# What are the most effective machine learning and statistical methods for data imputation in time-series datasets across different domains between 2020-2024?

Time-series data imputation between 2020-2024 was most effectively handled by a combination of RNNs for complex scenarios and KNN for simpler cases, with hybrid approaches offering balanced solutions.

## Abstract

This report analyzed 10 studies published between 2020-2024 that examined machine learning and statistical methods for time-series data imputation across healthcare, environmental monitoring, and industrial applications.

The reviewed studies found that deep learning methods, particularly Recurrent Neural Networks (RNNs), performed well in multiple scenarios, though effectiveness varied by context. Traditional methods like K-Nearest Neighbors (KNN) demonstrated comparable performance when handling datasets with low to moderate missing data rates.

The studies described hybrid approaches, such as MuSDRI, that combine statistical and machine learning techniques to address both short-term and long-term patterns in time series. Several studies utilized benchmarking tools like TSI-Bench to compare different imputation methods.

The reviewed literature indicated a consistent trade-off between imputation accuracy and computational requirements, with more complex methods generally requiring greater computational resources while potentially offering improved accuracy.

The studies suggested several areas for future research, including the development of adaptive imputation methods and standardized evaluation metrics. Additional cross-domain studies may help clarify how different imputation methods perform across various applications.

## Paper search

Using your research question "What are the most effective machine learning and statistical methods for data imputation in time-series datasets across different domains between 2020-2024?", we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 100 papers most relevant to the query.

## Screening

We screened in papers that met these criteria:

- **Time Series Data and Methods**: Does the study implement machine learning or statistical methods specifically for time-series data imputation?
- **Performance Evaluation**: Does the study include quantitative evaluation metrics and/or comparative analysis of the imputation performance?
- **Study Type and Methodology**: Is the study a primary research article, systematic review, or meta-analysis with clearly described methodology?

- **Empirical Implementation**: Does the study include practical implementation with clearly described implementation details?
- **Focus on Imputation**: Is the primary focus of the study on data imputation rather than real-time prediction or general data cleaning?
- **Missing Data Focus**: Does the study specifically address missing data imputation techniques rather than other data preprocessing tasks?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

## Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Imputation Method Type**:

  Identify and categorize the specific machine learning or statistical method used for data imputation. Look in the methods section for detailed descriptions. Classify the method into one of these categories:

- Machine Learning (specify type: e.g., RNN, Transformer, Deep Learning)
- Statistical Method (specify type: e.g., Kalman Filter, Matrix Factorization)
- Hybrid Method

If multiple methods are used, list all and indicate their specific roles in the imputation process. Be precise about the algorithmic approach and any unique characteristics of the method.

- **Data Characteristics and Imputation Context**:

  Extract details about the time-series dataset used in the study:

- Domain/Application (e.g., sensor data, financial, healthcare)
- Data Dimensionality (univariate, multivariate)
- Missing Data Characteristics:
  - Percentage of missing data
  - Pattern of missingness (random, systematic, block-wise)
  - Temporal resolution of the time series (e.g., hourly, daily, monthly)

If multiple datasets are used, provide information for each. If specific details are not explicitly stated, note this in your extraction.

- **Imputation Performance Metrics**:

  Identify and record all quantitative performance metrics used to evaluate the imputation method:

- Primary accuracy metrics (e.g., Mean Squared Error, Mean Absolute Error)
- Secondary performance indicators
- Comparative performance against other methods

Record the specific numerical values for these metrics. If multiple metrics are reported, list all with their corresponding values. If comparisons are made, note the methods compared and their relative performance.

- **Computational Considerations**:

Extract information about the computational aspects of the imputation method:

- Computational complexity
- Training time
- Inference/imputation time
- Hardware requirements (if specified)
- Any noted computational advantages or limitations

If specific numerical values are provided (e.g., processing time, computational overhead), record these precisely. If qualitative descriptions are used, quote or paraphrase the original text.

- **Unique Methodological Contributions**:

  Identify and describe any novel approaches or innovative techniques introduced in the imputation method:

- Specific algorithmic innovations
- Novel approaches to handling temporal dynamics
- Unique ways of addressing missing data challenges

Focus on methodological contributions that distinguish this approach from existing methods. Use direct quotes or paraphrasing to capture the essence of the innovation, and note how it addresses limitations in previous approaches.

# Results

## Characteristics of Included Studies

| Study | Full text retrieved | Study Design | Domain | Methods Evaluated | Missing Data Scenarios |
|---|---|---|---|---|---|
| Bansal et al., 2021 | Yes | Comparative study | Multiple (sensor, environmental, utility, sales) | DeepMVI (deep learning) | Missing Completely At Random, Missing Disjoint, Missing Overlapping, Black-out[^p1_q2_2] |

| Study | Full text retrieved | Study Design | Domain | Methods Evaluated | Missing Data Scenarios |
|---|---|---|---|---|---|
| Bella et al., 2024 | No | Comparative study | Environmental (air quality) | Mean, Mode, Median, Multiple Imputation by Chained Equations, Last Observation Carried Forward, K-Nearest Neighbors, Recurrent Neural Network, Long Short-Term Memory, Bidirectional Long Short-Term Memory, Gated Recurrent Unit | No mention found[^p2_q2_2] |
| Du et al., 2024 | No | Benchmark development | Multiple (not specified) | 28 deep learning algorithms | Diverse missingness scenarios[^p3_q2_2] |
| Kazijevs and Samad, 2023 | No | Review and benchmarking | Healthcare | Deep learning methods | Low missing rates, completely random[^p4_q2_2] |
| Khayati et al., 2020 | No | Experimental evaluation | Multiple (not specified) | 12 recovery algorithms | No mention found[^p5_q2_2] |

| Study | Full text retrieved | Study Design | Domain | Methods Evaluated | Missing Data Scenarios |
|---|---|---|---|---|---|
| Le et al., 2023 | No | Comparative study | Healthcare/Fitness | Mean, Median, Last Observation Carried Forward, K-Nearest Neighbors Imputation, Bidirectional Recurrent Imputation for Time Series, Transformer, Self-Attention-based Imputation for Time Series | Low to high missing rates[^p6_q2_2] |
| Niako et al., 2024 | No | Comparative study | Healthcare (blood pressure) | 10 imputation techniques | Missing Completely At Random, 10%, 15%, 25%, and 35% missingness[^p7_q2_2] |
| Porta et al., 2021 | No | Experimental study | Industrial and medical | Mixture of techniques | No mention found[^p8_q2_2] |
| Saad et al., 2020 | No | Comparative study | Multiple (not specified) | 7 deep learning methods, 3 machine learning ensembles | No mention found[^p9_q2_2] |
| Zhou et al., 2021 | No | Experimental study | Multiple (not specified) | MuSDRI (hybrid method) | No mention found[^p10_q2_2] |

Based on the information available in the abstracts and full texts we retrieved, we identified the following characteristics of the included studies:

- Study Design :
  - 5 comparative studies
  - 3 experimental studies
  - 2 benchmark studies (including 1 review and benchmarking study)
- Domain :

  - 6 studies focused on multiple domains
  - 3 studies focused on healthcare
  - 1 study focused on environmental data
  - We didn't find domain information in the abstract for 1 study
- Methods Evaluated :
  - Deep learning methods were the most common, with 37 instances across studies
  - We found mention of 10 machine learning methods and 8 traditional methods
  - 23 methods were not clearly specified in the abstracts or available full texts
  - 1 hybrid method was mentioned
- Missing Data Scenarios : We found specific information about missing data scenarios in 5 studies, while we didn't find mention of the scenarios used in the abstracts of 5 studies
- The studies varied widely in their approach, with some evaluating a single method (e.g., DeepMVI) and others comparing multiple techniques (up to 28 in one study)
- There was a notable focus on deep learning and machine learning methods for handling missing data in time series, particularly in healthcare and multi-domain contexts

## Comparative Effectiveness of Methods

### Traditional Statistical Methods

Traditional statistical methods serve as baselines for comparison with more advanced techniques in several studies:

- Bella et al. (2024) compared traditional methods like mean, mode, and median against more advanced techniques[^p2_q1_7]
- Le et al. (2023) included mean, median, and Last Observation Carried Forward in their comparative analysis[^p6_q1_9]
- Niako et al. (2024) evaluated several statistical methods including mean, Kalman filtering, and various interpolation techniques[^p7_q1_8]

While these methods are often outperformed by more sophisticated approaches, they remain relevant due to their simplicity and interpretability. For instance, Niako et al. (2024) found that mean imputation, Last Observation Carried Forward, and interpolation methods performed better for small rates of missingness.

### Machine Learning Approaches

Machine learning approaches, particularly K-Nearest Neighbors (KNN), have shown promising results in several studies:

- Bella et al. (2024) found that KNN imputation consistently outperformed other techniques across all models[^p2_q1_9]
- Le et al. (2023) reported that K-Nearest Neighbors Imputation was more efficient than some state-of-the-art methods at low to moderate missing rates (less than 30%)[^p6_q3_1]
- Niako et al. (2024) included KNN among the methods that performed well for small rates of missingness[^p7_q1_11]

These findings suggest that KNN, despite its relative simplicity compared to deep learning methods, can be highly effective for time series imputation, especially when the missing data rate is not too high. This insight highlights the importance of considering the missing data rate when selecting an imputation method, as simpler methods may outperform more complex ones in certain scenarios.

**Deep Learning Solutions**

Deep learning methods, particularly variants of Recurrent Neural Networks (RNNs), have emerged as powerful tools for time series imputation according to several studies in our review.

| Method Type | Average Performance | Computational Cost | Best-Suited Scenarios |
|---|---|---|---|
| DeepMVI | Reduces error by >50% in >half cases | Higher than simpler methods | Multidimensional time series, various missing patterns[^p1_q2_7] |
| Gated Recurrent Unit | Generally best for imputation | Varies based on time series type | Trend, seasonal, and combined time series[^p9_q5_3] |
| Long Short-Term Memory | Slightly better predictive accuracy | Higher than ARIMA | Small samples, complex time series[^p7_q4_3] |
| Bidirectional Recurrent Imputation for Time Series | Effective for time series imputation | No mention found | Multivariate time series[^p6_q2_7] |
| Transformer | Effective for time series imputation | No mention found | Multivariate time series[^p6_q2_8] |
| Self-Attention-based Imputation for Time Series | Best at higher missing rates ( 30%) | Reasonable execution time | High missing rates in multivariate time series[^p6_q2_9] |
| MuSDRI | Best performance among evaluated methods | No mention found | Time series with multiple seasonal patterns[^p10_q5_2] |

We analyzed 7 different method types for time series imputation:

- Performance varied across methods :
  - DeepMVI reduced error by >50% in more than half of cases
  - Gated Recurrent Unit was generally best for imputation
  - Long Short-Term Memory showed slightly better predictive accuracy
  - Bidirectional Recurrent Imputation for Time Series and Transformer were described as effective for time series imputation
  - Self-Attention-based Imputation for Time Series performed best at higher missing rates ( 30%)
  - MuSDRI had the best overall performance among evaluated methods
- Best-suited scenarios for these methods :
  - 3 methods were suited for multivariate time series
  - 1 method each was best for:

    - Multidimensional time series
    - Various missing patterns
    - Trend, seasonal, and combined time series
    - Small samples and complex time series
    - High missing rates
    - Multiple seasonal patterns
- Computational cost : We didn't find information about computational cost for 3 of the 7 methods in the abstracts or available full texts. For the others, costs ranged from "reasonable execution time" to "higher than simpler methods."

## Cross-Domain Performance Analysis

### Domain-Specific Effectiveness

Based on the studies we reviewed, the effectiveness of imputation methods appears to vary across different domains, suggesting the potential importance of considering domain-specific characteristics when selecting an imputation approach.

### Generalizability Findings

| Domain Type | Best Performing Methods | Performance Metrics | Limitations |
|---|---|---|---|
| Healthcare | Deep learning methods (cross-sectional and longitudinal imputation) | No mention found | Computationally expensive, requires high-performance computing resources[^p4_q4_1] |
| Environmental (Air Quality) | K-Nearest Neighbors | Mean Absolute Percentage Error: 0.186, Root Mean Square Error: 18.495 | Limited to one dataset (Beijing PM2.5)[^p2_q2_8] |
| Multivariate Time Series (General) | DeepMVI | >50% error reduction in >half cases | Slower than simpler methods[^p1_q4_4] |
| Wearable Device Data | K-Nearest Neighbors Imputation (low to moderate missing rates), Self-Attention-based Imputation for Time Series (higher missing rates) | Mean Absolute Error | Limited to one dataset (Crowd-sourced Fitbit)[^p6_q2_13] |
| Univariate Blood Pressure | Kalman smoothing, interpolation, moving average methods | No mention found | Limited to univariate time series[^p7_q2_8] |
| Multiple Domains | Gated Recurrent Unit variants | No mention found | Performance varies based on time series type[^p9_q5_9] |

| Domain Type | Best Performing Methods | Performance Metrics | Limitations |
|---|---|---|---|
| Multiple Domains | MuSDRI | No mention found | Limited evaluation (three real-world datasets)[^p10_q5_8] |
| Industrial and Medical | Mixture of techniques | No mention found | Limited details on specific datasets[^p8_q2_8] |
| Multiple Domains | Varies (no single method outperforms on all datasets) | No mention found | Limited comparability across datasets[^p5_q5_2] |
| Multiple Domains | 28 deep learning algorithms (TSI-Bench) | No mention found | Extensive evaluation, but specific results not provided[^p3_q5_1] |

Key insights from the cross-domain analysis:

- Diversity of best-performing methods : We found a variety of best-performing methods across different domains, suggesting that no single method is universally superior for time series imputation.
- Inconsistent reporting of performance metrics : Performance metrics were not consistently reported across studies, making direct comparisons challenging. Only 3 out of 10 studies provided specific performance metrics in their abstracts or available full texts.
- Domain-specific limitations : The limitations identified were often domain-specific, ranging from dataset constraints to computational limitations. This highlights the importance of considering these limitations when interpreting and generalizing results.
- Tailored approaches : The diversity in methods, metrics, and limitations across domains suggests that time series imputation techniques are often tailored to specific applications and datasets.
- Deep learning potential : While deep learning methods show promise across various domains, the effectiveness of simpler methods in specific scenarios suggests that a one-size-fits-all approach may not be optimal.
- Importance of missing data characteristics : Several studies emphasized the importance of considering the rate and pattern of missing data when selecting an imputation method, as performance can vary significantly based on these factors.

These findings underscore the complexity of selecting appropriate imputation methods for time series data across different domains. The effectiveness of a method appears to depend on various factors, including the specific domain, data characteristics, missing data patterns, and computational resources available.

# References

Agung Bella, Putra Utama, Wahyu Sakti, Gunawan Irianto, A. Wibawa, A. N. Handayani, and Amat Nyoto. "Improving Time-Series Forecasting Performance Using Imputation Techniques in Deep Learning." *2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, 2024.

J. Porta, Martín Ariel Domínguez, and Francisco Tamarit. "Automatic Data Imputation in Time Series Processing Using Neural Networks for Industry and Medical Datasets." *Symposium on Information Management and Big Data*, 2021.

Lien P. Le, Tu T. Do, and Thu Nguyen. "Data Imputation for Multivariate Time-Series Data." *International Conference on Knowledge and Systems Engineering*, 2023.

Maksims Kazijevs, and Manar D. Samad. "Deep Imputation of Missing Values in Time Series Health Data: A Review with Benchmarking." *Journal of Biomedical Informatics*, 2023.

Mourad Khayati, Alberto Lerner, and Zakhar Tymchenko. "Mind the Gap: An Experimental Evaluation of Imputation of Missing Values Techniques in Time Series," 2020.

Muhammad Saad, Lobna Nassar, F. Karray, and Vincent C. Gaudet. "Tackling Imputation Across Time Series Models Using Deep Learning and Ensemble Learning." *IEEE International Conference on Systems, Man and Cybernetics*, 2020.

Nicholas Niako, Jesus D. Melgarejo, Gladys E. Maestre, and Kristina P Vatcheva. "Effects of Missing Data Imputation Methods on Univariate Blood Pressure Time Series Data Analysis and Forecasting with ARIMA and LSTM." *BMC Medical Research Methodology*, 2024.

Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. "Missing Value Imputation on Multidimensional Time Series." *Proceedings of the VLDB Endowment*, 2021.

Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Fanxing Liu, Zepu Wang, Zina Ibrahim, et al. "TSI-Bench: Benchmarking Time Series Imputation." *arXiv.org*, 2024.

Yujue Zhou, Jie Jiang, Shuanghua Yang, Ligang He, and Yulong Ding. "MuSDRI: Multi-Seasonal Decomposition Based Recurrent Imputation for Time Series." *IEEE Sensors Journal*, 2021.