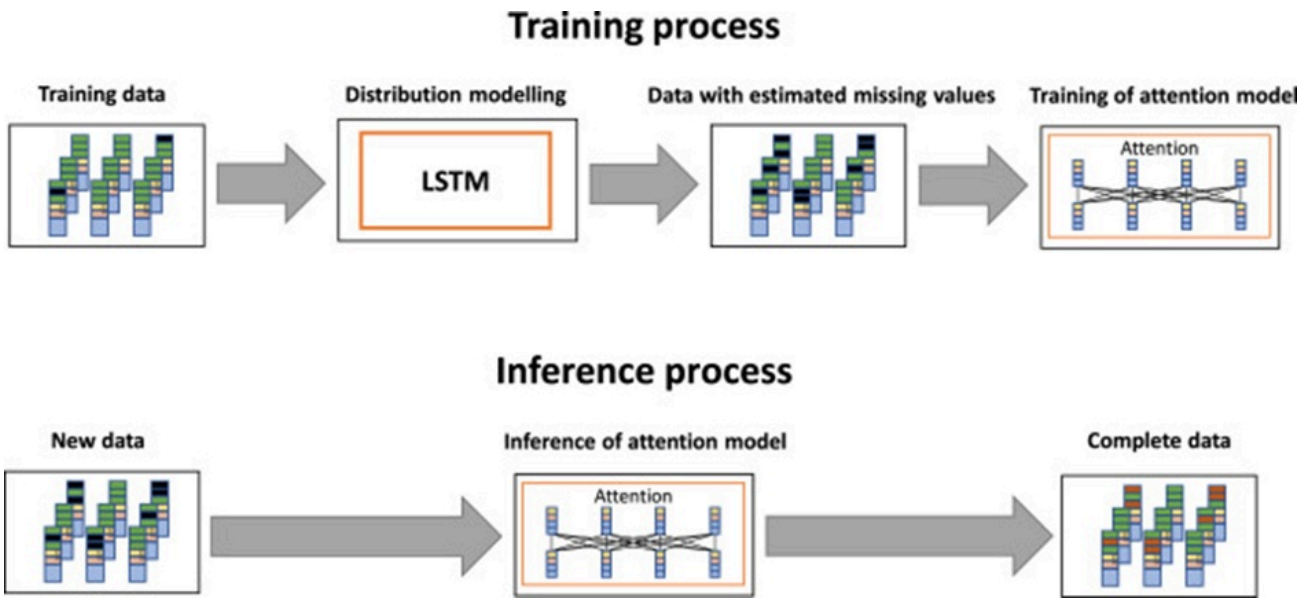


Figure 4. An abstract scheme for the model workflow. In the training process, missing data are generated artificially by the first-stage model and then further used to train the attention model. For performing inference on unseen data, only the second model is used to impute the actual missing values.

Modeling of missing data distribution

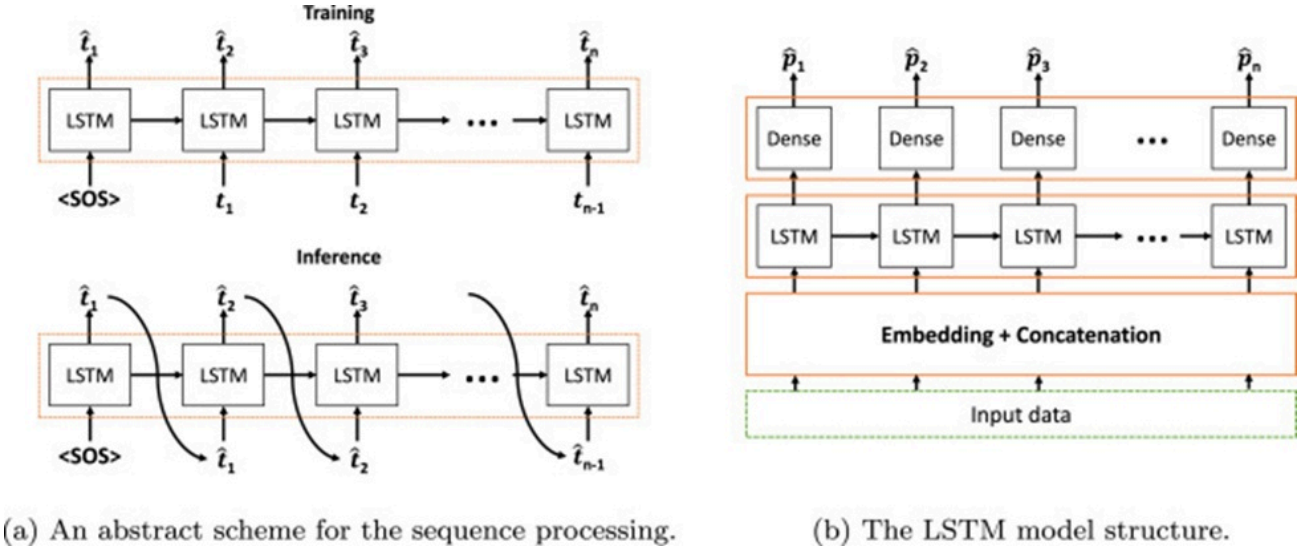
The primary phase of this analytical framework focuses on approximating how missing data points are distributed across the utilized dataset. To achieve this, a sequence model based on Long Short-Term Memory (LSTM) architecture is employed. The processing mechanism for time series inputs within this network is depicted in **Figure 5a**. Initially, a start-of-sequence (SOS) token is appended, signaling the algorithm to forecast the sequence's initial data point. The architecture then continues to forecast subsequent values sequentially. During training, a teacher-forcing methodology is applied; this technique feeds the actual ground-truth value into the network before the subsequent prediction step, thereby accelerating the neural network's convergence time significantly [21]. Conversely, during inference, the entire sequence is constructed solely based on the model's iterative outputs. **Figure 5b** illustrates the foundational design of this LSTM framework. Rather than ingesting the complete time series at once, the network operates on 24 h subsets, aligning to model data gaps within that targeted timeframe. The localized time series is fused with descriptive embeddings, such as positional encodings or spatial country indicators. Within the context of the MVTSI strategy, the input functions as a binary array detailing the presence or absence of individual elements. As an initialization token, the value −1 is implemented. Computationally, these inputs propagate through at least one LSTM layer before reaching a terminal Dense layer equipped with a sigmoid activation function:

$$S(x) = \frac{1}{1 + e^{-x}}. \quad (2)$$

Given that $(x) \in [0, 1]$, the network's generated output logically translates to the probabilistic likelihood of the impending value being absent. As previously highlighted, the training phase minimizes the binary cross-entropy loss metric through the implementation of teacher-forcing. When transitioning to inference, the algorithm relies exclusively on its autonomous

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

sequential forecasts to construct the full temporal sequence. To finalize the output, these continuous probability estimates are mapped into a discrete binary vector by sampling from a Bernoulli random variable corresponding to the predicted probabilities.



(a) An abstract scheme for the sequence processing. (b) The LSTM model structure.

Figure 5. The LSTM model for estimating the distribution of missing data. In the training process, the true data is used as an input for the next step, while in the inference process the predictions are fed back into the model. The model consists of several embedding and concatenation layers, followed by the bi-directional LSTM units. The final output is the probability of each value being missing.

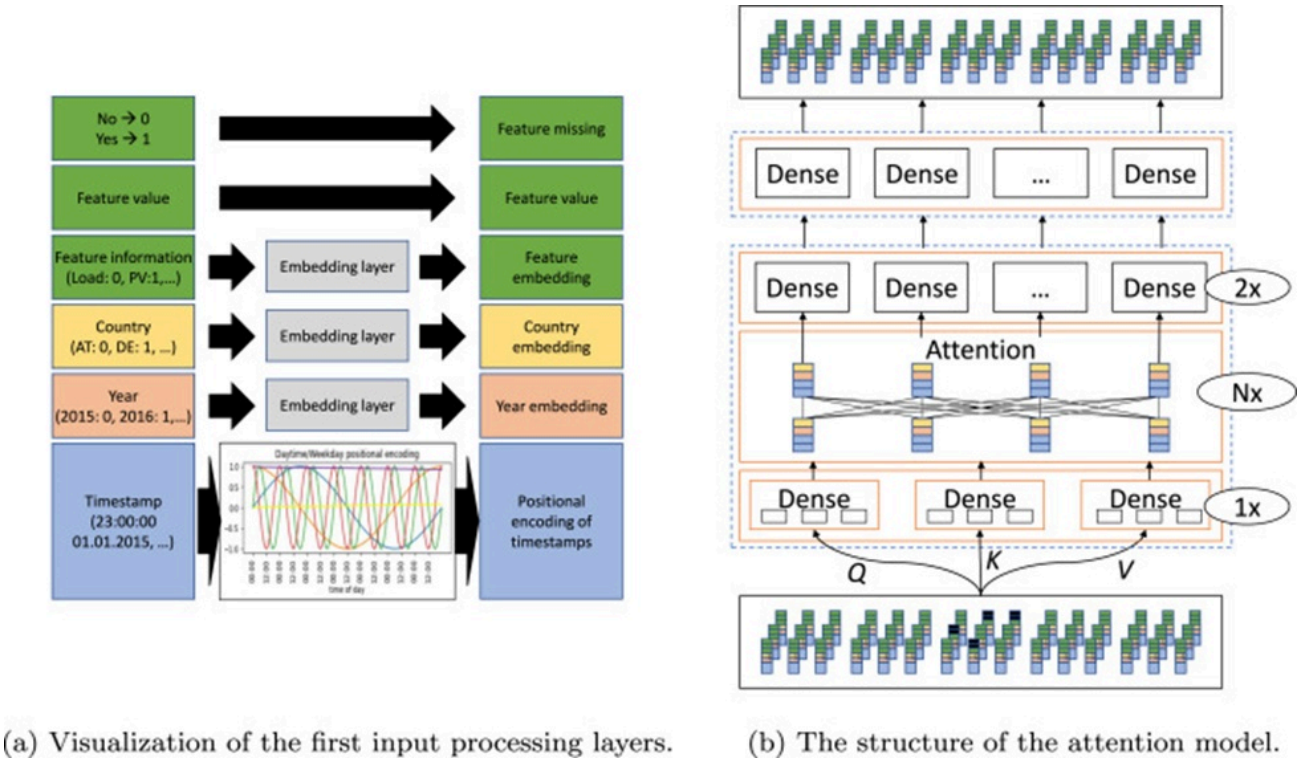
Attention model for imputation

This segment details the secondary phase of the methodology, namely the attention-driven imputation network. Unlike models that process time series sequentially, attention-based structures simultaneously map dependencies across every dimension and time step synchronously. To facilitate optimal pattern recognition, positional encodings are integrated into the dataset, supplying necessary contextual clues regarding the relative order of temporal elements. Furthermore, the architecture utilizes distinct categorical embeddings to represent the year, the country, and the feature, as illustrated in **Figure 6a**. These embedding mechanisms excel at deriving multi-dimensional representations from discrete categorical inputs.

Fundamentally, the core attention apparatus relies on the matrices Q, K, $V \in \mathbb{R}^{T,d}$ (representing the query, key, and value elements)—alongside an attention mask denoted as $M \in \mathbb{R}^{T,T}$. In this context, T denotes the dimensionality of the stacked inputs described in **Figure 6a,** whereas d acts as a hyperparameter defining the embedding dimension. Because the proposed methodology incorporates self-attention, the Q, K, V matrices are identical. Additionally, the standard causal attention mask is omitted; since the task requires an imputation approach, the architecture is “allowed” complete visibility of all data points. In predictive forecasting, by contrast, the mask M strictly prevents the leakage of future values into current estimations. The attention mechanism is subsequently formulated as $(Q, K, \ ) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$. This computation is broadly recognized as scaled dot-product attention [19]. Conceptually, this operation utilizes the softmax activation to enable the network to dynamically concentrate its “attention” on specific input segments depending on their contextual relevance. In an energy sector application, for instance, the algorithm might logically avoid applying attention to the Nuclear feature in Denmark, given that those metrics continuously remain at zero (Denmark operates no nuclear energy production). The structural layout of this attention framework is presented in **Figure 6b.** As previously detailed, the underlying data is generated via a 5-day sliding window mechanism, shifted by a single day, encompassing all variables

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

and geographic boundaries. For every generated sample, the preliminary first-stage model evaluates the probability distribution of data gaps occurring during the middle day. Leveraging this artificial dataset, the attention-based stage is trained to accurately fill these missing numbers by minimizing the mean squared error.



(a) Visualization of the first input processing layers. (b) The structure of the attention model.

Figure 6. The proposed attention model. Several embedding layers are used to implement categorical features, for example, the country, into the model. The data is fed through the attention layers, and the model learns to impute missing values in the middle day of the time window. In the setting of self-attention, the three matrices are equal and of dimension $Q, K, V \in \mathbb{R}^{T,d}$, where T is the input dimension, and d is the embedding dimension.

It should be emphasized that the current framework executes deterministic forecasting of the absent data and does not inherently evaluate the uncertainty surrounding its predictions. Transitioning toward a probabilistic forecasting paradigm, however, would be highly advantageous for quantifying the variance associated with these estimates. Numerous methodologies exist for transmuting rigid deterministic results into nuanced probabilistic forecasts. A straightforward option is the Single Gaussian technique [22], which utilizes the network's output to construct a predictive density distribution modeled on a normal distribution. Alternatively, Lakshminarayanan *et al*. [23] advocate deep ensembles, an algorithm in which multiple instantiations of a neural network are trained collectively, governed by a proper scoring rule [24] serving as the training standard.

Furthermore, Gal and Ghahramani [25] recommend integrating dropout during the neural network training phase to represent underlying model uncertainty successfully. While these illustrate just a selection of the specialized approaches tailored to deep neural models—and given that a majority of these methods function independently of the precise architectural design—this study deliberately centers its efforts on deterministic evaluation. Consequently, the incorporation of uncertainty prediction remains designated as a prospective subject for future research.

Implementation details

Each sample st undergoes normalization by subtracting the mean and dividing by the training set's standard deviation. All absent data points are converted to zero, as the neural network cannot process missing values that are not explicitly

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

encoded. The two stages of the model are trained using the Adam optimizer [26] to minimize mean squared error (MSE). The process employs a batch size of 32 alongside a tailored learning rate scheduler featuring early stopping capabilities. To achieve this, the dataset is partitioned into distinct training and validation subsets. Training halts automatically if the validation loss fails to show improvement over a predetermined number of iterations. Implemented using Python and TensorFlow, the architecture was trained across various GPUs via Google Colab. The cumulative execution time for all experimental runs remained under roughly 3 hours. The associated source code is accessible on GitHub.3

Results and Discussion

This segment outlines the experimental setup and the procedures for evaluating the model. Because the two distinct phases of the model are assessed independently, the evaluation is divided into two corresponding sections. Initially, the focus is placed on the first-stage architecture and its capacity to approximate the distribution of missing data. Subsequently, a comprehensive performance analysis of the attention-based model is presented using various metrics, benchmarking it against several baseline methods.

Evaluation of distribution modeling

To properly assess the initial stage, one must understand its specific estimation objective. The intention is not to overfit to the exact locations of missing values, as this would produce an estimator that merely memorizes data gaps. Instead, the objective is to generate a distribution closely resembling the natural pattern of the missing data. Because the model is stochastic, the generated missing values vary per iteration but should maintain the original data's core statistical properties. Since the architecture estimates a unique distribution for each time series, comparing every single estimated distribution to its original equivalent is impractical. Thus, summary statistics across all test instances are evaluated to verify whether the approach captures the fundamental characteristics of the true missing value distribution. Specific summary statistics were selected for this purpose. First, the daily count of missing values across all features is compared.

Furthermore, the total missing values per feature and per timestep are analyzed. Finally, the length of missing value sequences per feature and day is also evaluated. These statistics can be represented as histograms or empirical distribution functions. Consequently, specific distance metrics are required to measure the similarity between two (empirical) distributions. For this task, the earth mover's distance (also known as the first Wasserstein metric) and the Hellinger distance are utilized. Both yield a value of zero if the (empirical) distributions are identical. The Hellinger distance for discrete distributions is defined as

$$H(P,Q) = \frac{1}{\sqrt{2}} \| \sqrt{P} - \sqrt{Q} \|_2, \quad (3)$$

Where P, Q are probability vectors. In our case, they refer to the probability of a specific value being missing. (, Q) fulfills the property 0 ≤ (, Q) ≤ 1. Let U, V be distributions with cumulative distribution functions F_U , F_V, respectively. The first Wasserstein distance is then defined as:

$$W_1(U,V) = \int_{-\infty}^{\infty} |F_U(x) - F_V(x)| dx. \quad (4)$$

See Nguyen [27] for further information on characteristics of the Wasserstein metric and Meshgi and Ishii [28] for further ways to measure the similarity of histograms.

The next phase involves analyzing the performance of the LSTM model tasked with predicting the distribution of missing data. **Figure 7** presents the histogram and empirical cumulative distribution function (ECDF) for both the estimated and actual distributions. The visual similarity confirms the model's proficiency in mirroring the overall missing data patterns

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

present in the dataset. As previously detailed, distinct metrics are applied to quantify this similarity. **Table 3** displays the distance metrics for these summary distributions. With a value of zero indicating a flawless match, the results confirm the proposed architecture successfully generates a missing value distribution that faithfully replicates the underlying dataset's properties. Consequently, this trained model predicts missing test data, thereby generating the necessary data to train the attention model on the imputation task.

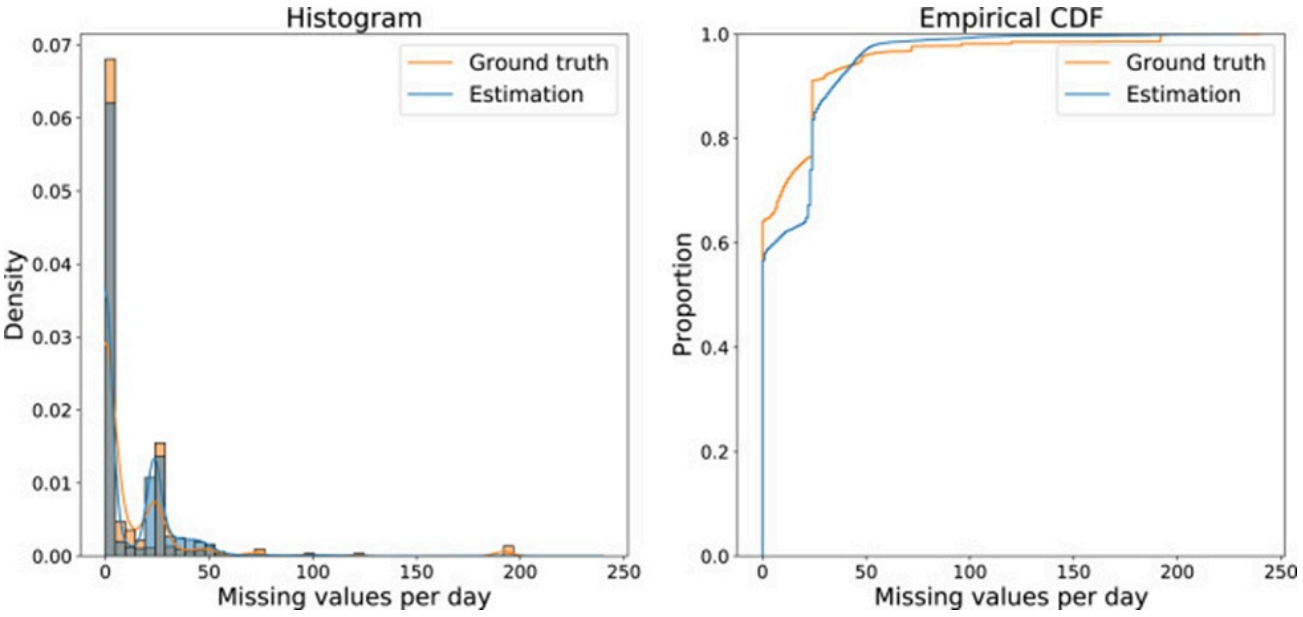


Figure 7. Both figures show the estimated and true (empirical) distribution of the amount of missing values per day in the test data. The left plot shows the values in a histogram, while the right plot shows the corresponding empirical cumulative distribution function.

Table 3. Distance measures between the model predictions and the test data over different variants of distribution characteristics. A smaller value indicates a higher similarity between the two distributions.

Missing values	Hellinger distance	Wasserstein distance
Per day	0.2978	0.0014
Per feature (average across all features)	0.1343	0.0048
Per timestep (average across all timesteps)	0.1231	0.0105
Sequence length (per feature and per day)	0.3391	0.0154

Evaluation of imputation model

Various metrics are applied to assess the quality of the data imputation methods and examine their relative performance. The mean squared error (MSE) and mean absolute error (MAE) serve as standard indicators for evaluating the deviation of predictions from their original values. For a rigorous evaluation, the dataset must be partitioned into training and testing subsets. Because the dataset consists of ordered time series, strictly random sampling is inappropriate. Conversely, isolating a single continuous time slice (e.g., one year) could introduce bias into the evaluation. Therefore, our methodology utilizes a test set with a monthly resolution. From a total of 1958 distinct monthly data samples (encompassing all unique countries, years, and months), 10% is designated as the test set. This data subsequently

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

undergoes the previously detailed rolling window transformation. This approach preserves the temporal sequence while ensuring sufficient randomness for a stable and robust evaluation process.

The initial-stage model is used to approximate the underlying missing-value distribution. Trained exclusively on the training data, it generates a missing value mask for the test set. This mask is applied to the data to define the locations of the missing values. Because some original ENTSO-E data points might actually be missing, the evaluation metrics are computed exclusively on points the LSTM model predicts to be missing, but that are not genuinely absent in the raw data. If the true value were actually unknown, evaluation would be impossible. During training, the LSTM model operates solely on incomplete data. The rationale is that forecasting fully complete days offers no utility for an imputation model. By learning exclusively from incomplete data, the model tends to over-predict the frequency of days with missing values compared to the raw dataset. However, because complete days are irrelevant to the core imputation objective, this methodology is justified. The attention-based framework is benchmarked against several time series imputation techniques, encompassing both traditional statistical approaches and deep learning methodologies. The primary “naive” baseline is the Last Observation Carried Forward (LOCF) method, which substitutes missing data with the immediately preceding observation. The k-nearest-neighbor (KNN) imputation method serves as the second comparative benchmark. Additionally, the framework is compared to a bidirectional LSTM, a deep learning alternative resembling the methodology introduced by Cao *et al*. [18].

Table 4. Result metrics and their respective standard deviations. The lowest result is highlighted in bold. For all models, the standard deviation is quite high because some time series are hard to replicate, leading to high errors and affecting the aggregated MSE.

Model	MAE	MSE
MVTSI	**0.0845** ± 0.2321	**0.0610** ± 0.5822
LSTM	0.1217 ± 0.2320	0.0687 ± 0.5810
LOCF	0.0959 ± 0.3062	0.1030 ± 0.9196
KNN	0.0927 ± 0.2412	0.0668 ± 0.5645

As noted, performance is measured via the mean squared error and the mean absolute error. **Table 4** presents these two error metrics alongside their corresponding standard deviations for each evaluated model. The attention-based framework achieves the lowest error rates for both MSE and MAE, maintaining standard deviations comparable to those of the benchmark models. Notably, all evaluated techniques exhibit relatively high standard deviations relative to their mean values. This phenomenon occurs because certain missing data points present extreme predictive difficulty, generating significant errors that inflate the aggregated variance. A plausible explanation is that specific data gaps occur entirely at random, heavily challenging any predictive model. Given that this difficulty impacts all models uniformly, comparative validity is preserved. **Table 5** details the MSE for each model categorized by specific data features. The attention model registers the lowest error across most features. Crucially, it demonstrates strong performance on variables containing a high volume of missing data, indicating its efficacy in managing the features responsible for the largest data gaps. The correlation between missing data volume and model precision is compelling and warrants further investigation. The first-stage architecture could be easily adjusted to simulate distributions with varying densities of missing values, enabling a systematic analysis of the imputation models’ responses. However, as this research specifically targets the ENTSO-E dataset and its inherent traits, such experiments are reserved as a direction for future exploration.

Table 5. Mean squared error calculated per individual feature, as well as share of total missing values. The best model for each feature is highlighted in bold.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Feature	KNN	LOCF	LSTM	MVTSI	Missing
Hard coal	0.0380	0.0369	0.0291	**0.0222**	23%
Other	0.0169	0.0209	0.0209	**0.0108**	21.46%
Pumped Storage	0.1929	0.3110	0.1785	**0.1662**	16.93%
Fossil Gas	0.0719	0.1150	**0.0607**	0.0636	16.11%
PV	0.0662	**0.0660**	0.0990	0.0778	10.45%
Hydro	0.0268	0.0237	0.0527	**0.0194**	9.49%
Biomass	0.0160	0.0144	0.0343	**0.0110**	5.97%
Day ahead price	0.0957	0.1865	**0.0942**	0.1012	5%
Wind Offshore	**0.1311**	0.3651	0.1429	0.2103	4.14%
Load	0.0074	0.0107	0.0205	**0.0060**	3.15%
Wind Onshore	0.1129	0.1845	**0.0871**	0.1107	2.52%
Nuclear	**0.0054**	0.0077	0.0338	0.0102	1.71%
Lignite	0.0209	0.0199	0.0251	**0.0040**	1.32%

Table 6. Different technologies for electricity production, as well as their respective greenhouse gas emissions [29], calculated for 2021 in Germany [30]. The right column shows the Model with the lowest MSE per feature.

Feature	Best model	CO_2 eq [Mt]	CO_2 eq [kg/MWh]
Lignite	MVTSI	121.55	1229
Hard Coal	MVTSI	56.95	1227.3
Fossil gas	LSTM	24.96	488.4
Biomass	MVTSI	8.68	201.4
PV	LOCF	4.15	85.8
Pumped storage	MVTSI	0.22	56.8
Wind Onshore	LSTM	2.43	27.1
Nuclear	KNN	0.76	11.7
Wind Offshore	LOCF	0.65	27.1
Hydro	MVTSI	0.07	4.6
Other	MVTSI	–	–

This final segment provides a brief perspective on the subsequent processing of the imputed dataset. To this end, the imputed features are contextualized with respect to their associated greenhouse gas emissions. **Table 6** details the

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

kilogram CO2 equivalent per MWh for specific generation technologies, alongside the best-performing model exhibiting the lowest MSE for each. Emission values are sourced from Xu *et al*. [29]. The MVTSI framework notably excels in modeling technologies with the highest carbon footprints, particularly Lignite and Hard coal. Conversely, its accuracy decreases for highly variable energy sources, such as wind and photovoltaic power. These findings indicate that the proposed methodology is well-suited for preparing data for energy system models that forecast future technology utilization and greenhouse gas impacts. However, expanding this application through a more detailed study extends beyond the scope of this paper.

Conclusion

Addressing missing data within multivariate time series remains a critical challenge across numerous disciplines, including energy sector modeling. Extensive research has focused on formulating effective imputation techniques, increasingly leveraging machine learning algorithms in recent years. This study introduces a dual-stage, attention-driven data framework designed to impute multivariate time series gaps. The initial component of the architecture deduces the statistical distribution of missing values from the raw dataset. This learned information is subsequently utilized to train an attention-based neural network tasked with the actual data imputation. While the primary application centers on the ENTSO-E dataset, the proposed structure offers flexibility, accommodating diverse multivariate time series across any chosen temporal horizon. Our experimental evaluations demonstrate that the proposed model surpasses selected baseline methods in predictive accuracy.

Beyond standard time series evaluation metrics, the study provides a granular analysis of model performance relative to specific electricity generation technologies. These findings reveal that our approach achieves the lowest error (in terms of MSE) across most features. However, this superiority does not extend to all variables and is heavily influenced by the underlying missing data distribution. Investigating the interplay between feature dynamics, model behavior, and missing data volume presents a valuable opportunity for future insights. Enhancing and fine-tuning the architecture could potentially yield improved accuracy for these specific features while preserving the model's broad generalization capabilities. As previously noted, multiple strategies exist to transition the framework toward a probabilistic imputation setting. Integrating these approaches to analyze the uncertainty in our model's predictions represents a highly promising direction for subsequent research.

Regarding the architectural design, incorporating cross-dimensional self-attention [6] could be a viable pathway to boost predictive performance while simultaneously lowering computational demands. Furthermore, alongside the evaluated metrics, we offer an initial perspective on leveraging the proposed approach to ready imputed data for downstream applications, such as integration into energy system models. Future investigations should thoroughly explore how the selection of a specific imputation technique ultimately impacts the analytical outputs of models that subsequently process the data.

Acknowledgements

Juan Gómez-Romero is partially supported by the projects IA4TES (MIA.2021.M04.0008, funded by the European Union – NextGenerationEU) and DeepSim (A.TIC.244.UGR20, funded by FEDER, Spain – Junta de Andalucía).

Conflict of interest

None

Financial support

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

None

Ethics statement

None

Received: 04 Jul 2023 Revised: 08 Sep 2023 Accepted: 25 Sep 2023
Published online: 15 June 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Zhang YF, Thorburn PJ, Xiang W, Fitch P. SSIM—A deep learning approach for recovering missing time series sensor data. IEEE Internet Things J. 2019;6(4):6618-28. https://doi.org/10.1109/JIOT.2019.2909038.

2. Yuan H, Xu G, Yao Z, Jia J, Zhang Y. Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18). New York (NY): Association for Computing Machinery; 2018. p. 1293-300. https://doi.org/10.1145/3267305.3274648.

3. Verma H, Kumar S. An accurate missing data prediction method using LSTM based deep learning for health care. In: Proceedings of the 20th International Conference on Distributed Computing and Networking (ICDCN '19). New York (NY): Association for Computing Machinery; 2019. p. 371-6. https://doi.org/10.1145/3288599.3295580.

4. Luo Y, Cai X, Zhang Y, Xu J, Xiaojie Y. Multivariate time series imputation with generative adversarial networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. Vol. 31. Red Hook (NY): Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper/2018/file/96b9bff013acedfb1d140579e2fbeb63-Paper.pdf.

5. Fortuin V, Baranchuk D, Rätsch G, Mandt S. GP-VAE: Deep probabilistic time series imputation. arXiv:1907.04155. 2020.

6. Ma J, Shou Z, Zareian A, Mansour H, Vetro A, Chang SF. CDSA: cross-dimensional self-attention for multivariate, geo-tagged time series imputation. arXiv:1905.09904. 2019.

7. Ruggles TH, Farnham DJ, Tong D, Caldeira K. Developing reliable hourly electricity demand data through screening and imputation. Sci Data. 2020;7(1):1-14. https://doi.org/10.1038/s41597-020-0503-4.

8. Pinson P. Wind energy: Forecasting challenges for its operational management. Statist Sci. 2013;28(4):564-85. https://doi.org/10.1214/13-STS445.

9. Giebel G, Brownsword R, Kariniotakis G, Denhard M, Draxl C. The state of the art in short-term prediction of wind power: a literature overview. 2nd ed. 2011. https://doi.org/10.13140/RG.2.1.2581.4485.

10. Akçay H, Filik T. Short-term wind speed forecasting by spectral analysis from long-term observations with missing values. Appl Energy. 2017;191:653-62. https://doi.org/10.1016/j.apenergy.2017.01.063.

11. PJM Data Miner. Data miner 2 [Internet]. 2021 [cited 2021 May 27]. Available from: https://dataminer2.pjm.com/list.

12. ENTSO-E. ENTSO-E transparency platform [Internet]. 2021 [cited 2021 May 27]. Available from: https://transparency.entsoe.eu/.

13. Hirth L, Mühlenpfordt J, Bulkeley M. The ENTSO-E Transparency Platform – A review of Europe's most ambitious electricity data platform. Appl Energy. 2018;225:1054-67. https://doi.org/10.1016/j.apenergy.2018.04.048.

14. Fung DS. Methods for the estimation of missing values in time series [dissertation]. 2006.

15. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York (NY): Springer; 2001.

16. Kim M, Park S, Lee J, Joo Y, Choi JK. Learning-based adaptive imputation method with kNN algorithm for missing power data. Energies. 2017;10(10):1668. https://doi.org/10.3390/en10101668.

17. Hsu HH, Yang AC, Lu MD. KNN-DTW based missing value imputation for microarray time series data. J Comput. 2011;6(3):418-25.

18. Cao W, Wang D, Li J, Zhou H, Li L, Li Y. BRITS: bidirectional recurrent imputation for time series. arXiv:1805.10572. 2018.

19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762. 2017.

20. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A transformer-based framework for multivariate time series representation learning. arXiv:2010.02803. 2020.

21. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1989;1(2):270-80. https://doi.org/10.1162/neco.1989.1.2.270.

22. Doubleday K, Van Scyoc Hernandez V, Hodge BM. Benchmark probabilistic solar forecasts: Characteristics and recommendations. Sol Energy. 2020;206:52-67. https://doi.org/10.1016/j.solener.2020.05.051.

23. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17). Red Hook (NY): Curran Associates Inc.; 2017. p. 6405-16.

24. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Amer Statist Assoc. 2007;102(477):359-78. https://doi.org/10.1198/016214506000001437.

25. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML '16). Vol. 48. JMLR.org; 2016. p. 1050-9.

26. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980. 2017.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

27. Nguyen X. Wasserstein distances for discrete measures and convergence in nonparametric mixture models. arXiv:1109.3250. 2011.

28. Meshgi K, Ishii S. Expanding histogram of colors with gridding to improve tracking accuracy. In: Proceedings of the 14th IAPR International Conference on Machine Vision Applications (MVA 2015). 2015. p. 475-9. https://doi.org/10.1109/MVA.2015.7153234.

29. Xu L, Wang Z, Yilmaz HU, Poganietz WR, Ren H, Guo Y. Considering the impacts of metal depletion on the European electricity system. Energies. 2021;14(6):1560. https://doi.org/10.3390/en14061560.

30. Fraunhofer ISE. Public net electricity generation in Germany in 2021: Renewables weaker due to weather [Internet]. 2022 [cited 2022]. Available from: https://www.ise.fraunhofer.de/en/press-media/news/2022/public-net-electricity-in-germany-in-2021-renewables-weaker-due-to-weather.html.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press