Deep Learning Models for Time Series Imputation: A Survey

Truong Tan Hung, Dang Xuan Hung, Tran Thang Dong
FPT University

***Abstract*—Missing or corrupted values are common in time series data. This survey reviews methods that aim to impute missing values and to enable reliable downstream tasks such as forecasting, classification, and event detection. We focus on both traditional and modern deep learning techniques including DTW-based methods, RNN variants, GAN-based models, hybrid pipelines and supporting preprocessing techniques.**

***Index Terms*—time series, imputation, DTW, GRU, LSTM, Transformer, deep learning, ensemble learning**

I. Introduction

Time series data appear widely across various domains such as energy, healthcare, finance, and environmental monitoring. In practice, they often contain missing values due to sensor failures, transmission errors, inaccurate measurements, harsh weather conditions, or manual maintenance. Missing data can be highly prevalent, sometimes reaching up to 90% in certain datasets.

Handling missing values is a crucial preprocessing step because most statistical and machine learning algorithms for time series require complete data. Ignoring or removing missing values can lead to loss of valuable information, biased or unreliable results, and incorrect interpretations. Therefore, imputation techniques are necessary to estimate missing values in a reasonable way while preserving the integrity of the dataset.

Traditional imputation methods are simple and fast but often fail to capture complex temporal dependencies. Recent advances in deep learning, particularly recurrent neural networks (RNNs), offer significant potential as they can exploit hidden temporal patterns within sequential data. These models have demonstrated substantial improvements in imputation tasks, especially for multivariate time series. This survey aims to provide a comprehensive evaluation and comparison of various deep learning architectures (such as GRU, LSTM, and Transformer) for time series imputation, highlighting the trade-off between model complexity and imputation accuracy.

II. Project Overview and Motivation

The main focuses of this study are:

- Imputing missing values in time series,
- Forecasting future values,
- Classification and event detection on time series.

The goal is to summarize state-of-the-art methods, discuss their strengths and limitations, and provide practical guidance for future implementation and experiments.

III. Objectives

At this stage, the objectives are:

- Collect and describe state-of-the-art methods for time series imputation.
- Explain how each method handles missingness and irregular sampling.
- Provide a reference baseline for later implementation and evaluation.
- Evaluate and compare the performance of various deep learning models, including conventional and hybrid architectures, to better understand the trade-off between model complexity and imputation accuracy.

IV. Dataset and Preprocessing

The experimental dataset used in this study is the Hanoi water level dataset, which contains water level measurements over an extended period. The dataset consists of two main columns:

- **Waterlevel:** A complete reference series with no missing values.
- **Average:** The primary target variable for imputation, containing numerous missing entries (represented as "NA" or blank cells).

A. Data Preprocessing Steps

The following preprocessing steps were applied to ensure data quality and consistency:

1) **Missing Value Parsing:** All non-numeric entries, "NA" values, and blank cells were converted to NaN for uniform treatment.
2) **Outlier Detection and Correction:** Visual inspection and statistical methods were used to identify and remove or correct anomalous values.
3) **Normalization:** MinMaxScaler was applied to normalize all feature columns to the range [0, 1], facilitating stable training for machine learning and deep learning models.
4) **Missing Value Mask:** A binary mask was generated to track the positions of missing values in the "Average" column. This mask is critical for fair evaluation, ensuring that performance metrics are computed only on originally missing positions.
5) **Gap Distribution Analysis:** The dataset exhibits a missing rate exceeding 10%, with gaps ranging from isolated points to long consecutive blocks. Histogram and time series plots reveal both scattered and structured missingness patterns, typical of real-world sensor networks.

V. Baseline Experimental Results

To establish a comprehensive baseline for comparison with deep learning models, we first evaluated classical machine learning methods and pattern-based DTWBI on the Hanoi dataset. All models were trained on available (non-missing) data and evaluated exclusively on the masked missing positions.

A. Performance Comparison of Baseline Models

Table I summarizes the imputation performance of baseline models using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and Similarity metrics.

Table I
Baseline Model Performance on Hanoi Dataset

Model	MAE	RMSE	R^2
Linear Regression	0.7785	1.0898	0.9999
KNN	1.5146	2.1837	0.9998
SVM	0.7946	2.2987	0.9998
Decision Tree	0.0604	0.2681	1.0000
Bagging	0.6730	1.0454	1.0000
Random Forest	0.6419	0.9820	1.0000
Extra Trees	0.0912	0.1374	1.0000
AdaBoost	8.1616	10.4113	0.9953
Gradient Boosting	0.4888	0.6051	1.0000
XGBoost	0.9622	1.5714	0.9999
Voting	0.7581	1.0009	1.0000
DTWBI	0.1276	97.19	–

B. Detailed Analysis of Baseline Results

a) Top Performers: Tree-Based Ensembles: **Decision Tree** achieved the lowest MAE (0.0604) and RMSE (0.2681) with perfect R^2 (1.0000), demonstrating exceptional fitting to the training data patterns. However, single decision trees are prone to overfitting, which may limit generalization to unseen gap patterns.

Extra Trees also performed remarkably well with MAE of 0.0912 and RMSE of 0.1374. The randomization in Extra Trees helps reduce overfitting compared to standard decision trees while maintaining high accuracy on local patterns.

Gradient Boosting showed strong and stable performance (MAE: 0.4888, RMSE: 0.6051), effectively capturing both linear trends and non-linear patterns through iterative boosting. This method is particularly robust for structured gaps.

b) Moderate Performers: Ensemble and Regression Methods: **Random Forest** and **Bagging** both achieved competitive results (MAE around 0.64-0.67), benefiting from ensemble averaging to reduce variance and improve stability across diverse gap types.

Linear Regression, **SVM**, and **Voting Regressor** showed moderate performance with MAE in the range of 0.76-0.95. These methods capture global trends well but struggle with highly non-linear local patterns, especially in long consecutive gaps.

c) Poor Performers: **AdaBoost** exhibited the worst performance (MAE: 8.16, RMSE: 10.41), likely due to its sensitivity to outliers and noisy patterns in the water level data. The sequential boosting strategy may have amplified errors from difficult samples.

KNN also underperformed (MAE: 1.51) due to the curse of dimensionality and difficulty in finding truly similar temporal neighbors in high-dimensional time series contexts.

d) Pattern-Based Method: DTWBI: **DTWBI** achieved excellent MAE (0.1276) and high similarity score (0.8959), indicating strong preservation of sequence shape and frequency characteristics. However, the extremely high RMSE (97.19) suggests occasional catastrophic failures on certain complex gaps where pattern matching fails. This highlights the trade-off between shape preservation and point-wise accuracy in DTW-based methods.

C. Key Observations

- Tree-based ensemble methods consistently outperform traditional linear models, demonstrating the importance of capturing non-linear temporal patterns.
- Perfect or near-perfect R^2 scores across most models suggest strong correlation between features and imputed values, though caution is needed regarding potential overfitting.
- Pattern-based DTWBI excels in structural fidelity but requires careful handling of edge cases to avoid large errors.
- These baseline results provide a robust benchmark for evaluating deep learning models, which are expected to further improve by learning complex temporal dependencies.

VI. Literature Review

A. Phan et al. (2017): DTWBI

Phan et al. proposed Dynamic Time Warping-Based Imputation (DTWBI) for univariate time series with long consecutive gaps. **Pipeline:** transform series into derivative space (DDTW), build a query before the gap, search for similar subsequences using cosine similarity and DTW cost, and fill the gap using the aligned continuation. **Dataset:** environmental and energy series with synthetic missing gaps. **Results:** DTWBI outperformed interpolation and Time Window Interpolation (TWI), particularly when missing occurred in the middle of the sequence. **Strengths:** preserves shape and seasonality. **Limitations:** computationally expensive.

B. Phan et al. (2019): eDTWBI

An extension of DTWBI that leverages both pre-gap and post-gap queries. **Pipeline:** build queries before and after the gap, find best matches using DDTW and feature extraction, and impute the gap by averaging aligned continuations. **Dataset:** CO_2, humidity, and temperature series. **Results:** eDTWBI achieved lower RMSE and better frequency preservation than DTWBI and seven other baselines. **Strengths:** uses both past and future context. **Limitations:** unsuitable for boundary gaps.

C. Bigand (2018): Elastic Matching Framework

Bigand extended DTW-based methods to multivariate and uncertain time series. **Pipeline:** fuzzy similarity measures (cosine, Euclidean, Sim) combined with DTW alignment. **Applications:** classification and imputation on environmental and phytoplankton data. **Results:** improved accuracy compared to single-measure methods.

D. Phan et al. (2020): Red River Forecasting

A hybrid ARIMA + machine learning model for river water level forecasting. **Pipeline:** ARIMA modeled linear trends; ML models (KNN, RF, LSTM) handled residuals. DTWBI was used to fill long gaps. **Results:** hybrid models reduced forecasting errors relative to ARIMA alone. **Implication:** DTWBI contributed to robust forecasting under missing data.

E. Fang and Wang (2020): Deep Learning Survey

A comprehensive survey of deep learning-based imputation. **Models:** GRU-D (decay-based RNN), BRITS (bidirectional imputation), NAOMI (multi-resolution), GRUI-GAN, and E2GAN. **Findings:** deep models capture complex dependencies but require large datasets and high computation.

F. Li et al. (2022): Photovoltaic Forecasting

Hybrid framework for multi-step photovoltaic power forecasting. **Pipeline:** TimeGAN for augmentation, Soft-DTW K-medoids for clustering daily patterns, CNN-GRU for spatio-temporal modeling. **Dataset:** PV power station data. **Results:** achieved RMSE = 0.927 MW, outperforming conventional models. **Implication:** integrating DTW with deep learning improves robustness.

G. Tsalikidis et al. (2023): Energy Load Forecasting

Hybrid ensemble for one-step-ahead energy load forecasting. **Pipeline:** Level-1 learners (CatBoost, LGBM, RF, XGBoost, MLP, LSTM) combined by a voting regressor. Preprocessing included interpolation and feature engineering. **Results:** ensemble achieved MAPE = 5.39%, outperforming individual models. **Strengths:** stable and accurate. **Limitations:** complex and resource-intensive.

VII. Methodology

This section details the Dynamic Time Warping (DTW)-based imputation methods and deep learning architectures employed in this study.

A. Dynamic Time Warping-Based Imputation (DTWBI)

The DTWBI method targets missing segments in univariate time series by dynamically aligning the query window, located near the missing data, with historical subsequences.

a) Core Idea: DTW computes the optimal matching path between two time series by minimizing the cumulative distance while allowing elastic shifts in the time axis. This makes it robust to temporal distortions often seen in real-world time series.

b) Detailed Steps:

1) **Transform to Derivative Space (DDTW):** The original time series is first transformed to utilize its derivative values, emphasizing the shape and dynamics rather than absolute magnitudes.
2) **Query Window Construction:** A fixed-length window immediately preceding the missing gap is extracted as the query.
3) **Subsequence Search:** The algorithm searches over the entire historical data for similar subsequences matching the query, applying a combined metric of cosine similarity and DTW cost.
4) **Gap Reconstruction:** The missing segment is imputed using the aligned continuation of the closest matching subsequence.

c) Benefits and Challenges: DTWBI excels at preserving seasonal patterns and shapes in time series, especially when data exhibits recurrent behaviors. However, it involves computationally expensive pairwise comparisons and depends on periodicity in the data for optimal performance.

B. Effective DTW-Based Imputation (eDTWBI)

eDTWBI refines DTWBI by incorporating dual query windows, leveraging both past and future information adjacent to missing segments to improve imputation accuracy.

a) Conceptual Overview: Instead of relying solely on a single pre-gap query, eDTWBI constructs two queries—one before and one after the gap—capturing the dynamic behavior surrounding the missing data on both sides.

b) Methodological Steps:

1) **Dual Query Construction:** Separate the full time series into two segments at the gap: pre-gap (Da) and post-gap (Db). Two fixed-length query windows are created: Q_b from Da, and Q_a from Db.
2) **Similar Subsequence Retrieval:** Using the DTWBI method, separately find subsequences most similar to Q_b within Da and to Q_a within Db, denoted Q_{bs} and Q_{as}.
3) **Imputation by Averaging:** The missing segment is reconstructed by averaging the aligned continuations derived from Q_{bs} and Q_{as}, balancing past and future context.

c) Advantages and Limitations: This dual query approach reduces estimation bias inherent in single-sided DTWBI, better preserves signal frequency characteristics, and provides more stable imputations. It may, however, be less effective near the boundaries where one query side is insufficient.

C. Deep Learning Architectures

Deep learning models for time series imputation leverage recurrent and attention-based architectures to capture long-range dependencies and complex temporal patterns.

a) GRU (Gated Recurrent Unit): GRU is a simplified variant of LSTM that uses gating mechanisms to control information flow. It is computationally efficient while maintaining strong performance on sequence modeling tasks. For imputation, GRU processes the time series with missing values and learns to predict missing points based on observed context.

b) LSTM (Long Short-Term Memory): LSTM networks use memory cells and three gates (input, forget, output) to maintain and update long-term dependencies. This architecture is particularly effective for capturing complex temporal patterns in time series with irregular missing patterns.

c) Transformer: Transformer models use self-attention mechanisms to weigh the importance of different time steps, enabling parallel processing and capturing global dependencies. Recent adaptations for time series imputation show promise in handling multivariate and long-sequence data.

d) Implementation Details: All deep learning models are implemented with:

- Sliding window approach for sequence input
- Masked loss functions focusing on missing positions
- Adam optimizer with learning rate scheduling
- Early stopping based on validation loss
- Hyperparameter tuning via grid search

VIII. Deep Learning Experimental Results

This section presents the core experimental results for deep learning models, which represent the main focus of this course project.

A. Performance of Deep Learning Models

Table II
Deep Learning Model Performance (To Be Completed)

Model	MAE	RMSE	Similarity
eDTWBI	(pending)	(pending)	(pending)
GRU	(pending)	(pending)	(pending)
LSTM	(pending)	(pending)	(pending)
Transformer	(pending)	(pending)	(pending)

B. Experimental Setup

Deep learning experiments are currently underway with the following configurations:

- **eDTWBI:** Implementation based on Phan et al. (2019) with dual query windows, cosine similarity threshold of 0.7, and k=2 best matches averaging.
- **GRU:** 2-layer architecture with 64 hidden units, dropout 0.2, sequence length 20, batch size 32.
- **LSTM:** 2-layer architecture with 64 hidden units, dropout 0.2, sequence length 20, batch size 32.
- **Transformer:** 4 attention heads, 2 encoder layers, feed-forward dimension 128.

Training is conducted with masked loss computation, focusing optimization only on missing positions to ensure fair comparison with baseline results.

C. Expected Performance

Based on literature review and preliminary experiments, we anticipate that:

- eDTWBI will achieve NMAE < 0.15 and similarity $>$ 0.90, improving upon DTWBI's structural preservation.
- GRU and LSTM will achieve MAE < 0.30, leveraging temporal context learning to outperform classical regressors.
- Transformer may require larger datasets and longer training but has potential for best overall performance on complex patterns.

Complete results, analysis, and comparison will be provided upon completion of training and validation phases.

IX. Discussion

The baseline experimental results establish that tree-based ensemble methods provide excellent accuracy for the Hanoi water level imputation task, with Extra Trees and Gradient Boosting achieving the best balance of accuracy and robustness. DTWBI demonstrates strong pattern preservation capabilities but requires careful tuning.

Deep learning models (GRU, LSTM, Transformer) are expected to further advance performance by learning complex temporal dependencies that classical methods cannot capture. These models are particularly promising for:

- Long consecutive gaps where local pattern matching fails
- Non-stationary time series with evolving dynamics
- Multivariate settings with cross-variable dependencies

Future work will integrate context features from DTW-based methods into deep learning architectures, explore attention mechanisms for adaptive pattern weighting, and extend experiments to multivariate time series and downstream forecasting tasks.

X. Conclusion

This study presents a comprehensive survey and experimental evaluation of time series imputation methods, from classical machine learning to modern deep learning architectures. Baseline experiments on the Hanoi water level dataset demonstrate that tree ensemble methods and pattern-based DTWBI provide strong performance. The core deep learning experiments (eDTWBI, GRU, LSTM, Transformer) are currently under completion and will provide insights into the effectiveness of neural architectures for capturing complex temporal patterns in missing data restoration.

The methodological foundation established through literature review and baseline benchmarking provides a solid framework for future research in time series imputation and related sequential data analysis tasks.

Acknowledgment

The authors gratefully acknowledge the guidance and support from faculty and peers at FPT University.

References

[1] T.-T.-H. Phan, É. P. Caillault, A. Lefebvre, and A. Bigand, "Dynamic time warping-based imputation for univariate time series data," *Pattern Recognition Letters*, 2017.

[2] T.-T.-H. Phan, É. P. Caillault, A. Lefebvre, and A. Bigand, "Effective dynamic time warping-based imputation for univariate time series," in *Proc. ICCSAMA*, 2019, pp. 310–324.

[3] A. Bigand, "Elastic matching for classification and modelisation of time series with applications to environmental data," Ph.D. dissertation, Université du Littoral Côte d'Opale, 2018.

[4] T.-T.-H. Phan, D.-T. Nguyen, É. P. Caillault, and A. Bigand, "Hybrid ARIMA and machine learning models for water level forecasting in the Red River," *Environmental Modelling & Software*, 2020.

[5] Z. Fang and H. Wang, "Missing data imputation for time series: A deep learning survey," *arXiv preprint arXiv:2005.11856*, 2020.

[6] X. Li, W. Zhang, J. Zhao, and L. Wang, "Hybrid deep learning framework for multi-step photovoltaic power forecasting with data augmentation and clustering," *Energy Reports*, vol. 8, pp. 2329–2341, 2022.

[7] I. Tsalikidis, N. Gounaris, and P. A. Mitkas, "One-step-ahead energy load forecasting using hybrid ensemble learning models," *Energy Informatics*, vol. 6, no. 1, pp. 1–14, 2023.