Journal of Civil Engineering, Science and Technology
Volume 16, Issue 2, September 2025, 148-153

EDITORIAL SCOPE: A PERSPECTIVE ON DATA LIMITATIONS AND MACHINE LEARNING APPLICATIONS IN CIVIL AND ENVIRONMENTAL ENGINEERING

Danial Jahed Armaghani* and Haleh Rasekh
School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia

Date received: 18/09/2025 **Date revised**: 23/09/2025 **Date accepted**: 24/09/2025 **Date published**: 30/09/2025
*Corresponding author's email: Danial.JahedArmaghani@uts.edu.au
DOI: 10.33736/jcest.10861.2025

Abstract — Although data-driven machine learning (ML) techniques have been widely applied in civil and environmental engineering, their performance depends heavily on the availability of large, high-quality, and reliable datasets. In civil engineering, researchers often face significant challenges due to limited, incomplete, or inaccessible data. This editorial scope discusses the main reasons behind these challenges and explores potential solutions. It also emphasises the urgent need for standardised data collection practices and the development of new ML models with higher levels of reliability, accuracy, and generalisation.

2462-1382 / © 2025 The Author(s). Published by UNIMAS Publisher. This is an open access article under the CC BY-NC-SA 4.0 license (http://creativecommons.org/licenses/by-nc-sa/4.0/).

***Keywords*:** machine learning, civil and environmental engineering, high-quality dataset, limited/incomplete data, generalisation capacity

1.0 INTRODUCTION

Nowadays, rarely can you find a topic within the scope of civil and environmental engineering with no touch of machine learning (ML) technologies. Compared to traditional theories or fundamentals in the area of civil engineering, ML models are faster, more efficient and accurate in general. These techniques due to their nature are able to provide non-linear relationships between input and output variables. In regression problems, they are able to estimate continuous outputs with a high degree of accuracy such as predicting tunnel boring machine (TBM) performance [1], concrete compressive strength estimation [2], or prediction of bridge deck deterioration [3]. In classification issues, they can easily classify data into various classes (e.g., slope stability [4], soil classification [5], and water quality classification [6]) according to the provided features or input variables. Time-series problems are effective for modelling temporal dependencies and forecasting future trends with improved reliability such as prediction of landslide displacement [7], forecasting upcoming scour depth around bridge piers [8], and analysing construction safety [9].

Despite these advancements, these areas of research generally face a basic challenge: the shortage of existing reliable data in various fields such as rock and soil mechanics, transportation geotechnics, and health monitoring. Data collection and preparation in civil engineering are often expensive, difficult, time-consuming, and, in some cases, unsafe [10–12]. For example, rockburst is defined as a sudden release of energy in the rock mass. Because of its rare occurrence and specific nature, collecting reliable rockburst data is time-consuming and extremely challenging. In another example, when collecting reliable data for TBM performance prediction, although TBM parameters are recorded by the machine, it is still necessary to revisit the same mechanised excavation project (at a high cost margin), observe the tunnel face for mapping purposes panel by panel, conduct field tests, and collect block samples for further assessment and for creating an ML model. In the case of laboratory work, transferring samples to the laboratory, preparing samples, and conducting tests require significant cost and time. However, even with all these efforts, only a limited database can be provided, which is not suitable for ML models. The lack of sufficient data not only limits the development of accurate predictive ML models but also restricts the generalisation ability of these techniques, which is crucial for future investigations.

This editorial scope discusses the challenges that remain when limited data are available in civil and environmental engineering, their impacts on ML models, and potential strategies to address them.

2.0 CHALLENGES OF LIMITED DATA IN CIVIL ENGINEERING

Although in some limited areas of civil engineering such as traffic flow, urban design/planning and water resources, large datasets may be available for ML modelling [13–15], many other fields of civil engineering still face significant challenges in preparing sufficient data. For example, in geotechnical engineering, shortage of data is a common fact because soil and rock properties are site specific, difficult to access, require a high cost to measure [16], collecting reliable data samples needs extensive field observations, monitoring and measurements [17]. In addition, laboratory testing and in-situ monitoring are time-consuming and expensive. In structural engineering, although numerous structures are monitored, the preparation of comprehensive data samples in the laboratory is very difficult as most of the tests must be conducted at full scale or half scale [18]. Unlike some areas of research such as computer vision or natural language processing where large-size databases are available, databases in civil engineering are limited for the following reasons:

2.1. High Cost and Time Requirements

One of the reasons for the limited availability of data in civil engineering relates to high-cost and time associated with data transfer, data collection, sample preparation, sample setup and so on. In addition, field observations, mappings, measurements and long-term monitoring systems require financial resources and specific equipment. Additionally, large-scale and physical tests need a well-planned schedule and often need to be repeated multiple times to account for different materials or situations, human or equipment errors, and incomplete testing setups. These factors make it difficult to establish sufficiently large and diverse datasets for ML model development and validation.

2.2. Safety and Accessibility Issues

Another major reason for the limited data availability in civil and environmental engineering is related to safety and accessibility. Many civil and environmental engineering applications require work in unsafe environments such as deep excavations and underground construction, slopes and landslides, high-rise structures, situations under dynamic conditions, flooded areas, contaminated sites, and extreme weather conditions. These environments may pose significant risks to personnel. In addition, difficult-to-reach locations can be very dangerous and also complex if there is a need to collect data. These concerns restrict the frequency and type of data that can be collected and due to them, preparation of comprehensive data samples for ML application is a challenging task.

2.3. Incomplete and Inconsistent Data Samples

Incomplete and inconsistent data samples are a common and practical limitation in the area of civil and environmental engineering. If the data are collected from various sites or at different time periods, there may be missing values for several parameters which lead to having incomplete datasets. Sensor malfunctions, human errors during field observations, and gaps in long-term monitoring programs further contribute to dataset incompleteness. Sometimes, a comprehensive database from an old laboratory is available, however, there is a missing information about the tests, procedure, and other details, making it unusable for an ML-based study. In other cases, the tests may have been conducted properly, but reports or expert justifications are not available. As a result, in this situation, extensive data cleaning must be applied to determine the final database to be used and unfortunately, in most cases the final database is insufficient to develop a reliable ML-based model.

2.4. Confidentiality and Ownership Restrictions

In the area of civil and environmental engineering, many datasets are related to real-world projects such as bridges, tunnels, road constructions, dams and so on. Typically, these projects are executed by private companies, government agencies, or research institutions, which often impose strict access limitations due to proprietary, legal, or security concerns. Even, when the data are available, sharing may be restricted to protect intellectual property, sensitive information, or individual privacy. This limited availability of data results in ML models that cannot be used in similar projects due to reduced generalisation capabilities.

2.5. Lack of Standardised Data Practices

Gathering data from multiple sources becomes difficult when the measurement techniques, instrumentation procedures, specific standards, and other factors are different. This inconsistency can lead to errors or misinterpretation of data especially in modelling and proposing ML models. In addition, the absence of consistent data formatting, labelling, and uniform standards prevents the development of comprehensive and high-quality databases. As a result, researchers must spend a significant amount of time and energy on data pre-processing stage which further limits the generalisation capacity of the proposed ML techniques.

3.0 IMPACT OF LIMITED DATA ON ML MODELS

It is obvious that ML models should learn from the data provided. This is an important stage (the training part) as the performance of the model depends significantly on how it has been trained. Typically, a larger portion of the available data is used for an effective training procedure. Lack of sufficient data in this part will result in poor model performance, reduced accuracy, and limited generalisation ability, which may ultimately compromise the reliability of the predictions. The implications of limited data are further discussed in the following paragraphs:

One of the foremost implications of relying on restricted datasets is the increased susceptibility of models to overfitting. When data availability is limited, machine learning algorithms are prone to memorising training examples rather than extracting meaningful, generalisable relationships, which in turn compromises their predictive performance on new and unseen scenarios [19]. Equally important, small datasets rarely encompass the inherent variability of real-world conditions—for instance, soil heterogeneity, structural response diversity, or fluctuations in traffic patterns. This lack of representativeness introduces systematic biases and substantially diminishes the models' capacity for generalisation.

A further concern relates to the impracticality of employing advanced modelling techniques. Contemporary approaches such as deep learning require extensive, high-quality datasets often numbering in the hundreds of thousands or even millions to yield dependable results [20]. In civil engineering, where such large-scale datasets are seldom attainable, researchers are frequently constrained to utilise simplified models with more limited predictive accuracy. Consequently, the absence of comprehensive datasets not only curtails the effective application of state-of-the-art AI methodologies but also impedes the pace of innovation and the broader advancement of scientific inquiry within the discipline.

4.0 STRATEGIES TO ADDRESS DATA LIMITATIONS

To overcome the challenges discussed, several ways/strategies can be implemented. Some of these strategies have been successfully implemented in civil and environmental engineering and others have been used in broader science and engineering fields and could potentially be adapted for civil and environmental applications. The use of such techniques can be a good opportunity for researchers in this field to propose ML models that are more reliable, accurate, and generalisable, and that can be effectively applied in future projects. In the following, these strategies will be discussed:

4.1. Synthetic Data Generation

To minimise the lack of comprehensive datasets in civil and environmental engineering, synthetic data generation is considered as one of the most practical opportunities. Monte Carlo sampling, physics-based simulations, and generative models (e.g., generative adversarial networks) can be used to generate artificial datasets that are very similar to the conditions of that example [21, 22]. For example, within civil and environmental applications, those related to risk analysis and costly experiments where conducting large-scale field trials are impractical or unsafe, could be good areas of research. Through systematic generation, these techniques aim to provide logical alternatives to address data scarcity and can yield more robust outcomes, particularly when applied to testing datasets.

4.2. Hybrid Modelling

To reduce dependency on comprehensive datasets, hybrid modelling of physics-based models and ML approaches could be another solution. This combination leverages specific knowledge in civil and environmental engineering such as constitutive laws, structural mechanics, or geotechnical principles to make ML models to be seem more

meaningful [23]. A common example is to combine finite element modelling with ML predictive models that enable experts to capture both mechanistic behavior and data-driven patterns [24]. Such models are particularly advantageous when experimental data is sparse but theoretical understanding of the system is well established.

4.3. Transfer Learning

Transfer learning provides the opportunity of the use of pre-trained models in other related domains within civil and environmental engineering [19, 25]. For instance, the trained models for vibration analysis in mechanical engineering could serve as input for research such as structural health monitoring with a different objective. In another example, the trained image recognition models could be used as input to detect cracks in concrete structures or rock materials. This solution allows researchers to start with trained models rather than starting from scratch, therefore data requirements for modelling will be reduced.

4.4. Data Augmentation and Bootstrapping

To artificially expand the size and diversity of available databases, data augmentation techniques can be used. In this way, the number of training datasets will be increased and possibly sufficient for modelling ML applications. Techniques like resampling, noise injection, feature perturbation, or geometric transformations have been used in other areas and can be applied to civil and environmental problems [26, 27]. Similarly, bootstrapping methods are able to improve model robustness by creating multiple resampled datasets from the original data. These techniques can help to minimise data overfitting and improve generalisation level when the original datasets are small.

5.0 DISCUSSION

The problem of limited data in civil and environmental engineering is challenging for some reasons such as difficulty of measurement, safety concerns, incomplete available sources and so on. However, there is a need to develop ML models which are generalised enough to be used in further research or real-world applications. The main problem of the most published studies in this area is their limitation for further research or similar situations. One reason is related to the nature of materials and applications such as geomaterials (soil or rock) which are site specific. Nevertheless, the main reason is because of nature of ML behaviour which is applicable within the range of input parameters and their conditions.

In the broader field of science and engineering, the discussed strategies have been applied to propose ML models which are more generalised. Particularly, in civil and environmental applications, researchers have sought strategies to address the challenge of limited data. For example, Asteris and Armaghani [28] tried to use empirical equations and their results in preparation of datasets in predicting peak particle velocity (PPV) resulting from blasting. Then, they successfully showed that the proposed ML model built by a new database achieved a better PPV prediction accuracy as well as more generalisation capacity. He et al. [27] presented a robust solution for the issue of data shortage and class imbalance in ML, specifically targeting blast-induced overbreak prediction in tunnel engineering. By utilising advanced data augmentation approach, this study demonstrated significant improvements in the reliability, accuracy and generalisation capacity of ML models.

It seems that the number of studies that have tried to discuss strategies for resolving the limited reliable data in civil and environmental engineering is not large. Therefore, there is a good opportunity for researchers to focus more on this area of research where we have a significant shortage of proposing general ML models in costly real-world projects in civil and environmental applications such as tunnelling, bridge construction, road maintenance, and structural health monitoring. The adoption of advanced computational methods should be accompanied by transparent reporting of dataset size, quality, and limitations. Without such practices, the reproducibility and credibility of ML-based research will remain questionable.

6.0 CONCLUSIONS AND FUTURE WORKS

One of the main constraints to expanding ML applications in civil and environmental engineering is restrictions in data availability. The primary obstacles in collecting robust, extensive and large-scale datasets, which impede the reproducibility which impedes the reproducibility and scalability of ML models are factors such as cost, safety concerns, privacy issues, and confidentiality restrictions. Alternate approaches like synthetic data generation, hybrid modeling with the ability to integrate physics-based and data-driven methods, and transfer learning from

related domains can be applied by researchers to overcome the challenges abovementioned. The advancement of ML in civil and environmental engineering requires collective data-sharing efforts, underpinned by consistent practices in data acquisition, documentation, and benchmarking.

Moving forward, accurate and physically reliable and explainable predictions in future studies r require emphasis on the development. Moreover, the trustworthiness of published research works can improve by the development of clear dataset size, quality, and limitations reporting protocols. Secure and ethical data exchange without breaching confidentiality requires enough effort towards open-access databases, federated learning approaches, and multi-institutional partnerships.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Armaghani, D. J., Mohamad, E. T., Narayanasamy, M. S., Narita, N., & Yagiz, S. (2017). Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition. Tunnelling and Underground Space Technology, 63, 29-43.

[2] Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P., & Pilakoutas, K. (2021). Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. Cement and Concrete Research, 145, 106449.

[3] Rashidi Nasab, A., & Elzarka, H. (2023). Optimizing machine learning algorithms for improving prediction of bridge deck deterioration: A case study of Ohio bridges. Buildings, 13(6), 1517.

[4] Asteris, P. G., Rizal, F. I. M., Koopialipoor, M., Roussis, P. C., Ferentinou, M., Armaghani, D. J., & Gordan, B. (2022). Slope stability classification under seismic conditions using several tree-based intelligent techniques. Applied Sciences, 12(3), 1753.

[5] Pham, B. T., Nguyen, M. D., Nguyen-Thoi, T., Ho, L. S., Koopialipoor, M., Quoc, N. K., & Van Le, H. (2021). A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. Transportation Geotechnics, 27, 100508.

[6] Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. Water Quality Research Journal, 57(3), 152-164.

[7] Yang, B., Yin, K., Lacasse, S., & Liu, Z. (2019). Time series analysis and long short-term memory neural network to predict landslide displacement. Landslides, 16(4), 677-694.

[8] Yousefpour N., Downie S., Walker S., Perkins N., & Dikanski H. (2021) Machine learning solutions for bridge scour forecast based on monitoring data. Transp Res Rec 2675(10):745–763

[9] Cao, H., & Goh, Y. M. (2019). Analyzing construction safety through time series methods. Frontiers of Engineering Management, 6(2), 262-274.

[10] Baghbani, A., Choudhury, T., Costa, S., & Reiner, J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. Earth-Science Reviews, 228, 103991.

[11] Liu, H., Su, H., Sun, L., & Dias-da-Costa, D. (2024). State-of-the-art review on the use of AI-enhanced computational mechanics in geotechnical engineering. Artificial Intelligence Review, 57(8), 196.

[12] He, B., Armaghani, D. J., Lai, S. H., He, X., Asteris, P. G., & Sheng, D. (2024). A deep dive into tunnel blasting studies between 2000 and 2023—A systematic review. Tunnelling and Underground Space Technology, 147, 105727.

[13] Cong, D. (2024). Research on Traffic Flow Prediction Using the MSTA-GNet Model Based on the PeMS Dataset. International Journal of Advanced Computer Science & Applications, 15(8), 529-539.

[14] Liang, Y., Ding, F., Liu, L., Yin, F., Hao, M., Kang, T., & Jiang, D. (2025). Monitoring water quality parameters in urban rivers using multi-source data and machine learning approach. Journal of Hydrology, 648, 132394.

[15] Refadah, S. S. (2025). Development in flood forecasting: A comprehensive review of complex and machine learning models. Physics and Chemistry of the Earth, Parts A/B/C, 103975.

[16] Kosmowski, F., Abebe, A., & Ozkan, D. (2020). Challenges and lessons for measuring soil metrics in household surveys. Geoderma, 375, 114500.

[17] Wang, X. F., Wang, C. J., Yue, W. V., Zhang, Z. J., & Yue, Z. Q. (2024). In situ digital testing method for quality assessment of soft soil improvement with polyurethane. Journal of Rock Mechanics and Geotechnical Engineering, 16(5), 1732-1748.

[18] Wu, Y., Chen, Y., Zeng, J., Jiang, Y., & Liu, Z. (2025). Long-Term Performance Analysis of Steel–Concrete Composite Beams Based on Finite Element Model Updating. Buildings, 15(8), 1374.

[19] Bao, N., Zhang, T., Huang, R., Biswal, S., Su, J., & Wang, Y. (2023). A Deep Transfer Learning Network for Structural Condition Identification with Limited Real-World Training Data. Structural Control and Health Monitoring, 2023(1), 8899806.
[20] El-Abbasy, A.A.A. (2025). Artificial intelligence-driven predictive modeling in civil engineering: a comprehensive review. J. Umm Al-Qura Univ. Eng.Archit. https://doi.org/10.1007/s43995-025-00166-5
[21] Unterlass, P.J., Erharter, G.H., Sapronova, A., & Marcher, T. (2023). A WGAN Approach to Synthetic TBM Data Generation. In: Gomes Correia, A., Azenha, M., Cruz, P.J.S., Novais, P., Pereira, P. (eds) Trends on Construction in the Digital Era. ISIC 2022. Lecture Notes in Civil Engineering, vol 306. Springer, Cham. https://doi.org/10.1007/978-3-031-20241-4_1
[22] Krüger, M. (2024). Synthetic Data Generation for the Enrichment of Civil Engineering Machine Data. In: Fottner, J., Nübel, K., Matt, D. (eds) Construction Logistics, Equipment, and Robotics. CLEaR 2023. Lecture Notes in Civil Engineering, vol 390. Springer, Cham. https://doi.org/10.1007/978-3-031-44021-2_18
[23] Demir, V., Uray, E., & Carbas, S. (2023). Modeling Civil Engineering Problems via Hybrid Versions of Machine Learning and Metaheuristic Optimization Algorithms. In: Bekdaş, G., Nigdeli, S.M. (eds) Hybrid Metaheuristics in Structural Engineering. Studies in Systems, Decision and Control, vol 480. Springer, Cham. https://doi.org/10.1007/978-3-031-34728-3_11
[24] Vadyala, S. R., Betgeri, S. N., Matthews, J. C., & Matthews, E. (2022). A review of physics-based machine learning in civil engineering. Results in Engineering, 13, 100316.
[25] Yano, M. O., Figueiredo, E., da Silva, S., & Cury, A. (2023). Foundations and applicability of transfer learning for structural health monitoring of bridges. Mechanical Systems and Signal Processing, 204, 110766.
[26] Li, L., & Betti, R. (2023). A machine learning-based data augmentation strategy for structural damage classification in civil infrastructure system. Journal of Civil Structural Health Monitoring, 13(6), 1265-1285.
[27] He, B., Armaghani, D. J., Lai, S. H., Samui, P., & Mohamad, E. T. (2024). Applying data augmentation technique on blast-induced overbreak prediction: Resolving the problem of data shortage and data imbalance. Expert Systems with Applications, 237, 121616.
[28] Asteris, P. G., & Armaghani, D. J. (2025). An empirical-driven machine learning (EDML) approach to predict PPV caused by quarry blasting. Bulletin of Engineering Geology and the Environment, 84(4), 1-17.