International Journal of Computing and Digital Systems
2025, VOL. 17, NO.1, 1-17

http://dx.doi.org/10.12785/ijcds/1571041166

A Systematic Review of Recurrent Neural Network Adoption in Missing Data Imputation

Nur Aqilah Fadzil Akbar[1], M. Izham Jaya*[1], Mohd Faizal Ab Razak[1] and Nurul Aqilah Zamri[1]

[1]*Faculty of Computing,Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia*

Received 21 June 2024, Revised 4 February 2025, Accepted 8 February 2025

Abstract: Missing data is a pervasive challenge in diverse datasets accross various domains. It is often resulting from human error, system faults, and respondent non-response. Failing to address missing data can lead to inaccurate results during data analysis, as incomplete data sequences introduce biases and compromise the distribution of the synthesized data, and cause a negative impact on the decision-making process. Over the past decade, deep learning methods, particularly Recurrent Neural Network (RNN), have been employed to tackle the problem. This study aims to comprehensively evaluate recent RNN methods for missing data imputation, focusing on their strengths and weaknesses to provide a detailed understanding of the current landscape. A systematic literature review was conducted on RNN-based data imputation methods, covering research articles from 2013 to 2023 that were identified in the SCOPUS database. Out of 362 relevant studies, 70 were selected as primary articles. The findings highlight that Long Short-Term Memory (LSTM) is the most adopted RNN method for data imputation due to its adaptability in processing data of varying lengths as compared to Gated Recurrent Units (GRU) and other hybrid methods. Performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Area Under the Receiver Operating Characteristic Curve (AU-ROC), Mean Squared Error (MSE), and Mean Relative Error (MRE) are commonly used to evaluate these models. Future development of a more robust RNN-based imputation methods that integrate optimization algorithms, such as Particle Swarm Optimization (PSO) and Stochastic Gradient Descent (SGD) will further enhance the imputation accuracy and reliability.

Keywords: Systematic literature review, missing values, data imputation, Recurrent Neural Network (RNN), data quality

1. Introduction

Grasping the complexities of data quality presents a significant challenge owing to its reliance on specific contexts and domains. The definition of data quality often revolves around its capability to fulfill user needs and its appropriateness for its designated purpose. These criteria resonate with established quality management principles [1], which underscore the criticality of defining and upholding quality standards to meet exacting consumer demands. Despite its paramount importance, the attainment of high data quality poses a formidable challenge due to the diverse quality dimensions of data across various applications.

Data completeness, within the framework of data quality dimensions, denotes the quantity of accessible data in a given dataset [2], assessed by the ratio of available data to total records [3]. An examination of data completeness underscores the critical challenge posed by missing data, a mainly noteworthy issue in real-world datasets, particularly those involving time-series data. Time-series models in machine learning are prone to encountering missing data owing to a range of factors, including human errors during data collection, system malfunctions, respondents' refusal to answer specific questions, withdrawal from study participation, and the inadvertent merging of disparate data sources [4].

Due to the various causes of missing data, it is an unavoidable issue in real-world scenarios. Missing data poses significant challenges in data analysis, leading to a degradation of accuracy and negatively impacting the decision-making process. This paper aims to achieve the following objectives which are to identify recent trends in data imputation using recurrent neural networks, evaluate the effectiveness of existing methods, highlight their strengths and weaknesses in addressing missing data, and provide insights for future research directions.

Missing mechanisms significantly impact on the efficacy and validity of imputation methods, [5] as a successful of the imputation method often depends on the underlying mechanism [6]. Many current imputation methods are developed under the assumption that missing data occurs

E-mail address: aqiladzan@gmail.com, izhamjaya@umpsa.edu.my, faizalrazak@umpsa.edu.my, nurulaqilah0012@gmail.com

based on Missing at Random (MAR) pattern [7]. While assuming that data is Missing Completely at Random (MCAR) provides a simple starting point for imputation analysis, this assumption is typically impractical for real-world scenarios, as missing data often arise from complex relationships among observed variables and may also be influenced by unobserved factors [5].

There are three types of missingness mechanism, which are MCAR, MAR and Missing Not Completely at Random (MNAR) [8]. MCAR occurs when the missingness has no connection to either observed or unobserved variables. MAR occurs when missingness is still random but is influenced by observed variables. In the case of MNAR, missingness depends on the unobserved value of the missing element itself [9].

Forecasting and classification tasks often experience a degradation in performance due to the influence of missing data within time-series datasets. The presence of missing data may introduce bias in parameter estimation and reduce the representativeness of the sample data [10], leading to disruptions in statistical analysis and hindering effective decision-making [11]. A 2015 diabetes management program in Australia serves as an illustration of these challenges, particularly in predictive analytics within healthcare. According to Liu et al. [12], the initiative aimed to enhance care for diabetic patients across 36 clinics by linking daily electronic medical record data and employing a centralized statistical engine to predict patients at risk of developing diabetes. However, incomplete data entry at the clinics ultimately led to the program's failure.

Therefore, the imperative of addressing missing data emerges as a critical consideration one that is frequently underestimated in the construction of a resilient time-series machine learning model [13]. Basic approaches like deletion can disturb the chronological continuity of the time-series dataset, leading to information loss and bias. Conversely, imputation entails more intricate procedures aimed at substituting missing data to uphold the integrity of the entire sequence.

Imputation method presents a means to fill in missing data within a dataset with the most pertinent value, potentially minimizing errors in subsequent time-series analyses. Recognizing that imputation has inherent limitations, it necessitates meticulous consideration due to the risk of introducing biases and inaccuracies, thereby compromising the credibility of data analysis results. The selection of an imputation method requires a data-driven assessment of factors such as the mechanism of missingness, data distribution characteristics, and research objectives [14].

Conventional approaches to time-series imputation using machine learning often rely on feature extraction prior to making predictions. However, this approach is limited in its ability to fully exploit the valuable information inherent in raw time sequence data [13]. An example of this challenge can be seen in the complexities of implementing imputation with K-Nearest Neighbors (KNN). Although KNN is a popular method for handling missing data, it is susceptible to drawbacks such as reduced accuracy and the potential introduction of spurious correlations, particularly in scenarios lacking genuine correlations [4].

Furthermore, conventional methods for addressing missing data in time-series datasets face substantial challenges when confronted with datasets characterized by a multitude of features or variables. Li et al. [15] argue that conventional imputation methods, such as KNN, begin to exhibit diminished performance and accuracy in high-dimensional data scenarios, particularly those involving datasets with a substantial number of columns or attributes. This emphasizes the necessity for more sophisticated methods that can directly harness the richness of unprocessed time-series data to improve predictive precision and reliability.

Numerous studies have explored a paradigmatic shift that integrates the imputation and prediction (I&P) processes within a unified imputation framework using Recurrent Neural Network (RNN). This progressive approach is exemplified by models such as AJ-RNN and LIME-RNN, which strive to concurrently tackle imputation and prediction tasks. Nevertheless, it is important to note that these methods often disregard horizontal correlations present within the time-series datasets. They primarily focus on the associations between an incomplete value and its nearest neighbors [16], overlooking the broader interrelationships among various variables at the same timestamp.

This study presents a critical examination of the recent landscape of RNN-based imputation methods for missing data in time-series datasets. The central focus of this evaluation is to illuminate the existing gaps by highlighting the strengths demonstrated by recent RNN-based imputation methods, while also addressing their limitations in handling missing data for time-series datasets. Through a comprehensive investigation, the objective is to provide a detailed understanding of the current landscape of RNN-based methods for missing data imputation in time-series datasets and contribute insights for advancing this field further.

The paper adopts a structured approach. Section 2 provides a review of related research to enhance comprehension of the field. Section 3 delineates the research methodology, offering a detailed overview of the study's execution. In Section 4, the empirical findings derived from the study are presented. Section 5 consolidates the research outcomes and suggests potential avenues for future exploration. Finally, Section 6 offers a robust summary, encapsulating the key insights and implications gleaned from the study.

2. Related Works

Liu et al. [17] conducted a rigorous analysis of diverse methodologies employed for data imputation within healthcare environments. The study meticulously assessed how

different data features exert a substantial influence on the selection and efficacy of the imputation algorithms. The effectiveness of imputation method is intricately linked to the degree of correlation among variables, highlighting the nuanced interplay between data structure and imputation performance. Deep learning imputation method such as RNN improve the ability to handle missing data specifically the imputation accuracy, time consumed and the computational cost [18].

On the other hand, RNNs are susceptible to the vanishing gradient problem, a phenomenon that hinders its ability to effectively learn long-term dependencies during the imputation process. To address this limitation, a more advanced method known as the Long Short-Term Memory (LSTM) has been adopted in RNN. The LSTM method enhances the RNN architecture by incorporating memory cells, enabling it to better capture and retain sequential information over extended periods, thus mitigating the vanishing gradient issue in time-series imputation. Cui et al. [19] introduced an innovative stacked model that combines bidirectional and unidirectional LSTM networks to predict the state of network-wide traffic. This hybrid LSTM model markedly enhanced prediction accuracy compared to traditional recurrent neural networks (RNNs), demonstrating its superior capability in capturing complex traffic patterns and dependencies. Shen et al. [20] proposed a graph attention recurrent neural network (GARNN), an LSTM-based imputation unit specifically designed for handling missing values in spatial-temporal data and outperformed other RNN-based imputation methods.

Mesquita et al. [21] conducted a literature survey on deep learning-based imputation methods for multivariate time series, emphasizing their importance in fields like healthcare and industry. The study highlights challenges posed by missing data which significantly impact forecasting and classification tasks. While various imputation techniques exist, comparative analysis across different missing data rates has been lacking. To address this gap, the study evaluates five deep learning-based imputation methods such as MRNN, US-GAN, GP-VAE, SAITS, and BRITS by using the Physionet Challenge 2012 dataset. The findings reveal that SAITS achieved the lowest average imputation error, whereas BRITS demonstrated lower error dispersion, underscoring the strengths of these approaches in handling missing MTS data.

Sun et al. [5] review the performance of deep learning imputation methods in comparison to conventional machine learning imputation methods such as MissForest and Multiple Imputation by Chained Equations (MICE). The study concludes that deep learning approaches demonstrate superior accuracy for imputing data with high proportions of missing values. Furthermore, deep learning imputation methods are adept at handling both temporal and spatial missing data, offering a more robust solution for complex imputation tasks. In contrast, conventional imputation methods are preferable for imputation in datasets with small sample sizes.

Kazijevs and Samad [22] conducted a comprehensive survey of deep learning approach for imputing missing values in time-series data. The study indicates that deep learning imputation methods, particularly LSTM, significantly enhance imputation performance for time-series datasets. Despite being more computationally demanding than conventional methods, LSTM-based imputation offers superior accuracy and robustness, making it a valuable tool for handling complex temporal data with missing values.

Deep learning based imputation methods, particularly RNN and LSTM, outperform traditional techniques in accuracy and robustness, especially for time-series and spatial-temporal data. Advanced architectures like hybrid LSTM and attention-based networks further enhance imputation performance. However, these methods are computationally intensive and require large datasets, while traditional methods remain preferable for smaller datasets and lower-resource settings.

This study builds on prior work by providing a structured comparison of deep learning imputation method, specifically RNN and give insight of the suitable method based on dataset characteristics to perform data imputation.

3. Methodology

Following the established methodology for systematic literature reviews as outlined by Kitchenham et al. [23], this study adopts a rigorous and transparent approach to critically evaluate the relevant literature. This comprehensive approach involves eight distinct stages, each meticulously designed to ensure thoroughness and reliability. These stages are summarized and illustrated in Figure 1, providing a clear framework for the review process.

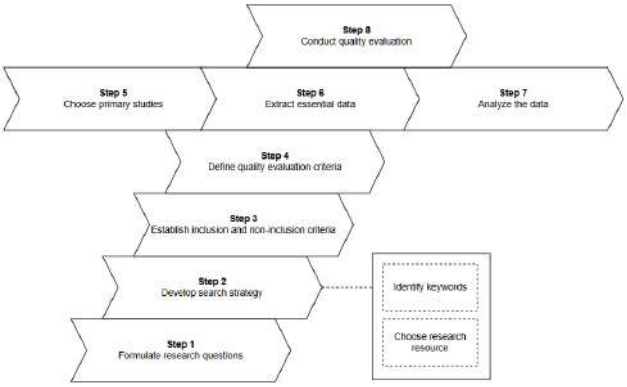

Figure 1. Framework for Review Process

The initial stage involves formulating research questions. These questions serve as the lens through which relevant data will be extracted and analyzed from the chosen primary studies throughout the review process. The subsequent stage revolves around developing a search strategy,

encompassing two key elements: identifying study-related keywords and selecting reputable and esteemed research resources, such as journals and conference proceedings, to serve as sources for the relevant studies. Logical operators will also be employed to facilitate the search process.

The third stage entails establishing criteria for inclusion and non-inclusion, ensuring that only pertinent studies are selected. The fourth stage entails defining criteria for evaluating study quality. By adhering to this criterion, it ensures that selected studies have the requisite information to address the research question and minimize potential biases.

The fifth stage involves selecting the primary studies, initially by examining their titles and abstracts to ascertain alignment with the outlined criteria before proceeding to a full-text evaluation. The subsequent stage entails extracting all essential data, which will then be subjected to analysis and synthesis to address all research inquiries. The final stage encompasses evaluating all selected primary studies against established quality standards. This systematic evaluation ensures adherence to predetermined standards and assigns dedicated marks to each criterion, ultimately leading to an overall quality evaluation for each study.

A. Research Question

This study provides a comprehensive review and evaluation of the adoption of RNN for imputing missing data in time-series datasets. The primary objective is to analyze and synthesize the effectiveness of RNN in addressing the challenges associated with missing data imputation. Emphasis is placed on identifying key performance metrics that contribute to achieving optimal imputation results. This systematic literature review aims to bridge existing research gaps by addressing the following research questions:

RQ1 How have RNN imputation methods evolved in the last decade?

RQ2 What are the most popular variants employed in RNN for handling missing data imputation?

RQ3 What are the features of the datasets utilized in RNN data imputation?

RQ4 What performance metrics are utilized for evaluating RNN imputation method?

RQ5 What are the advantages and potential drawbacks of the proposed RNN imputation method?

RQ1 explores the emerging trends in the adoption of RNNs for imputing missing values, highlighting the frequency and prevalence of RNN usage compared to other imputation methods. RQ2 examines the implementation processes and categorizes the various RNN imputation methods employed. RQ3 aims to analyze the features of the dataset used for RNN imputation. RQ4 investigates the metrics used to evaluate the performance of RNN imputation methods, providing insights into the key metrics that are essential for assessing their effectiveness. RQ5 focuses on identifying the strengths and limitations inherent in RNN methods, offering valuable guidance for future research and applications, particularly in the context of data imputation using RNNs.

B. Search Strategy

The search strategy is developed based on two primary components: the identification of relevant keywords and the selection of esteemed research resources. The selection of classifications and keywords for the search process is derived from a thorough examination of abstracts and research titles of sample literature deemed pertinent to the research questions. This study employs two primary classifications: ”missing values” and ”RNN” which ensure a focused and efficient search, capturing the most relevant studies for the review.

Table I provides a detailed categorization of the primary classifications used in this study, along with their associated keywords. To refine the search strategy, keywords related to ‘RNN’ are included, acknowledging that ’RNN’ may not be explicitly mentioned in all research titles and abstracts. Recognizing the diverse applications of RNN across various studies, a thorough identification of specific keywords relevant to missing data is conducted. This thorough approach allowed for the precise extraction of data from pertinent studies, thereby ensuring a comprehensive and focused analysis of literature.

TABLE I. Primary classification and associated keywords

Number	Classification	Associated Keywords
PC01	RNN	‘RNN’, ‘recurrent neural network’, ‘GRU’, ‘Gated Recurrent Unit’, ‘Long Short-Term Memory’, ‘LSTM’, ‘time-series’, ‘sequential.’
PC02	Missing data	‘missing data’, ‘data missingness’, ‘missing value’, ‘imputation’, ‘incomplete’

The decision to utilize the SCOPUS database as the primary source for the reviewed articles was based on several key considerations. Firstly, SCOPUS is a widely recognized database known for its wider journal coverage compared to Web of Science [24] that has extensive collection of peer-reviewed research articles, ensuring a comprehensive and reliable source of information [25]. Secondly, articles indexed in SCOPUS are considered to meet high-quality standards, having undergone rigorous quality assessments. Moreover, SCOPUS offers advanced search capabilities, including the use of logical operators such as OR and AND, which facilitated the refinement of search criteria by the

researchers. SCOPUS provides robust filtering features that allow for the narrowing down of search outcomes according to publication date, further enhancing the precision and relevance of the literature review.

C. Inclusion and Exclusion Criteria

A two-step process was thoroughly employed to identify relevant studies from the research database outlined in Table II.

TABLE II. Research database selection

Number	Database	Web Address
RN01	SCOPUS	https://www.scopus.com

The search process encompassed articles published between 2013 and 2023. Initially, a comprehensive search was conducted using predefined keywords and logical operators to ensure the inclusion of all potentially relevant articles. Subsequently, the titles and abstracts of these articles were meticulously screened and filtered to select studies that aligned with this research objectives. In the second iteration, full-text articles were reviewed in detail, and specific inclusion and exclusion criteria were applied to ensure a rigorous final selection of studies. This methodical approach significantly enhanced the validity and reliability of the research findings.

Inclusion criteria:

a. The study must have been published exclusively between January 1, 2013, and December 31, 2023.

b. The study's primary objective should revolve around resolving the problem of missing data within a dataset.

c. The proposed RNN method must be evaluated against other machine learning imputation methods.

d. The study must be composed in English to ensure clarity and accessibility for the research team.

e. The study must be published in a journal or conference proceedings indexed in SCOPUS, ensuring that it has undergone a rigorous peer-review process.

Non-inclusion criteria:

a) The study must not be a conference abstract or editorial, as these formats typically lack the comprehensive detail and rigorous methodology necessary for in-depth research analysis.

b) The study must refrain from prioritizing conventional imputation methods over RNN-based imputation.

c) The study's primary objective must be to improve data imputation performance, not to enhance other factors.

d) The study must focus on imputing missing values, not predicting a particular case.

This study employed a systematic search strategy to identify relevant articles for review. Table III delineates a comprehensive search query particularly crafted to harness the advanced search capabilities of the selected database. These queries were designed to incorporate relevant keywords and logical operators, thereby optimizing the search process to yield a precise and comprehensive selection of articles aligned with the research objectives.

TABLE III. Selected database search query

Database Name	Search Query
SCOPUS	(”missing data” OR “missing value” OR “data missingness” OR incomplete OR “imputation”) AND (”recurrent neural network” OR “RNN” OR “GRU” OR “Gated Recurrent Unit” OR “Long Short-Term Memory” OR “LSTM”) AND (“time-series”, “sequential”)

D. Quality Criteria

Prior to delving into the research questions, this section systematically evaluates the selected studies to confirm the requisite depth and detail of the selected articles for a thorough analysis. Each evaluation criterion is denoted by the abbreviation 'QAC', which indicates Quality Assessment Criteria. These criteria encompass a set of evaluation questions outlined as follows:

QAC1 Does the study implement RNN method for missing data imputation?

QAC2 Does the study clearly explicate their methodology and research purpose?

QAC3 Is the recurrent neural network approach assessed through comparisons against other RNN or machine learning-based methods?

QAC4 Are the performance metrics employed in the study clarified and explained by the researcher?

QAC5 Does the study elaborate on the comparative strengths and weaknesses of the recurrent neural network approaches employed?

E. Identification of Primary Articles Collection

The automated search conducted on the SCOPUS online database yielded a total of 362 articles centered on the application of RNN for missing value imputation. After a thorough assessment of their titles and abstracts, only 89 articles were considered pertinent during the initial screening. The remaining 19 studies were excluded due to their lack of relevance to the current investigation. Adhering strictly to the predetermined inclusion and exclusion criteria, a final

selection of 70 articles was made to specifically address the research inquiries, as depicted in Figure 2.

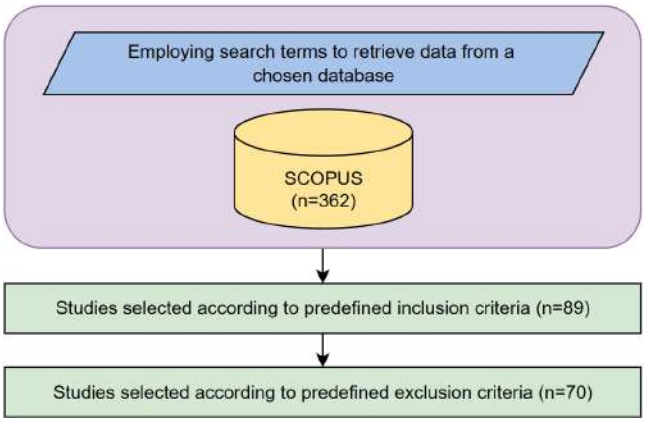

Figure 2. Primary Studies Screening Process

F. Data Extraction

The process of selecting pertinent articles involved an intensive search across various platforms including selected online database, reputable journal publishers, and relevant conference proceedings. Each article identified underwent a stringent evaluation process, following the categorization and subcategorization framework delineated in Table IV. These categories were derived from the research questions to ensure harmonization with the study's primary objectives. The data extraction process employed a blend of automated and manual search techniques, emphasizing a comprehensive and thorough approach to gather relevant information from the selected articles.

TABLE IV. Category and subcategory framework

Category	Sub-category	RQ
Paper information	Article's title Published year Author Publisher	RQ1
Research focus	Objective Methodology	RQ2
RNN method	Proposed method Selection method	RQ2
Dataset	Number of dataset Dataset sources Type of missing mechanism	RQ3
Performance evaluation	Evaluation metrics Evaluation methods	RQ4
Findings	Limitation Strength Future work	RQ5

The data extracted under the "Paper information" category was instrumental in addressing RQ1, which aimed to ascertain whether the articles were published within the past decade. This categorization, based on the publication year, offered valuable insights into the current trends and advancements in RNN-based missing data imputation, thereby significantly contributing to the resolution of RQ1. Moving on to the "Research focus" and "RNN method" category, a detailed examination of the specific RNN methods employed in the articles was conducted. Through the analysis of subcategories within this category, RQ2, which sought to identify the predominant RNN methods utilized for missing data imputation, could be effectively addressed, enriching the study's overall findings.

The 'Dataset' category aims to analyze the features of each dataset used in the primary studies. This approach provides a deeper understanding of how RNN imputation methods are applied to different datasets, representing the adoption of these methods in real-world scenarios, addressing RQ3. The "Performance evaluation" category played a pivotal role in addressing RQ4 by evaluating the metrics used to assess the effectiveness of the proposed RNN methods. This critical assessment of performance factors contributed significantly to understand the efficacy and reliability of RNN-based methods in handling missing data, aligning with the objectives of RQ4.

Lastly, the "Findings" category provided a comprehensive overview of the strengths and limitations of the proposed methods. These insights were crucial in addressing RQ5, which aimed to identify research gaps and potential avenues for future work in the domain of missing data imputation using RNNs. By pinpointing these gaps, the study not only contributed to the existing body of knowledge but also provided valuable guidance for researchers looking to advance this field further.

G. Data Synthesis

In the data synthesis phase, the extracted data from the preceding stages are combined to provide a comprehensive analysis. This integration employs two distinct approaches: quantitative descriptive analysis (QDA) and narrative synthesis. QDA is utilized to present a deeper understanding to research questions 1, 2, 3 and 4. In contrast, RQ5 is addressed through a narrative synthesis approach, which involves summarizing and integrating insights from various articles.

H. Data Quality Assessment

After applying rigorous inclusion and exclusion criteria, each article selected for inclusion underwent a comprehensive assessment using QAC. This step ensured that the selected articles aligned with the research questions and that any low-quality articles, which might introduce bias into the results, were systematically excluded [26]. This quality scoring framework assigned numerical values from 0 to 1 to assess the sufficiency of information available to address key domains outlined in the QAC: A marks of 1

suggest that an article completely fulfilled the criteria in question and provided abundant salient details; 0.5 denoted that the article only partially met the criteria or omitted some relevant information; and 0 signified that the article failed to address the criteria or research considerations at all. By tallying the quality scores achieved across all QAC dimensions, an aggregated quality benchmark could be calculated for each selected source.

Once evaluation of quality is performed, it can be observed that the total score for each QAC is predominantly greater than or equal to 80%, as illustrated in Figure 3. This signifies that the chosen relevant articles sufficiently cover the necessary information regarding the imputation of missing values using RNN.

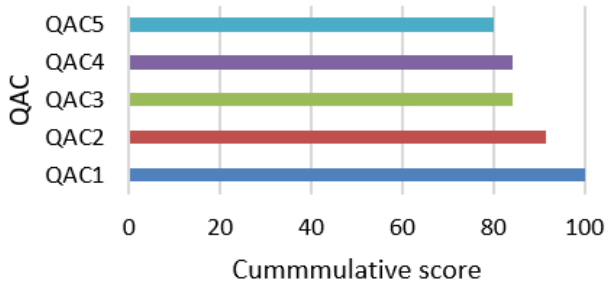

Figure 3. Total Score for Each QAC

4. Result and Discussion

This study places particular emphasis on articles concerning missing data imputation using the RNN method. Each selected article underwent a thorough analysis and data synthesis process, as detailed in the preceding section. The findings from this systematic literature review will be summarized to provide comprehensive answers to the research questions that are previously identified.

A. *RQ1. How Have RNN Imputation Methods Evolved in The Last Decade?*

Figure 4 depicts the annual publication count of research articles employing RNNs for missing value imputation. Interestingly, between 2013 and 2017, research remained very low and stable, with publication counts hovering between 0 and 1 paper per year. However, from 2018 to 2021, a remarkable increase was observed from 2018 to 2021, reaching 22 papers in 2021. As deep learning continues to gain traction, so does the adoption of RNNs for data imputation, with research in this area seeing a significant surge. The rise in deep learning research can be attributed, in part, to the emergence of high-level neural network Application Programming Interfaces (APIs) like TensorFlow and Keras. These user-friendly interfaces have addressed the computational hurdles of training deep learning models, as outlined by Ma et al. [27]. Their functionalities, such as automatic differentiation and efficient memory management, align with recommendations for optimizing training, including learning rate scheduling [28] and high-performance computing (HPC) communication techniques. In support of this, a SCOPUS database search reveals a substantial increase in studies in 2018 mentioning TensorFlow and Keras which are 500 and 111 articles respectively. Yet, the count has dipped slightly to less than 22 articles in 2022 and 2023, raising questions about the recent shift.

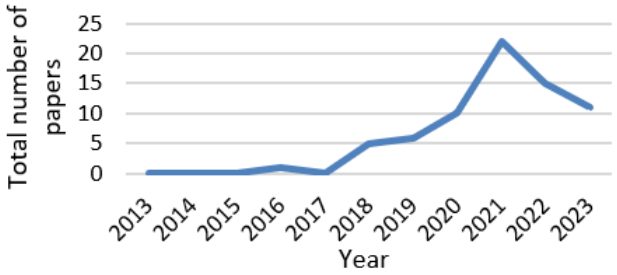

Figure 4. Year of Publication of Using RNN for Data Imputation

A large number of articles across 27 countries reveal a growing interest in RNNs for data imputation. As illustrated in Figure 5, the top five contributing nations stand out: China spearheads the field with 28 published journal articles and conference papers, closely followed by the United States with 23 articles. Australia and South Korea have 8 and 7 articles respectively, while Canada contributes 4 articles. This geographically diverse landscape of research underscores the increased popularity and efficacy of RNNs as a data imputation method, capturing the attention of researchers worldwide.

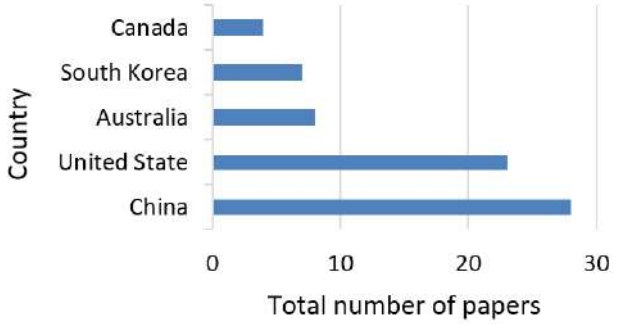

Figure 5. Top 5 Countries in RNN Data Imputation Research

Figure 6 illustrates the distribution of article types among research exploring missing data imputation using RNNs. Journal articles and conference papers emerge as the most prevalent form of publication, both accounting for 34 papers (49%) of the analyzed articles. Conference reviews and Review articles constitute a smaller portion, with 1 paper (1%) and 1 paper (1%), respectively. Figure 7 offers a more granular view, detailing the annual distribution of each document type.

Figure 7 illustrates the distribution of article types over the past decade for the RNN data imputation method. The data shows that prior to 2017, only conference articles

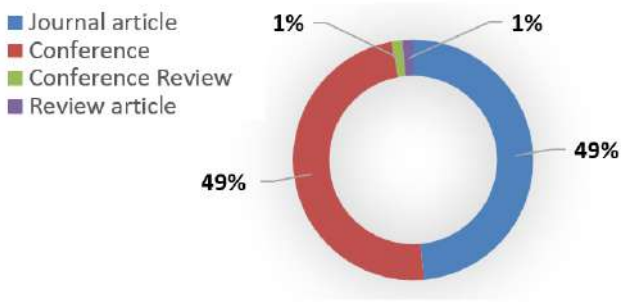

Figure 6. Document Type of the Primary Studies

were published, and in very low numbers. From 2018 to 2020, the number of conference articles increased and was higher than journal articles. However, starting in 2021, the number of journal articles has equaled or exceeded that of conference articles. This shift aligns with the findings from surveys conducted by Saydam et al. [29] revealing that researchers prioritize submitting to journal publications due to perceived reputation, prestige, and the assurance of undergoing a peer-review process—a crucial element in upholding research quality standards. Despite the abundance of articles and conference papers on this topic, there is a notable shortage of review papers. Only one review paper has been published, indicating a gap in the synthesized understanding of the existing literature.

Figure 8 depicts the top 5 subject areas (out of a total of 18) that encompass articles employing RNNs for data imputation. Within these top 5, computer science emerges as the predominant field, comprising 66 documents of the analyzed articles. Engineering follows with a substantial 32 articles, while mathematics contributes 20 articles. Medicine and decision sciences, with 14 articles and 11 articles respectively, round out the top 5.

B. *RQ2.What are the Most Popular Variants Employed in RNN for Handling Missing Data Imputation?*

Figure 9 illustrates the distribution of RNN variants employed for missing data imputation. LSTM variant dominates, accounting for 42% of utilized RNN variants across 30 articles. Following closely are GRU, commanding 35% of the landscape, represented by 25 articles. Twenty-three percent (16 papers) are categorized as hybrid, that combines two or more types of RNN variants [30] including Saad et al. [31] which utilized LSTM and GRU for missing data imputation. It is noteworthy that LSTM is the most popular variant in missing data imputation using RNN as it can handle data with varying length [32]. Although GRU has a simpler architecture with fewer gates than LSTM, making it faster to train, it is deemed suitable primarily for small datasets [33].

Table V shows the distribution of 14 articles that incorporate bidirectional RNN approach. Bidirectional RNNs is able to analyze temporal dependencies in both directions and grant them an advantage in handling sequential

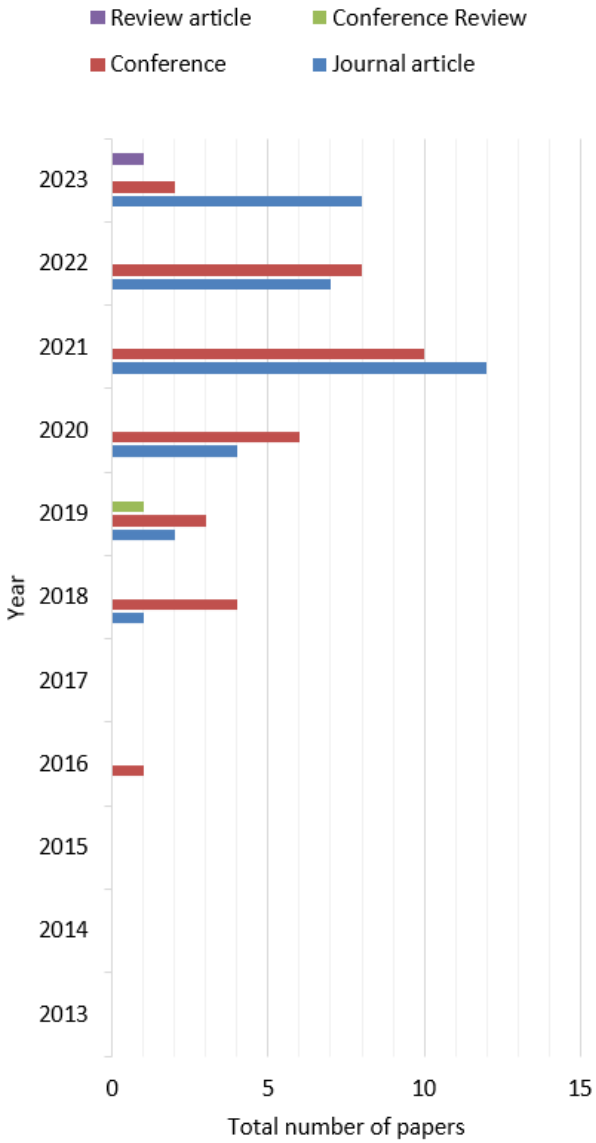

Figure 7. Document Type Trends of the Primary Studies

data, enabling them to effectively impute missing data by considering the broader context of the surrounding data points. This advantage is particularly beneficial compared to unidirectional RNNs, which only consider context from one direction. This trend is reflected in several recent studies employing bidirectional LSTMs for imputation tasks [34], [35], [19], [36], [37], [38], [39]. Bidirectional GRU architectures have also gained traction, as evidenced by several research [15], [40], [41], [42]. In line with Yang et al. [33], bidirectional approaches hold potential for accurate data imputation due to their ability to capture long-range context in sequential data.

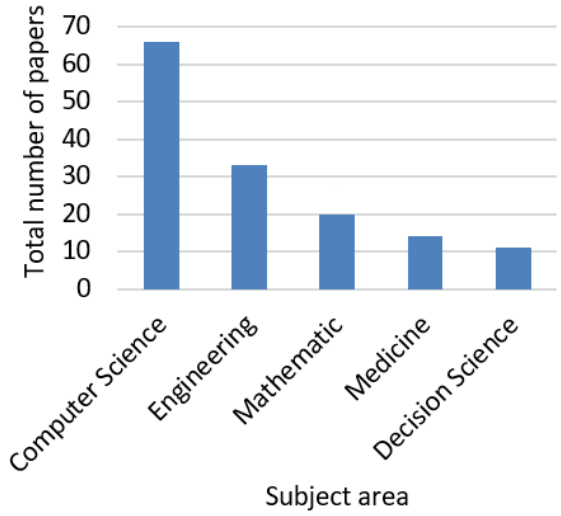

Figure 8. Top 5 Subject Area of Data Imputation Using RNN

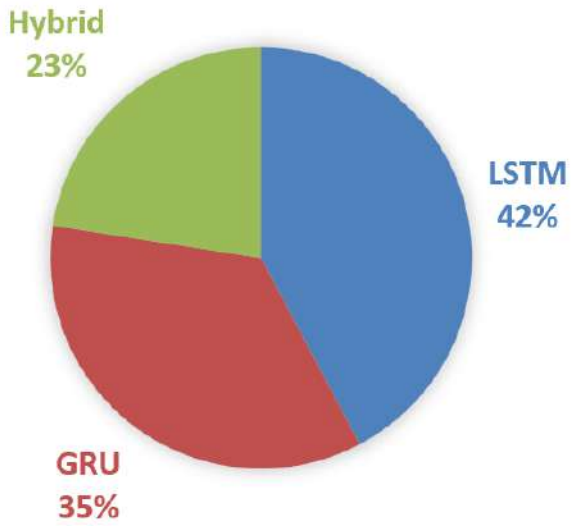

Figure 9. Prevalence of RNN Variants in Missing Data Imputation

C. *RQ3.What are The Features of The Datasets Utilized in RNN Data Imputation?*

The number of datasets used in each study is illustrated in Figure 10. The dataset counts range from 1 to 10. The most commonly used number of datasets per study is one or two, with each category comprising 22 studies. These datasets vary between synthetic and real-world data. A smaller proportion of studies, 12 studies utilized five or more datasets. Additionally, some studies did not specify the dataset count shown in the 'N/A' category.

Figure 11 illustrates the distribution of dataset domains used in primary studies on data imputation using RNN. The healthcare domain is the most frequently studied, accounting for 40% of the total, with data from 32 studies. This indicates a strong focus on healthcare data imputation

TABLE V. Number of bidirectional approach used in each type of RNN

RNN Variants	Number of Bidirectional Approaches
LSTM	7
GRU	4
Hybrid	3

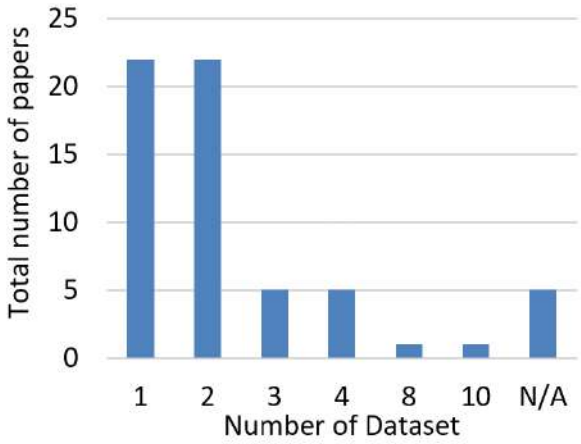

Figure 10. Total Number of Datasets Used in Primary Studies

in RNN research. The second most common domain is the environment, which makes up 20% of the studies, involving data from 16 studies. 21% of the datasets come from various other domains, also comprising 17 studies in total. The traffic domain follows, contributing 13% with 10 studies. Finally, 6% of the studies, representing 5 studies, did not specify the domain of their datasets, shown as "N/A" in the chart.

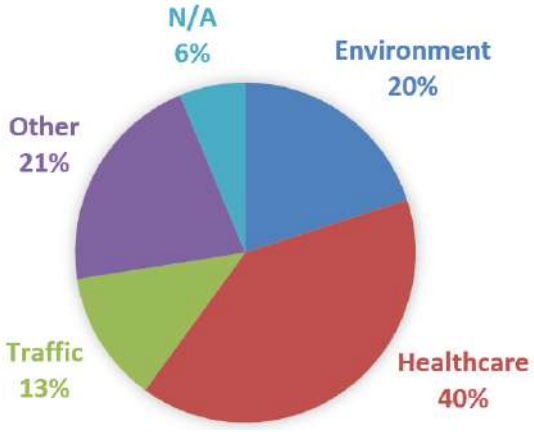

Figure 11. Distribution of Dataset Domains Used in Primary Studies

Table VI provides an insight of variety of dataset sources or repositories used in the primary studies of imputation using RNN. The most frequently used repositories are PhysioNet, which contains 10 studies. Followed by ADNI

and MIMIC, with 9 studies utilizing datasets from each, respectively. UCI repository was referenced in 5 studies. In addition, 27 studies used datasets from other sources, categorized as "Others." Finally, 15 studies did not specify the origin of their datasets, marked as "N/A."

TABLE VI. Total number of studies of dataset sources or repositories used in the primary studies

Dataset Repository	Number of Studies
Others	29
N/A	15
PhysioNet	10
ADNI	9
MIMIC	9
UCI	5

Most datasets in the primary studies do not explicitly categorize the missing mechanism. Instead, the missing mechanism is often assumed. For instance, the study by Shi et al. [43], which uses the MIMIC-III dataset, assumes that electronic health reocrd data follows MAR. Several other studies [7], [20] also assume MAR for their data. Conversely, a study involving microbiome data [44] assumes MNAR and MAR, even though the current experiment itself considers the MCAR, reflecting the impracticality of MCAR for longitudinal data. Similarly, Shen et al. [20] consider only MCAR and MAR, while studies [9], [45] assume their data is MCAR.

D. RQ4.What Performance Metrics are Utilized for Evaluating RNN Imputation Method?

Figure 12 provides a visual representation of the top 5 performance metrics employed in evaluating the effectiveness of the proposed RNN-based imputation method. Among these metrics, Mean Absolute Error (MAE), recognized as a widely adopted criterion for quantifying prediction error [46], emerged as the most prevalent, utilized in over 50% of the articles (n=38). Following MAE, the Root Mean Squared Error (RMSE), utilized to assess the variability of error [46], was employed in 24 articles. Additionally, the Area Under the Receiver Operating Characteristic Curve (AU-ROC) was identified in 14 articles, particularly in tasks involving classification. Other commonly utilized performance metrics include Mean Relative Error (MRE), featured in 12 articles, and Mean Squared Error (MSE), utilized to gauge the accuracy of the proposed method [46], found in 9 articles. It is worth noting that in the evaluation of RNN-based data imputation methods, error metrics such as MAE, RMSE, MRE, and MSE are frequently employed to quantify the reliability and accuracy of the imputed data.

Table VII provides an intricate breakdown of each article and its corresponding performance metrics, encompassing a diverse range of categorizations such as error metrics and classification metrics. Within the error metrics category, examples include MAE, RMSE, MSE, MRE, Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute

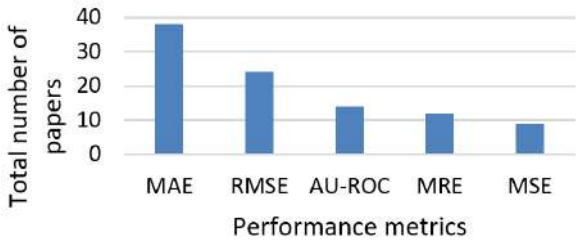

Figure 12. Top 5 Performance Metrics Used in Proposed RNN Method

Percentage Error (SMAPE), R-squared (R2), Absolute Difference Error (ADE), Final Displacement Error (FDE), Median Absolute Error (MEDAE), Normalized Mean Squared Error (NMSE), Normalized Root Mean Squared Deviation (NRMSD), and Root Mean Square (RMS). Conversely, classification metrics encompass metrics like Accuracy, AU-ROC, Area Under the Precision-Recall Curve (AUPRC), F1 score, Recall, and Precision.

Studies conducted by [34], [43], [84] exclusively utilized classification metrics such as F1 score, Recall, Precision, Accuracy, and AU-ROC for their evaluations, focusing on assessing imputation methods beyond just RNN and LSTM for handling missing data. In contrast, Li et al. [15] opted for an error-based evaluation approach to assess the effectiveness of their proposed RNN-based data imputation method. The proposed method aimed at evaluating the recovery of missing data by measuring the similarity difference between imputed values and true values. This distinction highlights the importance of aligning evaluation metrics with the specific objectives and focus areas of the respective research, ensuring a robust and relevant assessment framework tailored to the research goals.

Metrics like ROC-AUC, F-measure, and accuracy are not limited to evaluating data imputation. For instance, Vivar et al. [7] employed these metrics for their classification results while using RMSE to evaluate their imputation results. Additionally, in ablation experiments designed to assess the overall performance of a proposed method, AUC, F-measure, and accuracy were utilized. These examples demonstrate the versatility and importance of selecting appropriate metrics based on the specific evaluation context, whether for imputation accuracy or downstream tasks like classification.

To further analyze how these evaluation metrics are used, the MAE scores from studies that employed the MIMIC-III dataset are summarized. Table VIII presents the performance of different imputation methods at different percentages of missing in the MIMIC-III dataset, evaluated using MAE.

At lower missing rates, the method proposed by Mulyadi et al. [54] achieves the best performance with an MAE of 0.333 ± 0.005 at 5% missingness. Similarly, at 10%

TABLE VII. Performance metrics used in studies

Performance Metric	Total Used in Studies	References
MAE	38	[15], [16], [20], [31], [35], [19], [36], [37], [38], [39], [41], [42], [44], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71]
RMSE	24	[7], [15], [20], [31], [19], [41], [45], [50], [53], [56], [57], [63], [64], [65], [66], [67], [72], [73], [74], [75], [76], [77], [78], [79]
AU-ROC	13	[7], [34], [37], [43], [45], [49], [54], [58], [60], [72], [77], [79], [80]
MRE	12	[16], [35], [37], [38], [42], [54], [57], [59], [60], [62], [69], [71]
MSE	9	[42], [54], [59], [68], [69], [80], [81], [82], [83]
MAPE	6	[20], [19], [65], [66], [73], [78]
Accuracy	5	[7], [9], [34], [72], [84]
AUPRC	5	[54], [58], [60], [77], [79]
F1	3	[7], [34], [84]
Recall and Precision	3	[34], [43], [84]
R2	2	[66], [67]
ADE	1	[85]
FDE	1	[85]
MEDAE	1	[78]
NMSE	1	[66]
NRMSD	1	[86]
RMS	1	[87]
SMAPE	1	[15]

missing data, Zhou et al. [16] outperforms other methods with an MAE of 0.2706. As the missingness increases to 20% and 30%, Ni et al. [41] obtain MAEs of 6.26 ± 0.28 and 6.27 ± 0.17, respectively, reflecting the challenges faced when imputation methods handle greater missingness. Overall, this analysis shows that the percentage of missing data significantly affects the MAE metric, highlighting the challenges posed by increased missingness. Furthermore, MAE provides a clear indication of the average error magnitude in the imputed values, expressed in the same units as the data, which makes it easier to understand [88].

E. RQ5. What are the Advantages and Potential Drawbacks of the Proposed RNN Imputation Method?

One notable strength of the RNN-based data imputation method is its capability to generate highly accurate results, as highlighted in several scholarly articles [16], [52], [66], [72], [75], [79], [80], [81], [89]. Previous studies have also indicated that the variance of RNN-based data imputation methods, such as LSTM and GRU, consistently outperforms the baseline established by conventional RNN and machine learning imputation methods, as evidenced in works by researchers [9], [15], [34], [36], [67], [69], [84]. For instance, Li et al. [15] reported in the study that their attention-based RNN method surpassed other machine learning models, including ARIMA and KNN, achieving notably lower average MAE values of 0.083, 0.171, 0.055, and 0.105 across various real-world time-series datasets. This superior performance not only highlights the efficacy of RNN-based methods but also underscores their adaptability in handling diverse data types and imputation tasks, consistently delivering reliable imputation results across a spectrum of datasets.

Prior studies have delved into the realm of optimization algorithms to bolster the performance of RNNs for missing data imputation. For instance, Liang et al. [61] introduced a dynamic task weighting scheme grounded in dynamic gradient magnitude adjustments. This innovative approach aims to autonomously achieve balanced training across tasks, thereby addressing the challenges posed by the multi-task learning paradigm. Similarly, Li et al. [74] advocated for a smoothing regularization term to optimize the selection of hyperparameters in the RNN model. The integration of such optimization algorithms has led to a significant enhancement in the performance of RNN-based imputation methods, showcasing tangible improvements in their efficacy and accuracy.

A remarkable limitation observed in many proposed RNN-based imputation methods is their substantial computational cost, as evidenced by articles such as [47], [50], [54], [66], [70], [89]. Furthermore, several RNN-based imputation methods encounter limitations due to their inherent model complexity, as indicated in articles [16], [20], [34], [44], [47], [61], [79], [80]. For instance, Zhou et al. [16] bi-directional recurrent structure, characterized by a multitude of parameters, significantly increases the model's complexity, resulting in extended training times and potentially restricting its applicability to large-scale datasets. Additionally, it is noteworthy that a majority of the articles included in this study primarily focus their evaluation on a single problem domain and dataset. Only a limited number of studies, [15], [35], [44], [45], [50], [59], [69], [76], [80], have conducted robust evaluations of their proposed RNN-based imputation methods using multiple datasets from different problem domains. This delineation underscores the need for broader and more diverse evaluations to comprehensively assess the generalizability and robustness of RNN-based imputation methods across various contexts and datasets.

TABLE VIII. MAE result of imputation methods at different missing percentage

Study	5% Missing	10% Missing	20% Missing	30% Missing
[16]	N/A	0.2706	N/A	N/A
[54]	0.333±0.005	0.311±0.006	N/A	N/A
[41]	N/A	5.42±0.2	6.26±0.28	6.27±0.17
[58]	0.497±0.012	0.503±0.011	N/A	N/A
[62]	N/A	7.97±5.89	N/A	N/A

5. Findings

In this section, a comprehensive examination of the findings and analyses presented in the preceding section is undertaken. Our aim is to provide a detailed discussion on the current landscape surrounding the utilization of RNNs for the purpose of missing data imputation.

The surge in the adoption of RNNs since 2018 signifies a significant shift in the data analysis landscape, particularly in addressing the challenges posed by missing data. This trend is not merely a fleeting phenomenon but rather a reflection of the inherent capabilities and adaptability of RNNs in handling complex data scenarios. One of the key strengths of RNNs lies in their ability to capture temporal dependencies and sequential patterns, making them particularly well-suited for imputing missing data within time-series datasets. This capability has garnered considerable attention from both academic researchers and industry practitioners, leading to a notable uptick in RNN utilization.

The concurrent advancements in deep learning APIs, such as TensorFlow and Keras, have played a pivotal role in fueling the surge in RNN adoption. These robust frameworks have democratized the implementation of RNN-based solutions, making them more accessible and feasible for a broader range of applications. Moreover, the symbiotic relationship between the surge in RNN adoption and the availability of deep learning APIs is evident in their collaborative impact on research and practical applications. These APIs have not only streamlined the integration of RNNs into existing data analysis pipelines but have also catalyzed research by enabling researchers to explore novel methodologies and architectures.

The versatility of RNNs extends beyond mere data imputation; they have proven instrumental in a myriad of tasks, including natural language processing, time-series forecasting, and pattern recognition. This multifaceted utility positions RNNs as a cornerstone technology in the arsenal of deep learning tools, heralding a new era in data-driven decision-making and predictive modelling across diverse domains. However, it's crucial to note that the dominance of specific RNN variants, such as LSTM, in the realm of missing data imputation is also a result of their inherent advantages, notably their adaptability to data with varying lengths and their proficiency in capturing long-range dependencies within sequences. On the other hand, while the GRU presents simpler architecture and faster training times on smaller datasets, its limitations become apparent when dealing with larger and more complex datasets, highlighting the nuanced trade-offs inherent in selecting an appropriate RNN variant for specific missing data scenarios.

Determining the underlying missing mechanism is challenging, as different variables within the same dataset may follow different mechanisms. For example, one variable might exhibit MAR, while another could follow MNAR or a combination of mechanisms. This complexity is compounded by the fact that the missing mechanism is independent of the variable's role in the analysis. Whether the variable is an outcome or a covariate, the mechanism depends solely on the nature of the missingness itself [6]. Therefore, successful imputation requires not only a robust model like RNN but also an understanding of the missingness patterns across variables. While RNN has shown great promise, accurately identifying the missingness mechanism for all variables in large and complex datasets remains an ongoing challenge.

In evaluating the performance of RNN-based imputation methods, the selection and utilization of error metrics like RMSE, MAE, MSE, and MRE, along with classification metrics such as AU-ROC, are contingent upon the specific objectives and focus of the research at hand. These metrics serve as crucial tools for quantifying the accuracy, reliability, and effectiveness of RNN-based imputation methods, providing researchers with valuable insights into the strengths, limitations, and overall performance of these methods tailored to the research objectives. By strategically selecting and integrating these metrics into the evaluation framework, researchers can gain comprehensive insights into the performance and utility of RNN-based imputation methods across various missing data scenarios.

A. Future Research Directions

Drawing upon the insights gleaned from this study, several promising avenues for future research in the domain of RNN-based data imputation emerge.

1) Developing and Testing New RNN-Based Imputation Method: Future research could focus on designing RNN-based architectures specifically with the aim to enhance missing data imputation across various types of datasets. It focus on creating and testing flexible architectures or hybrid RNN techniques that adapt to diverse domains and missing data mechanisms. By developing another RNN-

based methods, the research can contribute to improving data completeness, ultimately making RNN-based imputation methods more robust in real-world applications.

2) Investigating the Effectiveness of Different RNN Architecture: The effectiveness of different RNN architectures for imputing missing values across diverse types of datasets is a crucial area for future exploration. While RNNs contribute to enhancing data quality by addressing the data completeness dimension, understanding the strengths and limitations of each architecture becomes paramount when dealing with varied datasets characterized by distinct domains and missing mechanisms. This necessitates the development of more flexible RNN models capable of handling missing values more effectively.

3) Adopting Optimization Algorithm in RNN-Based Data Imputation Method: The utilization of optimization algorithms such as Particle Swarm Optimization (PSO) and stochastic gradient descent (SGD) within the framework of RNNs presents a promising avenue for addressing the challenges associated with missing data imputation. While previous research has highlighted the potential of optimization, further exploration is required to fully exploit its capabilities. Investigating the impact of these optimization algorithm on model convergence, generalization, and computational efficiency, future research contributes to the development of more robust and effective RNN-based data imputation methods for handling missing data in various domains. Additionally, simplifying RNN architectures, such as reducing the number of trainable parameters and performing hyperparameter optimization with the help of these optimization algorithms, could help improve training efficiency and manage model complexity. To reduce computational costs and allow for broader use in resource-constrained settings, distributed methods of training and the use of hardware accelerations, like graphics processing unit or tensor processing unit, could also be explored.

4) Increased Diversity of Datasets and Domains: This study's scope is limited to RNN-based imputation methods, with experiments conducted on a few specific domains. Future research should aim to investigate a wider range of datasets from diverse domains to evaluate the generalizability and robustness of the method. Expanding this approach across various types of time-series datasets could enhance the versatility and applicability of RNN-based imputation for broader, real-world use.

6. Conclusion

The primary objective in this study was to critically analyse the recent landscape of RNN-based imputation methods for missing data in time-series datasets. The central focus of this evaluation is to illuminate the existing gaps by highlighting the strengths demonstrated by recent RNN-based imputation methods, while also addressing their limitations in handling missing data for time-series datasets. The analysis started with 362 articles from SCOPUS database and reduced to 70 articles after implementing the inclusion and exclusion criteria. In the review, RNN or its variant including LSTM and GRU is used to impute missing data in the dataset.

While the review provides valuable insights, it is essential to acknowledge a limitation. The study focuses primarily on time-series datasets, and the findings may not be directly applicable to datasets in other domains, such as cross-sectional or spatial data. Additionally, the absence of explicit analysis on datasets and the chosen evaluation metrics in relation to their specific domains for data imputation using the RNN method is a notable gap. Although dataset features have been addressed, further exploration is needed to examine how these features, along with the selected evaluation metrics within the context of RNN-based imputation techniques.

The prominence of LSTM in the reviewed studies underscores its popularity and effectiveness in handling missing values. However, the comparison between LSTM and GRU variants may not fully account for the impact of different hyperparameters or variations in dataset size and complexity, which can influence performance outcomes. Nevertheless, the review highlights a need for further advancements in RNN methodologies to enhance their capabilities in addressing the complexities associated with missing data.

In conclusion, the research on data imputation methods, particularly RNNs, is vital for mitigating challenges posed by missing values in datasets. This review serves as a foundation for future studies to refine and extend RNN-based approaches, ultimately contributing to the continuous improvement of data imputation techniques.

Acknowledgment

The authors gratefully acknowledge the financial support from the International Matching Grant with Project ID UIC241510 from the Universiti Malaysia Pahang Al-Sultan Abdullah (RDU242708). This support is gratefully acknowledged.

References

[1] F. G. Alizamini, M. M. Pedram, M. Alishahi, and K. Badie, "Data quality improvement using fuzzy association rules," in *2010 International Conference on Electronics and Information Engineering*. IEEE, 2010, pp. ssV1–468–V1–472.

[2] S. Issa, O. Adekunle, F. Hamdi, S. S.-S. Cherfi, M. Dumontier, and A. Zaveri, "Knowledge graph completeness: A systematic literature review," *IEEE Access*, vol. 9, pp. 31 322–31 339, 2021.

[3] M. Mohammed, J. R. Talburt, S. Dagtas, and M. Hollingsworth, "A zero trust model based framework for data quality assessment," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2021, pp. 305–307.

[4] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, p. 140, 2021.

[5] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: A review and comparative study," *Expert Systems with Applications*, vol. 227, p. 120201, 2023.

[6] T. F. Johnson, N. J. B. Isaac, A. Paviolo, and M. González-Suárez, "Handling missing values in trait data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51–62, 2021.

[7] G. Vivar, A. Kazi, H. Burwinkel, A. Zwergal, N. Navab, and S. A. Ahmadi, "Simultaneous imputation and classification using multigraph geometric matrix completion (mgmc): Application to neurodegenerative disease classification," *Artificial Intelligence in Medicine*, vol. 117, p. 102097, 2021.

[8] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, p. 693674, 2021.

[9] Q. Ma, S. Li, and G. W. Cottrell, "Adversarial joint-learning recurrent neural network for incomplete time series classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1765–1776, 2022.

[10] R. K. Pansari, M. Gupta, and R. Wadhvani, "Multivariate imputation by n neighbour mean and chained equation for time series missing data," in *Proceedings of 2023 IEEE 2nd International Conference on Industrial Electronics: Developments and Applications, ICIDeA 2023*, 2023, pp. 13–18.

[11] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decision Analytics Journal*, vol. 9, p. 100341, 2023.

[12] C. Liu, D. Zowghi, and A. Talaei-Khoei, "An empirical study of the antecedents of data completeness in electronic medical records," *International Journal of Information Management*, vol. 50, pp. 155–170, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026840121831065X

[13] P. B. Weerakody, K. W. Wong, and G. Wang, "Cyclic gate recurrent neural networks for time series data with missing values," *Neural Processing Letters*, vol. 55, no. 2, pp. 1527–1554, 2023.

[14] C. Li, X. Ren, and G. Zhao, "Machine-learning-based imputation method for filling missing values in ground meteorological observation data," *Algorithms*, vol. 16, p. 422, 2023.

[15] Y. Li, M. Du, and S. He, "Attention-based sequence-to-sequence model for time series imputation," *Entropy*, vol. 24, no. 12, 2022.

[16] X. Zhou, W. Xiang, and T. Huang, "A novel neural network for improved in-hospital mortality prediction with irregular and incomplete multivariate data," *Neural Networks*, vol. 167, pp. 741–750, 2023.

[17] Y. Liu, B. Li, S. Yang, and Z. Li, "Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods," *Expert Systems with Applications*, vol. 237, p. 121694, 2024.

[18] M. e. a. Alabadla, "Systematic review of using machine learning in imputing missing values," *IEEE Access*, vol. 10, pp. 44 483–44 502, 2022.

[19] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional lstm recurrent neural network for forecasting network-wide traffic state with missing values," *Transportation Research Part C: Emerging Technologies*, vol. 118, p. 102674, 2020.

[20] G. Shen, W. Zhou, W. Zhang, N. Liu, Z. Liu, and X. Kong, "Bidirectional spatial–temporal traffic data imputation via graph attention recurrent neural network," *Neurocomputing*, vol. 531, pp. 151–162, 2023.

[21] T. P. Mesquita, D. M. P. F. Silva, A. M. N. C. Ribeiro, I. R. R. Silva, C. J. A. Bastos-Filho, and R. P. Monteiro, "A comparative analysis of deep learning-based methods for multivariate time series imputation with varying missing rates," *IEEE Eighth Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, 2024.

[22] M. Kazijevs and M. D. Samad, "Deep imputation of missing values in time series health data: A review with benchmarking," *J Biomed Inform*, vol. 144, 2023.

[23] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – a systematic literature review," *Inf Softw Technol*, vol. 51, no. 1, pp. 7–15, 2009.

[24] V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, "The journal coverage of web of science, scopus and dimensions: A comparative analysis," *Scientometrics*, vol. 126, no. 6, pp. 5113–5142, 2021.

[25] L. Chen, M. A. Babar, and H. Zhang, "Towards an evidence-based understanding of electronic data sources," in *Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering, EASE'10*. Swindon, GBR: BCS Learning & Development Ltd., 2010, pp. 135–138.

[26] L. Yang *et al.*, "Quality assessment in systematic literature reviews: A software engineering perspective," *Inf Softw Technol*, vol. 130, p. 106397, 2021.

[27] M. Ma *et al.*, "Democratizing production-scale distributed deep learning," 2018, accessed: Jan. 04, 2024. [Online]. Available: http://arxiv.org/abs/1811.00143

[28] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," 2024, accessed: Jan. 04, 2024. [Online]. Available: https://www.researchgate.net/publication/317418674_Accurate_Large_Minibatch_SGD_Training_ImageNet_in_1_Hour

[29] S. Saydam and V. Kecojevic, "Publication strategies for academic career development in mining engineering," *Mining Technology*, vol. 123, no. 1, pp. 46–55, 2014.

[30] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Comput Sci*, vol. 2, no. 6, pp. 1–20, 2021.

[31] M. Saad, L. Nassar, F. Karray, and V. Gaudet, "Tackling imputation across time series models using deep learning and ensemble learning," in *Conf Proc IEEE Int Conf Syst Man Cybern*, vol. 2020-October, 2020, pp. 3084–3090.

[32] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci Rep*, vol. 8, no. 1, pp. 1–12, 2018.

[33] S. Yang, X. Yu, and Y. Zhou, "Lstm and gru neural network

performance comparison study: Taking yelp review dataset as an example," *Proceedings - 2020 International Workshop on Electronic Communication and Artificial Intelligence, IWECAI 2020*, pp. 98–101, 2020.

[34] Y. J. Kim and M. Chi, "Temporal belief memory: Imputing missing data during rnn training," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2018-July, pp. 2326–2332, 2018.

[35] W. Cao, H. Zhou, D. Wang, Y. Li, J. Li, and L. Li, "Brits: Bidirectional recurrent imputation for time series," in *Advances in Neural Information Processing Systems*. NeurIPS, 2018, pp. 6775–6785.

[36] Z. Wu *et al.*, "Brnn-gan: Generative adversarial networks with bi-directional recurrent neural networks for multivariate time series imputation," in *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*. IEEE, 2021, pp. 217–224.

[37] Q. Wang, S. Ren, and Y. Xia, "Mortality prediction with bidirectional coupled and gumbel subset network on irregularly multivariate time series," in *International Conference on Signal Processing Proceedings, ICSP*, vol. 2022-October, 2022, pp. 468–473.

[38] H. Gao, R. Li, J. Wang, H. Zhao, S. Yan, and L. Ma, "Research on bidirectional recurrent imputation of multivariate time series for clinical outcomes prediction," in *Proceedings - 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022*, 2022, pp. 954–960.

[39] M. Gupta, T. L. T. Phan, H. T. Bunnell, and R. Beheshti, "Concurrent imputation and prediction on ehr data using bi-directional gans: Bi-gans for ehr imputation and prediction," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2021*, 2021.

[40] D. Wen and X. Chen, "Research on multivariate time series prediction method based on missing data imputation," *Proceedings of the International Society for Optics and Photonics*, vol. 12791, pp. 541–548, 2023.

[41] Q. Ni and X. Cao, "Mbgan: An improved generative adversarial network with multi-head self-attention and bidirectional rnn for time series imputation," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105232, 2022.

[42] J. Xu, W. Lu, Y. Li, C. Zhu, and Y. Li, "A multi-directional recurrent graph convolutional network model for reconstructing traffic spatiotemporal diagram," *Transportation Letters*, pp. 1–11, 2023.

[43] D. Shi and H. Zheng, "A mortality risk assessment approach on icu patients clinical medication events using deep learning," *Computer Modeling in Engineering & Sciences*, vol. 128, no. 1, pp. 161–181, 2021.

[44] J. M. Choi, M. Ji, L. T. Watson, and L. Zhang, "Deepmicrogen: a generative adversarial network-based method for longitudinal microbiome data imputation," *Bioinformatics*, vol. 39, no. 5, p. btad286, 2023.

[45] P. B. Weerakody, K. W. Wong, and G. Wang, "Policy gradient empowered lstm with dynamic skips for irregular time series data," *Applied Soft Computing*, vol. 142, p. 110314, 2023.

[46] G. Doreswamy, I. Gad, and B. R. Manjunatha, "Performance evaluation of predictive models for missing data imputation in weather data," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1327–1334.

[47] M. Nguyen, N. Sun, D. C. Alexander, J. Feng, and B. T. T. Yeo, "Modeling alzheimer's disease progression using deep recurrent neural networks," in *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2018.

[48] W. Jung, A. W. Mulyadi, and H. I. Suk, "Unified modeling of imputation, forecasting, and prediction for ad progression," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11767 LNCS, 2019, pp. 168–176.

[49] M. Mehdipour Ghazi *et al.*, "Training recurrent neural networks robust to incomplete data: Application to alzheimer's disease progression modeling," *Medical Image Analysis*, vol. 53, pp. 39–46, 2019.

[50] L. Nassar, M. Saad, I. E. Okwuchi, M. Chaudhary, F. Karray, and K. Ponnambalam, "Imputation impact on strawberry yield and farm price prediction using deep learning," in *Conference Proceedings IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, vol. 2020-October, 2020, pp. 3599–3605.

[51] D. P. Dao, N. H. Ho, J. Kim, and H. J. Yang, "Improving recurrent gate mechanism for time-to-conversion prediction of alzheimer's disease," in *ACM International Conference Proceeding Series*. ACM, 2020, pp. 66–71.

[52] Y. Zhao, M. Berretta, T. Wang, and T. Chitnis, "Gru-df: A temporal model with dynamic imputation for missing target values in longitudinal patient data," in *2020 IEEE International Conference on Healthcare Informatics, ICHI 2020*, Nov 2020.

[53] Q. Ma *et al.*, "End-to-end incomplete time-series modeling from linear memory of latent variables," *IEEE Trans Cybern*, vol. 50, no. 12, pp. 4908–4920, 2020.

[54] A. W. Mulyadi, E. Jun, and H. I. Suk, "Uncertainty-aware variational-recurrent imputation network for clinical time series," *IEEE Trans Cybern*, 2021.

[55] V. A. Le, T. T. Le, P. L. Nguyen, H. T. T. Binh, R. Akerkar, and Y. Ji, "Gcrint: Network traffic imputation using graph convolutional recurrent neural network," in *IEEE International Conference on Communications*, Jun 2021.

[56] W. Zhong, Q. Suo, X. Jia, A. Zhang, and L. Su, "Heterogeneous spatio-temporal graph convolution network for traffic forecasting with missing values," in *Proc Int Conf Distrib Comput Syst*, vol. 2021-July, 2021, pp. 707–717.

[57] W. Jung, E. Jun, H. I. Suk, and A. D. N. Initiative, "Deep recurrent model for individualized prediction of alzheimer's disease progression," *Neuroimage*, vol. 237, p. 118143, 2021.

[58] E. Jun, A. W. Mulyadi, J. Choi, and H. I. Suk, "Uncertainty-gated stochastic sequential model for ehr mortality prediction," *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 9, pp. 4052–4062, 2021.

[59] Y. Zhou, J. Jiang, S. H. Yang, L. He, and Y. Ding, "Musdri: Multi-seasonal decomposition based recurrent imputation for time series," *IEEE Sens J*, vol. 21, no. 20, pp. 23 213–23 223, 2021.

[60] D. Li, P. Lyons, J. Klaus, B. Gage, M. Kollef, and C. Lu, "Integrating

static and time-series data in deep recurrent models for oncology early warning systems," in *International Conference on Information and Knowledge Management, Proceedings*, 2021, pp. 913–922.

[61] W. Liang, K. Zhang, P. Cao, X. Liu, J. Yang, and O. Zaiane, "Rethinking modeling alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network," *Comput Biol Med*, vol. 138, p. 104935, 2021.

[62] Z. Shi *et al.*, "Deep dynamic imputation of clinical time series for mortality prediction," *Inf Sci (N Y)*, vol. 579, pp. 607–622, 2021.

[63] M. Adib, U. Zaman, D. A. Du, U. Zaman, and D. A. Du, "A stochastic multivariate irregularly sampled time series imputation method for electronic health records," *BioMedInformatics 2021*, vol. 1, no. 3, pp. 166–181, 2021.

[64] Z. Ji, W. Zhu, and W. Zhongchang, Ji, "A traffic data imputing method based on multisource recurrent neural network," *Proceedings of SPIE*, vol. 12260, pp. 90–95, 2022.

[65] X. Deng, H. Zhou, and C. Ye, "Multi-feature fusion strategy for missing values filling in traffic prediction," in *Proceedings - 2022 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Autonomous and Trusted Vehicles, Scalable Computing and Communications, Digital Twin, Privacy Computing, Metaverse*, 2022, pp. 227–234.

[66] K. Chreng, H. S. Lee, R. P. Pradana, T. Q. Trong, I. D. G. Arya Putra, and H. Nimiya, "Imputation of missing values for generating typical meteorological year (tmy) with data decomposition and recurrent neural networks," *IOP Conf Ser Earth Environ Sci*, vol. 1007, no. 1, p. 012020, 2022.

[67] S. Shan *et al.*, "A deep-learning based solar irradiance forecast using missing data," *IET Renewable Power Generation*, vol. 16, no. 7, pp. 1462–1473, 2022.

[68] Y. Huang, Y. Tang, J. VanZwieten, and J. Liu, "Reliable machine prognostic health management in the presence of missing data," *Concurr Comput*, vol. 34, no. 12, p. e5762, 2022.

[69] Y. Chen, Z. Li, C. Yang, X. Wang, G. Long, and G. Xu, "Adaptive graph recurrent network for multivariate time series imputation," in *Communications in Computer and Information Science*, vol. 1792, 2023, pp. 64–73.

[70] N. H. Ho, H. J. Yang, J. Kim, D. P. Dao, H. R. Park, and S. Pant, "Predicting progression of alzheimer's disease using forward-to-backward bi-directional network with integrative imputation," *Neural Networks*, vol. 150, pp. 422–439, 2022.

[71] Z. Wu, C. Ma, X. Shi, L. Wu, Y. Dong, and M. Stojmenovic, "Imputing missing indoor air quality data with inverse mapping generative adversarial network," *Build Environ*, vol. 215, p. 108896, 2022.

[72] G. Vivar, A. Zwergal, N. Navab, and S. A. Ahmadi, "Multi-modal disease classification in incomplete datasets using geometric matrix completion," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11044, pp. 24–31, 2018.

[73] Z. Khan, S. M. Khan, K. Dey, and M. Chowdhury, "Development and evaluation of recurrent neural network-based models for hourly traffic volume and annual average daily traffic prediction," *Transportation Research Record*, vol. 2673, no. 7, pp. 489–503, 2019.

[74] D. Li, L. Li, X. Li, Z. Ke, and Q. Hu, "Smoothed lstm-ae: A spatio-temporal deep model for multiple time-series missing imputation," *Neurocomputing*, vol. 411, pp. 351–363, 2020.

[75] M. Zymbler, Y. Kraeva, E. Latypova, S. Kumar, D. Shnayder, and A. Basalaev, "Cleaning sensor data in smart heating control system," in *Proceedings - 2020 Global Smart Industry Conference, GloSIC 2020*, 2020, pp. 375–381.

[76] E. Oh, T. Kim, Y. Ji, and S. Khyalia, "Sting: Self-attention based time-series imputation networks using gan," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2021, pp. 1264–1269.

[77] B. Li, Y. Shi, L. Cheng, Z. Yan, X. Wang, and H. Li, "Mtssp: Missing value imputation in multivariate time series for survival prediction," in *Proceedings of the International Joint Conference on Neural Networks*, 2022.

[78] M. W. Saif-ul Allah, M. A. Qyyum, N. Ul-Haq, C. A. Salman, and F. Ahmed, "Gated recurrent unit coupled with projection to model plane imputation for the pm2.5 prediction for guangzhou city, china," *Frontiers in Environmental Science*, vol. 9, p. 816616, 2022.

[79] B. Li *et al.*, "Mvira: A model based on missing value imputation and reliability assessment for mortality risk prediction," *International Journal of Medical Informatics*, vol. 178, p. 105191, 2023.

[80] J. Zhao, C. Rong, C. Lin, and X. Dang, "Multivariate time series data imputation using attention-based mechanism," *Neurocomputing*, vol. 542, p. 126238, 2023.

[81] A. Rawat and A. Solanki, "Sequence imputation using machine learning with early stopping mechanism," in *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, 2020, pp. 859–863.

[82] O. S. Ezeora, J. Heckenbergerova, P. Musilek, and J. Rodway, "Sampling control in environmental monitoring systems using recurrent neural networks," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–4.

[83] T. Deng, M. Wan, K. Shi, L. Zhu, X. Wang, and X. Jiang, "Short term prediction of wireless traffic based on tensor decomposition and recurrent neural network," *SN Appl Sci*, vol. 3, no. 9, p. 779, 2021.

[84] S. Park, Y. Seonwoo, J. Kim, J. Kim, and A. Oh, "Denoising recurrent neural networks for classifying crash-related events," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 2906–2917, 2020.

[85] R. Fujii, J. Vongkulbhisal, R. Hachiuma, and H. Saito, "A two-block rnn-based trajectory prediction from incomplete trajectory," *IEEE Access*, vol. 9, pp. 56 140–56 151, 2021.

[86] Q. Suo, L. Yao, G. Xun, J. Sun, and A. Zhang, "Recurrent imputation for multivariate time series with missing values," in *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, 2019.

[87] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Appl Energy*, vol. 212, pp. 372–385, 2018.

[88] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (mae) into systematic and unsystematic components," *PLoS One*, vol. 18, no. 2, p. e0279774, 2023.

[89] Y.-F. Zhang, P. J. Thorburn, M. P. Vilas, and P. Fitch, "Machine learning approaches to improve and predict water quality data," 2019.