Effect of Missing Data on Machine-Learning Algorithms for Real-time Safety Monitoring in Scaffolds

Laura Alvarez[1], Mahendra Ghimire[2] and JeeWoong Park[3]

[1,2,3]Department of Civil and Environmental Engineering and Construction, University of Nevada Las Vegas, USA
alvarezq@unlv.nevada.edu, ghimim2@unlv.nevada.edu, jee.park@unlv.edu

Abstract –

In the context of real-time data acquisition and processing, dealing with missing data (MD) is a common challenge that can compromise the quality and effectiveness of machine learning (ML) algorithms. Previous research focuses on creating a real-time safety monitoring system that predicts safety conditions in scaffolds by analyzing strain measurements from sensors placed in the structure's columns. However, it does not address the effect of sensor failures and the resulting MD. This paper explores how the presence of MD, caused by faulty sensors, affects the performance of eight ML algorithms in a safety monitoring scaffolding system: gaussian naive Bayes (GNB), random forest (RF), multi-layer perceptron (MLP), support vector machine (SVM), decision tree (DT), XGBoost (XGB), logistic regression (LR), and linear support vector classification (LSVC). This study identifies how these algorithms perform when processing datasets with missing values. As the amount of MD in the datasets increases, there is a consistent negative influence on the performance of each algorithm, resulting in reduced predictive accuracy. Among all the tested ML algorithms, RF and DT have shown to be the most sensitive to MD.

Keywords –

Scaffolds; classification; missing data; machine learning.

1 Introduction

Real-time data acquisition and processing often face the challenge of missing data (MD), impacting data quality and machine learning (ML) algorithm performance [1], impacting pattern identification [2]. Strategies to handle MD are crucial [3] because its presence introduces a risk of process failures, failing to accurately represent the true reality of the system [4]. The collective findings from several studies [5–8] underscore the significant influence of MD on the ML algorithms' performances, and stress the necessity of handling MD effectively to ensure accurate and dependable ML results.

Previous research focuses on real-time monitoring of intricate scaffolds using ML techniques to forecast safety conditions [9]. This study delves into a technique for categorizing instances of scaffold failure and accurately predicting safety conditions, using data from strains installed on the scaffold columns. The authors successfully improved the accuracy of ML classification through a self-multiplication technique [10]; nonetheless, prior research did not account for the influence of sensor failures and the subsequent absence of data on the system's acquisition. In this specific case, MD can cause an incorrect prediction in the scaffold structure's safety conditions. For instance, if the scaffold structure is about to overturn due to unbalanced loads, the MD in a sensor measurement can result in a safe classification instead of an overturning one.

These real-time sensing systems should demonstrate high accuracy in safety monitoring to promptly detect temporary structures' potential structural failures. The primary purpose of research toward reliable safety predictions is to safeguard the lives of workers [10]. Accurate predictions help to identify potential hazards or structural failures in advance, allowing for timely interventions or preventive measures to protect workers from injuries. This could involve reinforcing a structure, evacuating an area, or adjusting working conditions to prevent accidents. In addition, an accurate safety prediction can save costs associated with medical expenses, structural or property damage, legal liabilities, and potential project delays.

In response, this project studies the problem of MD in a scaffold dataset. It investigates MD effects on the performance of eight ML classification algorithms: gaussian naive Bayes (GNB), random forest (RF), multi-layer perceptron (MLP), support vector machine (SVM), decision tree (DT), XGBoost (XGB), logistic regression (LR), and linear support vector classification (LSVC).

2 Literature Review

To study the effects of MD on data mining processes, it is crucial to understand MD mechanisms [11]. The

methods used to handle missing values are often based on assumptions tied to the underlying mechanisms causing the MD [12]. Four mechanisms of the occurrence of MD are defined [13]: when the entry is not supposed to have a value in a given field, it is said that it is structurally missing (SMD) [14]. Missing completely at random (MCAR) is when the cause of data absence is independent from the observed and unobserved entries. Missing at random (MAR) is when the cause of data absence is related to the other observed entries. Finally, missing not at random (MNAR) is when the cause of data absence is related to the missing entry and the other observed entries. According to the previous definitions, this study deals with MCAR data, because the missing values originate when a sensor failure occurs due to a communication error or device malfunction. This event is not related to any other measured variable in the system.

The influence of MCAR on ML algorithms for safety monitoring cannot be overstated. Studies [5,8] highlight how the presence of MCAR significantly distorts the integrity of datasets, leading to biased predictions and compromised algorithmic performance. This bias stems from the randomness of MD, affecting the statistical properties of the dataset and, consequently, the MS models' training and generalization capabilities [7].

In a comprehensive review [15] of 152 ML-based clinical prediction model studies, the reporting quality regarding MD was generally poor, aligning it with similar reviews. Excluding participants with MD and insufficient details was found to be a common practice of MD handling methods. Other studies [16] utilized techniques like mean imputation or complete-case analysis for healthcare data. The review highlighted the pressing need for improved reporting guidelines, adherence, and understanding the repercussions of improper MD handling in ML-based prediction studies in healthcare.

The gaps and challenges related to MD are evident in ML-based heart disease prediction models [17]. The limited exploration of MD effects reveals a lack of comprehensive understanding regarding model performance. The study demonstrates by varying accuracy percentages in predicting heart disease, yet fail to explicitly address how to handle MD or its influence on these outcomes. This highlights the need for a more nuanced comprehension of data preprocessing methods. Another study [6] centers on how MD affects ML algorithms used in hydrologic predictions and proposes a method to fill in the gaps. A comprehensive overview [18] of MD in ML emphasizes the importance of appropriately addressing and evaluating various imputation techniques. Additionally, the study notes the use of smaller, domain-specific datasets, accenting the need for exploring MD handling in larger, real-world datasets with diverse features.

Applications in construction [19] focus on critically evaluating concrete strength predictions for enhanced sustainability. However, challenges arise due to MD, noise, and model interpretability. The authors used a dataset of manufactured sand concrete and various ML algorithms to demonstrate predictive performance. They found it necessary to bridge MD concerns and enhance interpretability for reliable concrete strength predictions in construction applications.

In safety applications, researchers addressed MD issues on high-plateau flights [20] by proposing an improved method based on least squares support vector machines (LS-SVM). This method approaches the challenges placed by missing or abnormal quick access recorder data due to harsh environmental conditions. Through advanced ML techniques, this method enhances the reliability and accuracy of flight data processing and analysis, contributing to improved aviation safety.

Estimations of increased risk of crashes on freeways need to be accurate and reliable, but the utilization of real-time traffic data in proactive safety management systems is lacking due to MD. To fill these gaps, past research [21] has proposed a framework for real-time risk assessment on freeways by integrating data from multiple detection systems, real-time weather, and roadway geometry. The development of the framework mitigates the effects of MD, contributing to the system with high estimation accuracy, robustness, and reliability.

To detect high-severity accidents in the construction industry, accurate prediction models are needed. While occupational accidents are common in construction, the challenge lies in determining the combination of preprocessing techniques that yield the most accurate severity prediction, considering issues such as MD, outliers, feature scaling, and imbalanced class distribution. Specifically, in dealing with MD, the researchers [22] have experimented with different scenarios of preprocessing techniques to determine the best combination. One of the scenarios involved not removing MD, indicating that missing values were retained in the dataset rather than being imputed or deleted. This approach allowed the model to learn from the available data without discarding potentially valuable information.

Prior studies developed a real-time safety monitoring system [9] based on strain measurements from sensors installed in the columns of the structure but did not consider the implications of MD introduced by sensor malfunctions. The previous articles contribute to understanding the influence of MD and how to improve model accuracy, comparability, and reliability in different areas of study. These contributions are used as guidelines, as the nature of the described problems present similar challenges to the ones found when a faulty sensor generates MD in a scaffold safety

monitoring system's dataset. Leveraging existing knowledge from various domains helps in developing strategies to handle MD effectively in the context of scaffold safety monitoring systems. Understanding how MD affects the performance of various ML algorithms contributes to the development of strategies to address specific types of MD and the mitigation of their impacts on algorithmic performance.

3 Objective and Scope

This research aims to assess the influence of MD, stemming from sensor failures, on the ML algorithms' predictive accuracy and performance in a real-time safety monitoring system for scaffolds. This paper evaluates the performance of eight ML algorithms: GNB, RF, MLP, SVM, DT, XGB, LR, and LSVC when exposed to datasets that feature MD caused by the simulation of faulty sensors. This paper expands upon a previous study [10] through additional investigation, evaluation, and discussions of MD across various ML models. By employing a diverse set of classification algorithms, the study explores different modeling approaches to handle the same type of data and capture the patterns. While reviewing different safety-related studies conducted with ML application, it was observed that GNB [23–25], RF [26–28], MLP [26,28,29], SVM [26,28,30], DT [28,31,32], XGB [33–35], LR [32,36,37] and LSVC [26,28,38] are mostly employed by researchers. Therefore, these eight ML models are used in this investigation.

4 Approach

The proposed approach comprised four fundamental steps, as illustrated in Figure 1. General approach of the studyFirst, it conducts an in-depth analysis of the initial dataset, establishing a benchmark for understanding the complete data's characteristics. Subsequently, simulated MD is introduced to replicate sensor faults, enabling the evaluation of ML algorithms' performance under these conditions. Then, the algorithms are assessed for their predictive capabilities using both the original complete dataset and the data affected by missing values.

A crucial aspect involves progressively evaluating algorithm performance as the number of faulty sensors or MD increases, highlighting the repercussions of sensor failures on algorithm accuracy. This comprehensive approach provides insights into the ML classification algorithms' behavior in scenarios involving faulty sensors, aiding in understanding their robustness.

The results may contribute to making decisions about algorithm selection and implementation in real-world applications. The previous steps are explained and justified as follows:

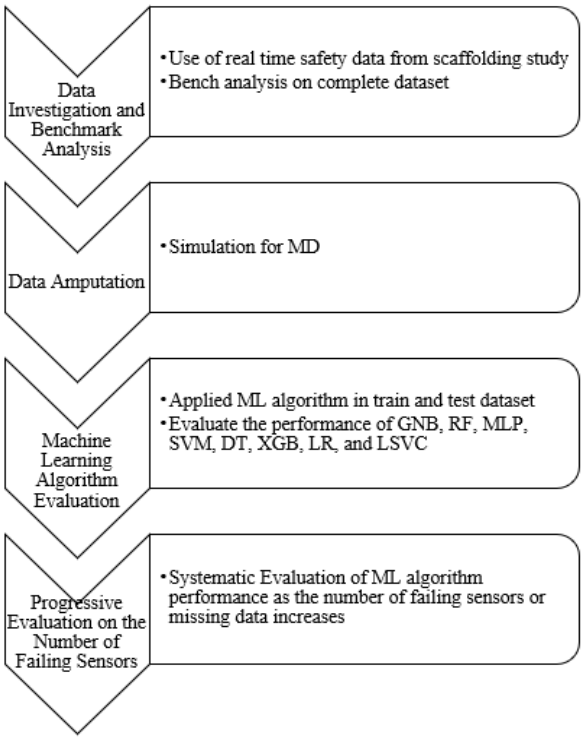

Figure 1. General approach of the study.

Step 1: Data analysis. This step involves thoroughly examining the characteristics of the original complete dataset. It is necessary to establish a baseline understanding of the nature of the system, including data distribution, variability of measurements, and any underlying patterns. This analysis provides essential context for subsequent steps and helps to identify potential issues or anomalies in the dataset.

Step 2: Data amputation. Simulated MD is introduced to replicate sensor faults, replicating real-world scenarios where data may be incomplete due to sensor malfunctions. This step is crucial to evaluate ML algorithms' performance in the presence of MD. By simulating sensor faults, it is possible to estimate how well the algorithms may handle these conditions and whether they can effectively make predictions despite incomplete information.

Step 3: ML algorithm evaluation. In this step, ML algorithms are trained and tested using the original complete datasets. This allows the evaluation of the algorithms ability to make accurate predictions, while operating under ideal conditions before introducing MD. This step is a starting point assessment to compare algorithm performance under different conditions and determine how MD affects predictive accuracy. It provides a quantifiable perception of the algorithms' robustness and their reliability in real-world applications.

Step 4: Progressive evaluation with increasing MD. As the number of faulty sensors or MD increases, algorithm performance is progressively evaluated. This step is essential to understand how algorithm accuracy changes as data quality deteriorates due to sensor failures. By systematically increasing the severity of MD scenarios, it is possible to identify thresholds where algorithm performance significantly degrades. This evidence helps to understand the limitations of ML algorithms in handling MD and informs decision-making regarding algorithm selection and deployment.

4.1 Real-time safety monitoring system

This study incorporates a real-time safety monitoring system designed for scaffold structures, based on a previous investigation [10]. The system relies on strain sensors embedded within the columns to gather crucial data indicative of potential scaffold failures. Figure 2 (Figure 5 in [10]), illustrates the scaffold's configuration, comprising 10 columns distributed across 3 stories with 20 sensors strategically positioned at various locations. The sensor measurements are used to predict the scaffold safety condition as overload, uneven, sideways, and safe by processing the data with ML classification algorithms in Python (GNB, RF, MLP, SVM, DT, XGB, LR, and LSVC). The algorithms are a built-in function from sklearn package, except for XGB, which has its own package called xgboost. Training was carried out with complete datasets without including MD.

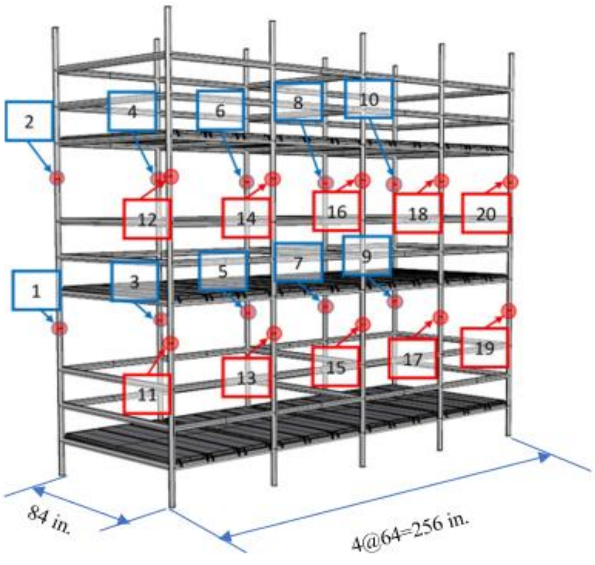

Figure 2. Scaffolding Structure with sensor measurement (Figure 5 in [10]).

4.2 Dataset

The original complete dataset was obtained from a previous study [9]. The dataset for strain measurement was created based on structural conditions observed in scaffold usage and distinguished between safe and unsafe conditions. The unsafe category was divided into global and local failures. Global failures involved overturning in both lateral X and Y directions. Local failures included uneven settlement and overloading issues. The scaffold model comprised 10 vertical members, each equipped with 20 evenly distributed strain-measuring sensors placed on them, as shown in Figure 2 (Figure 5 in [10]). The output is divided into 23 classes representing different safety conditions. The dataset contains 1,000 samples for each of the 23 classes.

4.3 MD generation and follow-up analysis

The initial dataset was generated based on the automated monitoring system's optimal operation conditions without accounting for MD [10]. Therefore, data amputation is needed to simulate the MCAR scenario.

This paper considers the progressive occurrence of failure. In this progression, the number of failing sensors starts from 1 to 20, equal to the entire sensors. While this may not make sense in a practical manner, it is important that this research investigates the effect of incremental sensor failures on the classification accuracy from a theoretical perspective. To simulate faulty sensors, or MD, null values were introduced, which are considered as 0.000με as an indication of absence of measurement.

The conducted analysis involved observing how different ML algorithms responded to the introduced failure cases in the dataset and thus overall prediction performance. By systematically introducing these cases across all sensors progressively, how each algorithm adapted to and handled the simulated errors was evaluated. algorithm presents a summary of the analysis output by showing the prediction accuracy in relation to the number of faulty sensors the entire dataset across all the tested ML algorithms.

The evaluation metric used for ML classification algorithms is accuracy instead of other metrics because is straightforward to understand. It represents the proportion of correctly predicted classes out of the total classes in the dataset and intuitively captures how well the ML model performs overall. In this specific case of study, accuracy is an appropriate metric due to the balanced dataset, i.e., the dataset contains 1,000 samples for each of the 23 classes.

Accuracy is evaluated by dividing the strain dataset into two portions: a training set and a testing set. This splitting is done to assess how well the ML algorithms perform in making predictions on new, unseen data. Specifically, the testing set comprises 20% of the entire dataset. The purpose of this separation is to use the larger portion (80%) as the training data to teach the model how

to make predictions based on patterns and information within that data. The remaining 20% is set aside as the testing data, which is kept separate and not used during the training phase.

After training with the training dataset, the obtained model is then evaluated using the testing dataset. The predictive accuracy is determined based on how well it predicts or classifies the outcomes within this separate testing data.

The following sections further discuss these results.

5 Results and Discussion

Prior to discussing the results, Table 1. ML algorithm performance in optimal operation conditions shows the accuracy achieved by each studied ML classification algorithm on a complete strain dataset. As this is based on a complete dataset, high accuracies are anticipated.

Table 1. ML algorithm performance in optimal operation conditions.

ML Algorithm	Accuracy
RF	0.9998
XGB	0.9998
DT	0.9991
SVM	0.9989
LR	0.9974
MLP	0.9941
GNB	0.9937
LSVC	0.9937

The high accuracy percentages reported in the study for various ML algorithms indicate their effectiveness in correctly classifying a complete dataset without MD presence. The reported accuracies in Table 1 are achieved under optimal operating conditions without any faulty sensor performing. In such conditions, these algorithms perform up to 99%, achieving near-perfect accuracy (100%). Note that more complex algorithms with a larger number of hyperparameters are prone to overfitting and have difficulty handling MD, especially if not appropriately tuned or validated.

Figure 3. Effect of MD generated from faulty sensors on the performance of ML classification algorithms while processing a strain dataset of a scaffold structure.illustrates how the accuracy of the ML classification algorithms is negatively affected by the inclusion of MD in the dataset. The accuracy of GNB drops gradually as the number of faulty sensors increases. With all sensors functioning, it achieves an accuracy of 99.37%. However, as the number of faulty sensors increases, the accuracy decreases progressively, reaching 4% when all sensors are faulty. It still presents at least 50% accuracy with about 8 faulty sensors out of 20 sensors.

RF shows a similar trend, but after the second sensor fails, the accuracy is more negatively affected than that of GNB.

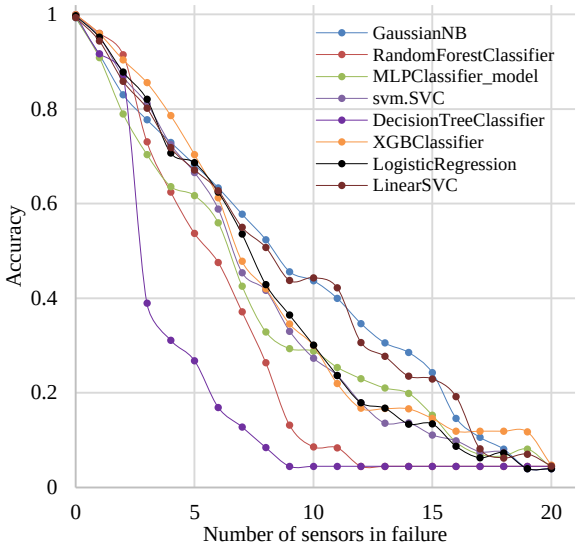

Figure 3. Effect of MD generated from faulty sensors on the performance of ML classification algorithms while processing a strain dataset of a scaffold structure.

DT shows a pattern where accuracy significantly drops from the third faulty sensor. The accuracy drops from analyzing with 2 sensors in failure to 3 sensors in failure, exhibiting the most detrimental rate at about 50%.

In general, all ML algorithms experience a decrease in accuracy as the number of faulty sensors increases. Except for DT's case, about 50% accuracy was still achieved when 5-7 failing sensors were included in the analysis. Overall, GNB and LSVC appear to be relatively more robust against MD compared to the other algorithms listed here. DT and RF exhibit higher sensitivity to MD, showing a significant decrease in accuracy as the number of faulty sensors increases.

Several reasons could contribute to RF and DT being more sensitive to MD compared to other algorithms, and they can be more susceptible to noisy or inconsistent data. DT create biased nodes when encountering MD, affecting subsequent decision-making and accuracy. MD, which can be considered a form of noise, might be challenging for these algorithms to handle effectively. These algorithms might lack the robustness to handle MD compared to other algorithms like GNB or LSVC, which can handle missing values more effectively due to their underlying mechanisms.

The study also evidences the decline in accuracy as the number of faulty sensors or amount of MD increases. This decrease in accuracy indicates that the algorithms are sensitive to MD, which may be due to their inability

to effectively handle such inconsistencies.

It is important to note that the behavior of these ML algorithms concerning MD can depend on various factors, including the specific dataset used in training. Tuning hyperparameters or using specific techniques for MD handling might help to mitigate these algorithms' sensitivity to missing values.

6 Conclusions

This study investigates the effect of MD caused by simulated sensor failures on the performance of ML classification algorithms used in a real-time safety monitoring system for scaffolds. The research focuses on assessing the predictive accuracy of eight ML algorithms when confronted with a dataset containing different amounts of MD.

It was evident that MD has a significant negative influence on ML algorithms' performance, and there is a need to effectively handle it to ensure accurate results. Previous studies in various domains have stated the challenges posed by MD and the need for improved reporting guidelines and understanding the repercussions of improper MD handling in ML-based prediction studies. In the context of scaffold safety monitoring systems, developing strategies to handle MD effectively is required, considering the influence MD has on algorithmic performance.

Results indicate that most ML algorithms achieve over 99% accuracy on the complete dataset, and RF, DT, and XGB exhibit the highest accuracy. However, when introducing progressively incremental MD, all ML algorithms experience a decrease in accuracy. Notably, GNB and LSVC appear relatively robust to MD, while DT and RF exhibit higher sensitivities to MD, showing a significant decrease in accuracy as the number of faulty sensors increases.

To conclude, MD significantly affects ML algorithms' performance, particularly DT and RF, which show higher sensitivity to MD. Possible reasons for this sensitivity include susceptibility to noisy data and a lack of robustness in handling MD compared to that of other algorithms. Dataset characteristics and proper handling techniques must be considered to mitigate the algorithms' sensitivities to MD.

7 Limitations

The study primarily focuses on theoretical simulations of sensor failures, introducing null values to simulate MD. This approach allows for controlled experimentation and may not fully capture the complexities of real-world sensor malfunctions. Although the assumption of sensor failing one by one does not reflect a realistic scenario, this study conducts a progressive analysis to systematically assess MD's effects. By simulating failures in a progressive manner, the study can observe the incremental degradation in algorithm performance with each additional failing sensor. This helps to understand MD's cumulative effect on the algorithms' reliability for safety monitoring applications.

As the study's primary objective is to investigate MD's influence on ML algorithms' performance, the hyperparameter tuning is not considered. The performance comparison of the ML algorithms is made with default hyperparameter settings; this approach provides a baseline for comparison and allows for assessing the algorithms' robustness without additional tuning. To extend and generalize the results, validation with real-time acquisition systems and actual sensor failures are required.

8 Future Work

As future work, MD handling techniques like imputation could be applied to fill in the MD. The real-time safety monitoring system's accuracy and precision could be improved by optimizing these methods. To do this, tuning hyperparameters or data preprocessing could be used. Once the imputation stage is complete, it would be possible to evaluate the imputed values' effects on the predicted safety conditions in the real-time safety monitoring system.

References

[1] Schauer, J. M., Diaz, K., Pigott, T. D., and Lee, J., Exploratory Analyses for Missing Data in Meta-Analyses and Meta-Regression: A Tutorial. *Alcohol and Alcoholism*, Vol. 57, No. 1, 2022, pp. 35–46. https://doi.org/10.1093/ALCALC/AGAA144

[2] Alvarez Quiñones, L. I., Lozano-Moncada, C. A., and Bravo Montenegro, D. A., Machine Learning for Predictive Maintenance Scheduling of Distribution Transformers. *Journal of Quality in Maintenance Engineering*, Vol. 29, No. 1, 2023. https://doi.org/10.1108/JQME-06-2021-0052

[3] Sharma, S., Chmaj, G., and Selvaraj, H., Sensor Data Restoration in Internet of Things Systems Using Machine Learning Approach. Vol. 611 LNNS, 2023, pp. 21–30. https://doi.org/10.1007/978-3-031-27470-1_3

[4] Rioux, C., and Little, T. D., Missing Data Treatments in Intervention Studies: What Was, What Is, and What Should Be. *International Journal of Behavioral Development*, Vol. 45, No. 1, 2021, pp. 51–58. https://doi.org/10.1177/0165025419880609

[5] Blomberg, L. C., and Ruiz, D. D. A., Evaluating the

Influence of Missing Data on Classification Algorithms in Data Mining Applications. *Brazilian Symposium on Information Systems*, 2013, pp. 734–743. https://doi.org/10.5753/SBSI.2013.5736

[6] Gill, M. K., Asefa, T., Kaheil, Y., and McKee, M., Effect of Missing Data on Performance of Learning Algorithms for Hydrologic Predictions: Implications to an Imputation Technique. *Water Resources Research*, Vol. 43, No. 7, 2007. https://doi.org/10.1029/2006WR005298

[7] Marlin, B. M., Missing Data Problems in Machine Learning. 2008. Retrieved 5 November 2023

[8] Radišić, B., Seljan, S., and Dunđer, I., Impact of Missing Values on the Performance of Machine Learning Algorithms. 2023. Retrieved 5 November 2023

[9] Cho, C., Park, J., Kim, K., and Sakhakarmi, S., Machine Learning for Assessing Real-Time Safety Conditions of Scaffolds. 2018. https://doi.org/10.22260/isarc2018/0008

[10] Sakhakarmi, S., Park, J., and Cho, C., Enhanced Machine Learning Classification Accuracy for Scaffolding Safety Using Increased Features. *Journal of Construction Engineering and Management*, Vol. 145, No. 2, 2019, p. 04018133. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001601/ASSET/534952B9-0A71-450E-982D-8B7E10F06AD2/ASSETS/IMAGES/LARGE/FIGURE8.JPG

[11] Enders, C. K., and London, N. Y., Applied Missing Data Analysis the Guilford Press. 2010.

[12] Kambach, S., Bruelheide, H., Gerstner, K., Gurevitch, J., Beckmann, M., and Seppelt, R., Consequences of Multiple Imputation of Missing Standard Deviations and Sample Sizes in Meta-Analysis. *Ecology and Evolution*, Vol. 10, No. 20, 2020, pp. 11699–11712. https://doi.org/10.1002/ECE3.6806

[13] Bo, N., Little, R. J. A., and Rubin, D. B., Statistical Analysis with Missing Data. *Population (French Edition)*, Vol. 43, No. 6, 1988, p. 1174. https://doi.org/10.2307/1533221

[14] Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., and Spangenberg, L., Evaluation of Different Approaches for Missing Data Imputation on Features Associated to Genomic Data. *BioData Mining*, Vol. 14, No. 1, 2021. https://doi.org/10.1186/s13040-021-00274-7

[15] Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., Asselbergs, F. W., Moons, K. G. M., and Debray, T. P. A., Missing Data Is Poorly Handled and Reported in Prediction Model Studies Using Machine Learning: A Literature Review. *Journal of Clinical Epidemiology*, Vol. 142, 2022, pp. 218–229. https://doi.org/10.1016/J.JCLINEPI.2021.11.023

[16] Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R., Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, Vol. 338, No. 7713, 2009, pp. 157–160. https://doi.org/10.1136/BMJ.B2393

[17] Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., and Siddique, Z., Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, Vol. 9, No. 3, 2021. https://doi.org/10.3390/TECHNOLOGIES9030052

[18] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O., A Survey on Missing Data in Machine Learning. *Journal of Big Data*, Vol. 8, No. 1, 2021. https://doi.org/10.1186/S40537-021-00516-9

[19] Lyngdoh, G. A., Zaki, M., Krishnan, N. M. A., and Das, S., Prediction of Concrete Strengths Enabled by Missing Data Imputation and Interpretable Machine Learning. *Cement and Concrete Composites*, Vol. 128, 2022, p. 104414. https://doi.org/10.1016/J.CEMCONCOMP.2022.104414

[20] Chen, N., Sun, Y., Wang, Z., and Peng, C., Improved LS-SVM Method for Flight Data Fitting of Civil Aircraft Flying at High Plateau. *Electronics (Switzerland)*, Vol. 11, No. 10, 2022. https://doi.org/10.3390/electronics11101558

[21] Ahmed, M., and Abdel-Aty, M., A Data Fusion Framework for Real-Time Risk Assessment on Freeways. *Transportation Research Part C: Emerging Technologies*, Vol. 26, 2013. https://doi.org/10.1016/j.trc.2012.09.002

[22] Koc, K., and Gurgun, A. P., Scenario-Based Automated Data Preprocessing to Predict Severity of Construction Accidents. *Automation in Construction*, Vol. 140, 2022. https://doi.org/10.1016/j.autcon.2022.104351

[23] Munian, Y., Martinez-Molina, A., Miserlis, D., Hernandez, H., and Alamaniotis, M., Intelligent System Utilizing HOG and CNN for Thermal Image-Based Detection of Wild Animals in Nocturnal Periods for Vehicle Safety. *Applied Artificial Intelligence*, Vol. 36, No. 1, 2022. https://doi.org/10.1080/08839514.2022.2031825

[24] Sibarani, J. N., Sirait, D. R., and Ramadhanti, S. S., Intrusion Detection Systems Pada Bot-IoT Dataset Menggunakan Algoritma Machine Learning. *Journal Masyarakat Informatika*, Vol. 14, No. 1, 2023. https://doi.org/10.14710/jmasif.14.1.49721

[25] Xue, L., Jiang, H., Zhao, Y., Wang, J., Wang, G., and Xiao, M., Fault Diagnosis of Wet Clutch Control System of Tractor Hydrostatic Power Split Continuously Variable Transmission. *Computers and Electronics in Agriculture*, Vol. 194, 2022. https://doi.org/10.1016/j.compag.2022.106778

[26] Qi, C., Fourie, A., Ma, G., and Tang, X., A Hybrid Method for Improved Stability Prediction in Construction Projects: A Case Study of Stope Hangingwall Stability. *Applied Soft Computing Journal*, Vol. 71, 2018. https://doi.org/10.1016/j.asoc.2018.07.035

[27] Liu, Z., and Li, S., A Sound Monitoring System for Prevention of Underground Pipeline Damage Caused by Construction. *Automation in Construction*, Vol. 113, 2020. https://doi.org/10.1016/j.autcon.2020.103125

[28] Antwi-Afari, M. F., Li, H., Seo, J. O., and Wong, A., Y. L. Automated Detection and Classification of Construction Workers' Loss of Balance Events Using Wearable Insole Pressure Sensors. *Automation in Construction*, Vol. 96, 2018. https://doi.org/10.1016/j.autcon.2018.09.010

[29] Hu, J., Huang, M. C., and Yu, X., Efficient Mapping of Crash Risk at Intersections with Connected Vehicle Data and Deep Learning Models. *Accident Analysis and Prevention*, Vol. 144, 2020. https://doi.org/10.1016/j.aap.2020.105665

[30] Liu, P., Xie, M., Bian, J., Li, H., and Song, L., A Hybrid Pso–Svm Model Based on Safety Risk Prediction for the Design Process in Metro Station Construction. *International Journal of Environmental Research and Public Health*, Vol. 17, No. 5, 2020. https://doi.org/10.3390/ijerph17051714

[31] Abbasianjahromi, H., and Aghakarimi, M., Safety Performance Prediction and Modification Strategies for Construction Projects via Machine Learning Techniques. *Engineering, Construction and Architectural Management*, Vol. 30, No. 3, 2023. https://doi.org/10.1108/ECAM-04-2021-0303

[32] Zhu, R., Hu, X., Hou, J., and Li, X., Application of Machine Learning Techniques for Predicting the Consequences of Construction Accidents in China. *Process Safety and Environmental Protection*, Vol. 145, 2021. https://doi.org/10.1016/j.psep.2020.08.006

[33] Alkaissy, M., Arashpour, M., Golafshani, E. M., Hosseini, M. R., Khanmohammadi, S., Bai, Y., and Feng, H., Enhancing Construction Safety: Machine Learning-Based Classification of Injury Types. *Safety Science*, Vol. 162, 2023. https://doi.org/10.1016/j.ssci.2023.106102

[34] Koc, K., Ekmekcioğlu, Ö., and Gurgun, A. P., Integrating Feature Engineering, Genetic Algorithm and Tree-Based Machine Learning Methods to Predict the Post-Accident Disability Status of Construction Workers. *Automation in Construction*, Vol. 131, 2021. https://doi.org/10.1016/j.autcon.2021.103896

[35] Geng, X., Wu, S., Zhang, Y., Sun, J., Cheng, H., Zhang, Z., and Pu, S., Developing Hybrid XGBoost Model Integrated with Entropy Weight and Bayesian Optimization for Predicting Tunnel Squeezing Intensity. *Natural Hazards*, Vol. 119, No. 1, 2023. https://doi.org/10.1007/s11069-023-06137-0

[36] Halabi, Y., Xu, H., Long, D., Chen, Y., Yu, Z., Alhaek, F., and Alhaddad, W., Causal Factors and Risk Assessment of Fall Accidents in the U.S. Construction Industry: A Comprehensive Data Analysis (2000–2020). *Safety Science*, Vol. 146, 2022. https://doi.org/10.1016/j.ssci.2021.105537

[37] Makki, A. A., and Mosly, I., Predicting the Safety Climate in Construction Sites of Saudi Arabia: A Bootstrapped Multiple Ordinal Logistic Regression Modeling Approach. *Applied Sciences (Switzerland)*, Vol. 11, No. 4, 2021. https://doi.org/10.3390/app11041474

[38] Baker, H., Hallowell, M. R., and Tixier, A. J. P., AI-Based Prediction of Independent Construction Safety Outcomes from Universal Attributes. *Automation in Construction*, Vol. 118, 2020. https://doi.org/10.1016/j.autcon.2020.103146