Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal

Cite this: DOI: 10.1039/d4va00367e

AI for enhanced water quality data imputation: a deep learning perspective

Ishan Prasad Banjara, [a] Suman Poudel, [a] Kalam Pariyar, [a] Deepesh Upreti, [a] Antigoni Zafeirakou [b] and Shukra Raj Paudel *[a]

Water quality data, a crucial resource for scientific water resource management practices (*e.g.*, irrigation), engineering solutions (*e.g.*, process control of both water and wastewater treatment plants), *etc.*, are often hindered in their utility due to missingness within the dataset. Addressing this challenge, this perspective article underscores the necessity of missing data imputation. Along with highlighting the imputation strengths and limitations of different statistical and machine learning models, this article highlights deep learning (DL) models, and their underlying major limitations as well as potential resolutions. This study embodies novelty by proposing a robust model, integrating diverse solutions with an aim to set new standards in terms of accuracy, efficiency and adaptability in the domain of water quality data analysis. The paper presents the real-world implementation of the proposed framework along with its limitations and potential resolutions. Finally, the study concludes by calling forth coordinated efforts from researchers of diverse disciplines for developing a novel, generalized, and memory-efficient deep learning architecture.

Received 21st October 2024
Accepted 5th April 2025

DOI: 10.1039/d4va00367e

rsc.li/esadvances

Environmental significance

The manuscript addresses critical issues in water resource management through advanced deep learning techniques for robust water quality data imputation. Accurate and complete water quality data are pivotal for informed decision-making in environmental policy, ecosystem protection, and public health management. Existing deep learning algorithms, while effective, often face limitations related to computational demands, adaptability across diverse datasets, and handling spatial–temporal complexities. This study critically reviews these limitations, highlighting their consequences, and proposes an innovative framework integrating spatial–temporal analysis, dynamic ensemble modelling, and an I/O-aware mechanism to enhance accuracy and efficiency. Ultimately, it contributes to achieving the sixth goal of UN's Sustainable Development Goals by promoting access to safe, clean, and reliable water.

Background

Water quality data are vital to the effective management of water resources as well as to make informed decisions for environmental policy, ecosystem protection, and public health management. For instance, Dissolved Oxygen (DO) is a critical parameter for assessing the health of the aquatic ecosystem and its biodiversity. A low level of DO in rivers could result into the death of aquatic fauna as well as causing ecological disruption. It is important to utilize the historical water quality data to assess the level of DO in water and develop a suitable model for early warning of potential water quality risks.[1] Similarly, in the context of reclaimed water reuse in agriculture, water quality data are fundamental for developing policy level documents such as agricultural water quality guidelines. Using the water quality data in the field, regulatory bodies can enforce compliance within the wastewater reuse projects and ensure public health safety while consumption of crops irrigated *via* reclaimed water.[2] The sixth goal of the United Nations Sustainability framework addresses the importance of sustainable management of water and sanitation for all aiming towards pollution reduction as well as enhancement of water quality. Access to good quality water and sanitation is also the foundation for achieving other sustainable development goals.[3] Realizing such fundamental importance of safe water, many nations have established their separate organizations devoted to collecting and maintaining water quality data which serves to facilitate the monitoring and scientific management of water resources. However, in underdeveloped or developing nations, although such separate bodies exist, the act of maintaining water quality data often remains only on paper rather than systematic implementations. This issue arises due to the finance and labor intensity required for such monitoring programs as many developing nations lack adequate capital, technology, and skilled human resources for the water quality data collection system and its management. For instance, despite the rampant prevalence of emerging contaminants, pharmaceutical compounds, and micro-

[a]*Environmental Engineering Program, Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Pulchowk, Lalitpur, Nepal. E-mail: 076bce080.ishan@pcampus.edu.np; 076bce176.suman@pcampus.edu.np; 076bce084.kalam@pcampus.edu.np; 076bce067.deepesh@pcampus.edu.np; srpaudel@ioe.edu.np; Tel: +977-9851220184*

[b]*Department of Civil Engineering, School of Engineering, Aristotle University of Thessaloniki, Greece. E-mail: azafir@civil.auth.gr*

pollutants, there has not been any investigation in Nigerian water sources for the past two decades. The use of technologies like GIS for water quality monitoring and its examination is also quite rare in the region.[4] Also, even with proper investment in advanced water quality data recording equipment, the problem of missing data is prevalent, particularly in real-time and automated processes,[5] which is a challenge faced even by developed nations. For instance, a water quality dataset obtained from the Great Barrier Reef monitoring program has about 60% of missing data for its nitrate parameter even though it is the most influential parameter for the stability of the Great Barrier Reef's health.[6]

Even if there are water quality parameters with such high rate of missingness, researchers often do not explicitly address or pinpoint the rate of missing data in their studies. As a result, conclusions drawn from such studies can be biased/mislead and anomalies in water quality data could go undetected, thereby questioning the reliability of those studies. Furthermore, decisions based on such studies may lead to erroneous strategies that might exacerbate the problems rather than solve them.[7] Therefore, it is extremely important for a call to develop an enhanced water quality data imputation model to impute the missing data with higher accuracy while also addressing the computational restraints which is a common issue in underdeveloped or developing countries.

Traditional statistical and classical machine learning models

Traditional statistical models such as mean imputation, regression imputation, and Expectation-Maximization (EM)

Ishan Prasad Banjara

Ishan Prasad Banjara is an innovative civil engineer with a graduate degree from the Institute of Engineering, Pulchowk Campus, Tribhuvan University, Nepal. He worked as a research assistant (RA) under Prof. Paudel for about 2 years. His expertise bridges traditional civil engineering with advanced technologies, focusing on machine learning, computer vision, and data-driven hydrology. Ishan has led projects such as an AI-based surveillance system for real-time identification and contributed to earthquake-resistant structural designs. A multiple national awardee and scholarship holder, Ishan is dedicated to integrating cutting-edge solution into civil engineering for sustainable development.

Suman Poudel

Suman Poudel is a civil engineering graduate from Pulchowk Campus, IOE, TU, Nepal. He has worked as a research assistant (RA) under Prof. Paudel for about 3 years. His expertise spans climate change, water quality, and water and wastewater treatment technologies, and is actively engaged in research at the intersection of environmental and water resources, Beyond traditional civil engineering, Mr Suman has developed deep learning models for interdisciplinary applications, including vehicle detection, flood image segmentation, rainfall prediction, and scientific machine learning. His work reflects a strong commitment to integrating data-driven approaches with engineering challenges.

Kalam Pariyar

Kalam Pariyar is a civil engineering graduate from Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal. He has worked as a research assistant (RA) under Prof. Paudel for about 2 years. He has contributed to research work focusing on developing Machine Learning and Deep Learning models to address challenges in the water quality domain. His research interest areas are data-driven decision making processes using deep learning, climate change and water quality, sustainable infrastructure, and their intersection area, etc.

Deepesh Upreti

Deepesh Upreti is a recent graduate in Civil Engineering from Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal. He worked as a RA under Prof. Paudel for about 2 years, and also worked on a project to survey, design and model the Dhobikhola Bridge of Thakre VDC, Dhadhing, Nepal, which was funded by the Government of Nepal, Local Roads and Bridge Support Units. He has been actively involved in research activities and projects in numerous interdisciplinary fields ranging from water resource management, environmental engineering, structural engineering, machine learning, and deep learning algorithms.

have been widely adapted for most of the imputation tasks due to their simplicity, less computational demand, and ease of implementation.[8] However, in many cases, these methods often fall short of addressing the complexities in the datasets and their inherent pattern of missing data.[9] Their limitations become more evident as the nature of the datasets becomes larger, often leading to inaccurate or biased results. To address the limitations of traditional statistical methods, classical machine learning models like Support Vector Regression (SVR), Decision Trees, Random Forest, XGBRgresssor, and MICE have been employed, owing to their greater efficacy in various imputation studies. However, even machine learning models aren't devoid of limitations. The complexity of SVR is dependent on the kernel function which is not always linear. Additionally, there is no proven method to select appropriate or optimize the parameters involved.[10] Decision trees are often prone to overfitting and errors for imbalanced data.[11] The XGBRegressor model demands substantial computational resources, and is quite unstable due to its sensitivity to small changes in the training sets. Random Forest is also computationally complex on large datasets as well as fails to capture the temporal nature of the dataset.[11] Similarly, MICE, even when integrated with other machine learning models, fails to capture the temporal dimension throughout the imputation process.[12] Moreover, as the number of datasets becomes enormous, the performance of these models often falls short compared to deep learning models.[13]

Antigoni Zafeirakou

Antigoni Zafeirakou is an Associate Professor at the Department of Civil Engineering, Aristotle University of Thessaloniki, Greece. She focuses on environmental engineering, sustainable water, wastewater and storm water management, and circular economy. She is a member of organizing and scientific committees of international conferences, is a journal and conference paper reviewer, is a guest editor in Special Issues and is a reviewer of national and international research projects. She is a member of the Technical Commerce of Greece and the Special Committee for the Environment, Solid Waste Management Association in Greece, Committee for Sustainable Development of A.U.Th., Hellenic Hydrotechnical Association, UNESCO-INWEB, and Water Footprint Network.

Shukra Raj Paudel

Shukra Raj Paudel is an Associate Professor of Environmental Engineering at the Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal. He has been working in the areas of energy, resource recovery, climate change, water quality, water and wastewater treatment technologies, & data-driven process control of the treatment systems. He is actively involved in teaching and research in environmental engineering at the university. He is a member of the editorial advisory board of Renewable Energy Focus, International Journal of Ambient Energy, *and* Discover Applied Sciences, *and also serves as an associate editor of* H2Open Journal, *and some domestic journals. He also worked as a guest editor of* ACS ES&T Water.

Deep learning models

While decision trees and their ensembles are not inherently designed to capture temporal patterns, neural networks—especially with deep architecture—encapsulate the correlation between the variables involved as well as the non-linear temporal patterns in the dataset. For instance, a neural network developed by Zhang *et al.*, 2019 primarily focuses on retrieving two most important aspects of water quality data: (1) temporal information between data gaps *via* a long short-term memory neural network and (2) the global attention mechanism to focus on distinct parts of input. This is the reason for the widespread adoption of deep learning architecture in the majority of state-of-the-art imputation models.

However, there are limitations even within such robust models which call for innovative approaches to enhance their performance. The first limitation of the deep learning models is that existing models for missing water quality data imputation predominantly focus on temporal datasets, neglecting the crucial spatial dimensions that often influence water quality parameters. As a result of this, the deep learning models cannot relate the dataset of a station and the stations neighboring it, thus producing less accurate results. For instance, Huan *et al.*[14] compared the performance of spatial–temporal deep learning models with common DL models such as LSTMs and GRU for river water quality prediction. The paper presented the superiority of spatial–temporal models as well as accentuating the importance of incorporating spatial datasets for improving the prediction accuracy of the model. Similarly, Masolele *et al.*[15] compared the performance of temporal, spatial, and spatio-temporal deep learning models for the land-use classification task and revealed that spatio-temporal models achieved a substantially higher F1-score that other models. This study showcases the value of spatial–temporal deep learning models for effectively addressing the tasks of identifying intricate data patterns in a scalable and cost effective manner. Second, deep learning models perform differently under different scenarios like missingness patterns as well as dataset size. For instance, the Generative Adversarial Imputation Network (GAIN) is shown to excel in the case of Missing Completely at Random (MCAR) and crumple in the cases of Missing at Random (MAR) and Missing Not at Random (MNAR). Similarly, conventional models like MICE and missForest are well suited to handle missing data having limited sample size in comparison to deep

learning models like GAIN and VAE.[13] This inconsistency can become a burden for individuals, who do not have specific knowledge in the AI field while selecting the best suitable imputation model as per the scenarios. The third limitation is that the DL model tends to consume more computational resources when compared to other models. For instance, the self-attention mechanism in deep learning models requires more computational time and memory usage which increase quadratically with each time step posing a significant challenge to developing countries with limited computational resources.[16]

Proposed framework

The above-mentioned limitations, thus, necessitate an effective approach to tackle them, and the section herewith presents the solution. As it is evident that the spatial aspect is crucial to be integrated with existing water quality data imputation, one way for solving this issue is assuming each water quality recording station as a node and implementing an attention mechanism to provide weights to those stations neighboring the concerned stations. An analogy can be drawn from the study done by Wu *et al.*[17] in traffic studies. The location of the sensors in the study is analogous to the water quality data recording stations and the predicted variables of traffic flow conditions are analogous to the water quality parameters. This study revealed that incorporating spatial–temporal aspects resulted in their model to outperform the other state-of-the-art baseline models in the traffic conditions predictions. This underscores how implementation of spatial–temporal aspects in water quality data imputation can significantly increase the model performance.

Similarly, as discussed, the sensitivity of the model depends upon the different missing scenarios of the data. Preferably, some models perform best based on the type of missingness or dataset size or source of data collection. A potential resolution for this issue has been presented by Choi *et al.*[18]. The study presented a framework of the dynamic weighing ensemble model which incorporated ten different imputation models that showed the least RMSE value in comparison to standalone state-of-the-art imputation models in water quality data of different dataset sizes and hydrological sources. This highlights that incorporating a dynamic weighing ensemble model can be a way to tackle the issue of model adaptability for different missingness scenarios.

Since the above-described approaches have imbued improved accuracy and robustness within vanilla deep learning models, it is natural to consider that combining both approaches yields even a powerful model. A recent study by Zhao *et al.*[19] revealed the apparent superiority of this combination for predicting land subsidence. Because their model considered the possible spatial–temporal heterogeneity in the land and combined four deep learning models into an ensemble, they concluded that the spatial–temporal ensemble model was superior to those four individual models. Additionally, the study by Chen *et al.*[20] also revealed that an automated machine-learning-assisted ensemble framework (AutoML-Ens) which incorporates the spatial–temporal environmental changes and dynamic weighing ensemble approach was superior to other conventional statistical and ensemble models. These studies prove not only compatibility but also strong performance when combining spatial–temporal data with dynamic ensemble models. While these kinds of ensemble models are computationally demanding,[21] these studies have not made any efforts towards mitigating the heavy memory and computational power usage.

The prospects of making deep learning models more computationally efficient have been explored in different domains. For instance, Sun *et al.*[22] focused on introducing linear or near-linear approaches for reducing computation which is the solution to reducing the high CPU usage. However, this method is not promising as in general, ensemble models tend to bottleneck memory constraints rather than CPU usage. Additionally, computation of spatial–temporal data also requires high memory space due to the increment in the number of input parameters. In order to resolve the memory issue as such, Dao *et al.*[23] proposed an I/O (Input/Output) aware mechanism method which focuses on reading and writing of data in different levels of memory, resulting in reduced memory usage and accelerated computation. In fact, the concept of an I/O aware mechanism has been implemented in a Large Language Model (LLM), MAMBA which demonstrated 4–5 times faster processing speed during inference than the baseline model of ChatGPT, Transformers.[24] This improved efficiency

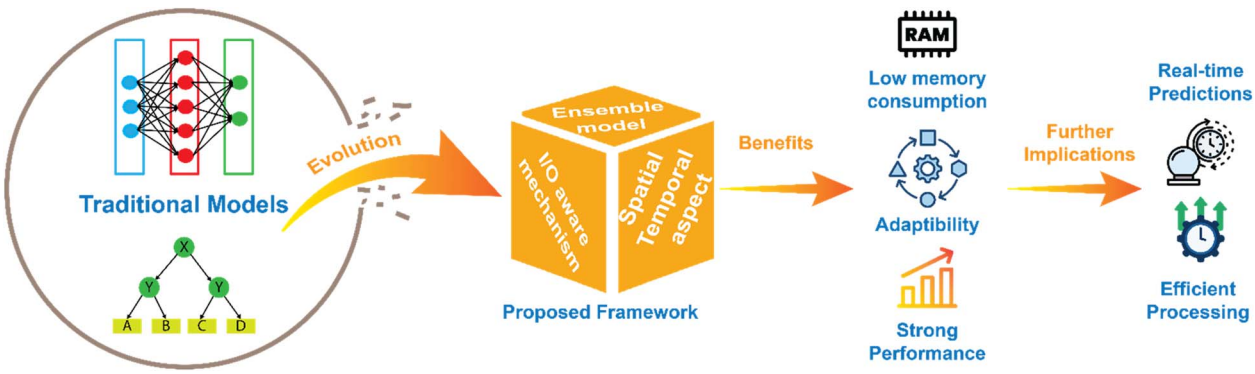

Fig. 1 Proposed framework with benefits and future implications.

can be attributed to hardware level coding through low-level languages and access to memory at different levels during reading and writing variables, which is not possible in common python interfaces such as PyTorch and Tensorflow.

Current water quality imputation models fail to simultaneously address the spatial–temporal dependencies, dynamic adaptability, and computational constraints in water quality data imputation despite their importance. There are fragmented approaches but they restrict the progress in the field, making water quality data imputation context-dependent and difficult to generalize. To break this cycle, we propose a paradigm shift—a novel framework unifying three points of a triangle—spatial temporal aspect, Dynamic ensemble modelling, and I/O aware mechanisms into a cohesive, scalable solution (Fig. 1). Unlike previous methods that are case-dependent, this framework redefines water quality data imputation as an intelligent, adaptive, and resource-efficient process. This framework represents a new standard in AI-driven environmental data reconstruction, offering a transformative blueprint for future water resource management.

Real-world implementation of the framework

Deep learning models are gaining popularity in the domain of water quality data imputation. Lee *et al.*[25] utilized the LSTM model for imputation of missing dissolved oxygen data in an eel recirculating aquaculture system. Another recent study assessed the potential of KNN in imputing real-time missing data in an aquaculture tail water treatment plant. This study also presented the importance of a correct imputation model for better real-time water quality prediction.[26] Similarly, Zhang *et al.*[5] implemented ten different imputation models for the water quality dataset of the Great Barrier Reef catchment and Iowa Water Quality Information System in near real-time through cloud based data imputation. However, as discussed in the previous section such deep learning models suffer from issues related to accuracy, generalizability, and memory consumption. For instance, Bai *et al.*[27] highlighted that in comparison to normal deep learning models like LSTM and GRU, graph neural networks which incorporate spatial–temporal data perform better prediction for the groundwater level in terms of all evaluation metrics. Similarly, Yee Wong *et al.*[28] created a novel ensemble model by stacking various machine learning models and compared its performance with ten different standalone models for classifying the water quality index based on 23 input variables. The stacked ensemble model outperformed all other machine learning models in tackling imbalanced water quality data in terms of generalizability and robustness of the model. Additionally, Khan *et al.*[29] also stated that ensemble models like Random Forest (RF) consumes a lot of memory while training on spatial–temporal datasets. The paper developed a novel ensemble of different machine learning models like CNN, LSTMs, RF, and GBM for water quality assessment using spatio-temporal data. While the model outperformed the complex machine learning models, it consumed the highest amount of training time and memory. The authors have only suggested generic solutions for reducing computational time and cost without any efforts towards implementing the solution.

Our proposed framework suggests the mitigation of these limitations *via* incorporation of novel concepts such as spatial–temporal datasets, dynamic ensemble modelling and I/O aware mechanisms. As depicted in Fig. 2, combination of these three concepts can yield a model showcasing excellent accuracy, and adaptability while consuming low computational resources. These strengths make our model suitable for areas with diverse geographic and economic conditions as well. Developed nations, with their advanced computational resources, can train and refine our model to develop a highly optimized version for water quality data imputation. This pre-trained model can then be deployed in underdeveloped regions, where data availability and computational capacity are limited. By leveraging a transfer learning approach, these regions can adapt the model to their specific needs while benefiting from its inherent adaptability and low memory requirements. Additionally, similar to other traditional imputation models, our system, owing to its lightweight nature, can be integrated with cloud-based platforms for real-time processing, making it more accessible as well as cost friendly to the users. This approach ensures advancements in addressing the current water quality data imputation problem with minimal computational demand, promoting equitable access to cutting-edge technological advancements even in resource-constrained settings, along with additional benefits depicted in Fig. 1.

The integration of this model into the national water management strategies can significantly benefit the environmental authorities worldwide. A lack of accurate data on critical water quality parameters often leads to severe ecological degradation. For instance, imbalances in oxygen production and oxygen consumption in an ecosystem can harm aquatic life, like fish and shrimp.[30] By imputing missing data in real-time,

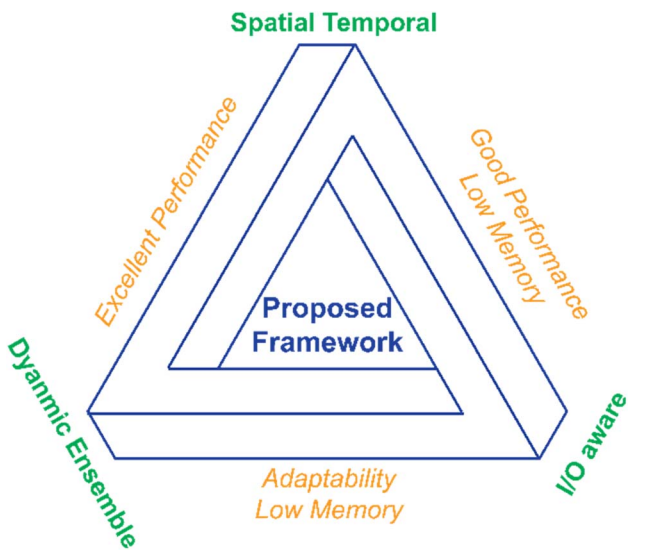

Fig. 2 Benefit of incorporating spatial–temporal data, dynamic ensemble model, and an I/O aware mechanism within the proposed framework.

agencies such as U.S. Environmental Protection Agency (EPA) and India's Central Pollution Control Board (CPCB) can make more informed decisions on pollution control and water safety. However, conventional models often lack the generalizability required to accurately impute the missing data in all scenarios. For example, Random Forest is less suitable for regression, or the SVM model is often challenging to scale up to large datasets and is more sensitive to parameter choices.[1] The lack of generalizability in these models can also necessitate the expert knowledge for implementation and maintenance.

In contrast, the proposed framework, with its adaptability and inclusion of spatial–temporal information (Fig. 2), can accurately impute the missing water quality data regardless of the missing data's nature. Furthermore, its generalizability makes it a user-friendly tool for water resource managers without extensive AI expertise. This approach enables policymakers to set higher standards for data-driven decision-making, ensuring compliance with global environmental policies, such as the United Nations Sustainable Development Goals (SDG 6: Clean Water and Sanitation) and the EU Water Framework Directive. Thus, this framework can serve as a robust foundation for evidence-based policymaking at both national and international levels.

Future research direction

The framework suggested in this perspective aims to resolve some of the critical issues faced while applying deep learning models in the domain of water quality data. Adhering to the philosophy of improving accuracy, generalizability and low memory consumption, we propose using spatial–temporal data, within dynamic weighting ensemble models, *via* an I/O aware mechanism. Although our framework can serve as a silver bullet in the domain of water quality data imputation, it is not devoid of limitations. First of all, spatial–temporal models are computationally intensive, and adding a dynamic weighing ensemble framework could significantly increase the memory requirements. Careful implementation of the framework is required such that increased memory demand does not nullify the gains obtained from implementing the I/O aware. Alternatively, to decrease the memory demand from spatial–temporal data, one can implement a statistical approach such as feature engineering, to simplify the input variables in the model. Second, the management of spatial–temporal dependencies along with dynamically adjusting ensemble weights requires proper interaction between these components. However, this can be mitigated by developing proper architectural designs to ensure flawless coordination between those components and preventing any conflicts. Lastly, I/O aware optimizations may require low level programming and hardware dependency which limits the applicability of the model to cross platforms. The problem can be resolved by designing portable I/O interfaces through standard abstractions by the hardware and software experts, especially during the initial stage of the project deployment. Although there are few limitations even within our proposed frameworks, its resolution can be readily implemented *via* coordinated efforts from researchers across diverse disciplines. Thus, we urge the researchers to implement the proposed framework (Fig. 1) and pave the pathway towards developing a novel, generalized, and memory-efficient deep learning architecture. By refining these aspects, researchers can unlock its full potential, not only in water quality management but also in broader applications such as wastewater analysis, meteorology, and energy systems, where accurate data imputation is essential.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Author contributions

Writing original draft, visualization, literature review, data collection and analysis: Ishan Prasad Banjara; writing original draft, visualization, literature review, data collection and analysis: Suman Poudel; writing original draft, visualization, literature review, data collection and analysis: Kalam Pariyar; writing original draft, visualization, literature review, data collection and analysis: Deepesh Upreti; literature-review and critical comments: Antigoni Zafeirakou; conceptualization, supervision, writing – review and critical editing: Shukra Raj Paudel.

Conflicts of interest

The authors declare they have no competing interests.

Acknowledgements

The authors acknowledge Pulchowk Campus Administration, Department of Civil Engineering, Pulchowk Campus and all the universities/institutes for their unwavering support during the preparation of this article.

References

1 J. Dong, Z. Wang, J. Wu, J. Huang and C. Zhang, A water quality prediction model based on signal decomposition and ensemble deep learning techniques, *Water Sci. Technol.*, 2023, **88**(10), 2611–2632.
2 S. Poudel, A. Shrestha, N. Kandel, S. Adhikari and S. R. Paudel, A review of reclaimed water reuse for irrigation in South Asian countries, *ACS EST Water*, 2023, **3**(12), 3790–3806.
3 J. Alcamo, Water quality and its interlinkages with the Sustainable Development Goals, *Curr. Opin. Environ. Sustain.*, 2019, **36**, 126–140.
4 J. O. Ighalo and A. G. Adeniyi, A comprehensive review of water quality monitoring and assessment in Nigeria, *Chemosphere*, 2020, **260**, 127569.
5 Y. Zhang and P. J. Thorburn, Handling missing data in near real-time environmental monitoring: A system and a review

of selected methods, *Future Gener. Comput. Syst.*, 2022, **128**, 63–72.
6 Y. F. Zhang, P. J. Thorburn, W. Xiang and P. Fitch, SSIM—A Deep Learning Approach for Recovering Missing Time Series Sensor Data, *IEEE Internet Things J.*, 2019, **6**(4), 6618–6628.
7 D. Sierra-Porta, Assessing the impact of missing data on water quality index estimation: a machine learning approach, *Discover Water*, 2024, **4**(1), 11.
8 W. C. Lin and C. F. Tsai, Missing value imputation: a review and analysis of the literature (2006–2017), *Artif. Intell. Rev.*, 2020, **53**(2), 1487–1509.
9 H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen, Methods for imputation of missing values in air quality data sets, *Atmos. Environ.*, 2004, **38**(18), 2895–2907.
10 Q. Shang, Z. Yang, S. Gao and D. Tan, An Imputation Method for Missing Traffic Data Based on FCM Optimized by PSO-SVR, *J. Adv. Transp.*, 2018, **2018**, 1–21.
11 Y. Ali, F. Hussain and M. M. Haque, Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review, *Accid. Anal. Prev.*, 2024, **194**, 107378.
12 R. Ratolojanahary, R. Houé Ngouna, K. Medjaher, J. Junca-Bourié, F. Dauriac and M. Sebilo, Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset, *Expert Syst. Appl.*, 2019, **131**, 299–307.
13 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 Syst. Appl.*, 2023, **227**, 120201.
14 J. Huan, W. Liao, Y. Zheng, X. Xu, H. Zhang and B. Shi, A deep learning model with spatio-temporal graph convolutional networks for river water quality prediction, *Water Supply*, 2023, **23**(7), 2940–2957.
15 R. N. Masolele, V. De Sy, M. Herold, D. Marcos, J. Verbesselt, F. Gieseke, *et al.*, Spatial and temporal deep learning methods for deriving land-use following deforestation: A pan-tropical case study using Landsat time series, *Remote Sens. Environ.*, 2021, **264**, 112600.
16 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, Attention Is All You Need, *arXiv*, 2017, preprint, arXiv:1706.03762, DOI: **10.48550/arXiv.1706.03762**, [cited 2024 Aug 13], available from: **https://arxiv.org/abs/1706.03762**.
17 Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang and C. Zhang, Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, ACM, CA USA, 2020, pp. 753–763, [cited 2024 Aug 7], available from: DOI: **10.1145/3394486.3403118**.
18 J. Choi, K. J. Lim and B. Ji, Robust imputation method with context-aware voting ensemble model for management of water-quality data, *Water Res.*, 2023, **243**, 120369.
19 B. Zhao, G. Wu, J. Li, Q. Wu and M. Deng, Spatio-Temporal Heterogeneous Ensemble Learning Method for Predicting Land Subsidence, *Appl. Sci.*, 2024, **14**(18), 8330.
20 H. Chen, T. Wang, Y. Zhang, Y. Bai and X. Chen, Dynamically weighted ensemble of geoscientific models via automated machine-learning-based classification, *Geosci. Model Dev.*, 2023, **16**(19), 5685–5701.
21 A. Mohammed and R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, *J. King Saud Univ. Comput. Inf. Sci.*, 2023, **35**(2), 757–774.
22 Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, *et al.*, Retentive Network: A Successor to Transformer for Large Language Models, *arXiv*, 2023, preprint, arXiv:2307.08621, DOI: **10.48550/arXiv.2307.08621**, [cited 2025 Mar 7], available from: **https://arxiv.org/abs/2307.08621**.
23 T. Dao, D. Y. Fu, S. Ermon, A. Rudra and C. Ré, FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, *arXiv*, 2022, preprint, arXiv:2205.14135, DOI: **10.48550/arXiv.2205.14135**, [cited 2024 Aug 7], available from: **https://arxiv.org/abs/2205.14135**.
24 A. Gu and T. Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces, *arXiv*, 2023, preprint, arXiv:2312.00752, DOI: **10.48550/arXiv.2312.00752**, [cited 2025 Mar 7], available from: **https://arxiv.org/abs/2312.00752**.
25 S. Lee, D. Jeong, J. Choi, S. Jo, D. Park and J. Kim, LSTM model to predict missing data of dissolved oxygen in land-based aquaculture farm, *ETRI J.*, 2024, **46**(6), 1047–1060.
26 Z. Deng, J. Wan, G. Ye and Y. Wang, Data-driven prediction of effluent quality in wastewater treatment processes: Model performance optimization and missing-data handling, *J. Water Process Eng.*, 2025, **71**, 107352.
27 T. Bai and P. Tahmasebi, Graph neural network for groundwater level forecasting, *J. Hydrol.*, 2023, **616**, 128792.
28 W. Yee Wong, K. Hasikin, M. K. A. Salwa, S. Abdul Razak, H. Farzana Hizaddin, M. M. Istajib, *et al.*, A Stacked Ensemble Deep Learning Approach for Imbalanced Multi-Class Water Quality Index Prediction, *Comput. Mater. Continua*, 2023, **76**(2), 1361–1384.
29 K. M. Karthick Raghunath, S. B. Khan, P. Govindarajan, T. R. Mohammad Alojail, M. Alojail and T. R. Gadekallu, Machine learning-driven intelligent water quality assessment for enhanced drinking safety and real-time consumer awareness, *Hydrol. Res.*, 2025, **56**(2), 136–152.
30 Z. Xiao, L. Peng, Y. Chen, H. Liu, J. Wang and Y. Nie, The Dissolved Oxygen Prediction Method Based on Neural Network, *Complexity*, 2017, **2017**, 1–6.