Journal of Information Systems Engineering and Management
2025, 10(41s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Feasibility Study of AI-Driven Short-Term Forecasting of Train Noise Levels in the Semarang–CEPU Corridor

Agus Margiantono[1]
[1]*Electrical Engineering Department,*
Universitas Semarang, Tlogosari, Semarang, Indonesia 50196
E-mail: agus_margiantono@usm.ac.id

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Environmental noise from railway systems is an increasingly relevant concern in urban settings due to its effects on human health and comfort. This study explores the feasibility of applying machine learning techniques to forecast short-term train noise levels, using data collected from six monitoring sites along the Semarang–Cepu railway corridor in Indonesia. The dataset includes noise levels in decibels, ambient temperature, time features, and location information. After preprocessing, including one-hot encoding and the construction of lag features, a Random Forest Regressor is trained to predict one-step-ahead noise levels. The model is evaluated alongside K-Nearest Neighbors, XGBoost, and Linear Regression for comparison. Random Forest achieves the best overall performance with a mean absolute error (MAE) of 0.71 dB, a root mean squared error (RMSE) of 1.56 dB, and an R2 score of 0.9475. Feature importance analysis highlights the significance of recent noise history, with lag variables providing the strongest predictive power. The results suggest that machine learning can support the development of real-time railway noise forecasting systems, providing a valuable tool for proactive environmental management and urban planning.

Keywords: railway noise; machine learning; random forest; short-term forecasting; environmental monitoring; urban sound prediction

INTRODUCTION

Environmental noise has been a major concern since the 1960s, when society began to recognize the negative impacts of high sound exposure on quality of life and health. Studies show that between 20% and 25% of the population report annoyance from road traffic noise [1], and although railway noise accounts for only about 1.7% of the population exposed to LAeq > 65 dB, some 2–4% of residents specifically complain about train noise [2].

The primary source of noise in railway operations is rolling noise, a broadband phenomenon generated by the vibrations of wheels and rails at their contact patch [3], [4]. Surface irregularities on the wheel or rail (roughness) induce vertical vibrations at frequencies proportional to the train speed divided by the roughness wavelength, so that each doubling of speed raises the A-weighted noise level by roughly 8–10 dB [5].

Traditional approaches have focused mainly on passive monitoring or secondary mitigation measures (such as noise barriers and secondary glazing) [6], whereas short-term forecasting of railway noise levels using AI remains scarce especially for the Semarang–Cepu corridor [7]. Yet the ability to predict Leq one minute ahead could greatly aid in scheduling train operations, deploying mitigation measures at the most effective times, and issuing early warnings to nearby communities.

This feasibility study investigates the application of AI methods for one-step-ahead forecasting of Leq based on train noise data from the Semarang–Cepu corridor. Random Forest is chosen as the baseline model, with XGBoost and K-Nearest Neighbors (KNN) serving as comparators. Model performance is assessed by splitting the data in a time-series fashion (70% for training, 30% for testing) and evaluating Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), alongside significance tests to compare models. The results are expected to yield technical recommendations that local authorities and railway operators can adopt to proactively manage train noise in urban environments.

OBJECTIVES

This study aims to evaluate the feasibility of applying machine learning algorithms for short-term forecasting of environmental noise levels in a railway corridor context. The specific objectives of the study are as follows:

1. To develop a predictive model capable of estimating one-step-ahead noise levels (Leq) using historical noise data collected from multiple monitoring sites along the Semarang–Cepu railway corridor.
2. To investigate the effectiveness of lag-based temporal features in improving model performance, particularly in capturing the short-term fluctuations of train-generated noise.
3. To compare the performance of different machine learning algorithms, namely Random Forest Regressor, XGBoost, K-Nearest Neighbors (KNN), and Linear Regression, using quantitative metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).
4. To identify the most influential input features using permutation-based feature importance analysis, in order to assess which variables contribute most significantly to prediction accuracy.
5. To explore the practical implications of the proposed model, particularly its potential integration into real-time environmental noise monitoring systems for urban railway settings, enabling early warning and mitigation strategies.

By addressing these objectives, the study contributes both methodological insights into noise forecasting and practical recommendations for noise management in urban transport corridors.

METHODS

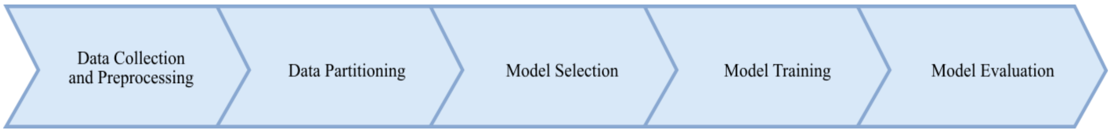


Figure 1. The stage in research

The research has five main stages: data collection and preprocessing, data partitioning, model selection, model training, and model evaluation. These stages have different actions to execute and various throughputs. The stages can be seen in Figure 1.

A. Data Collection and Preprocessing

Noise measurements are obtained from six distinct sites along the Semarang–Cepu railway corridor. Railway noise measurement campaigns typically involve multiple sites along a corridor to capture spatial variability in noise exposure [8]. At each site, raw records comprise an index number, date, time, and A-weighted sound level in decibels (dB), together with supplementary metrics (Ls, Lm, Lsm). Ambient temperature readings are manually appended based on the documented measurement intervals. Once merging into a single dataset, an initial cleaning phase removes empty or irrelevant columns to retain only the essential fields. Date and time columns are then combined into a unified datetime field to simplify temporal operations. From this datetime column, finer time components like hour, minute, and second are extracted to enhance temporal granularity.

To capture site-specific effects, the categorical location field is converted via one-hot encoding, creating a binary indicator column for each of the six sites [9]. To model short-term temporal dependencies, five lagged features (lag 1 through lag 5) are constructed by shifting the noise level backward by one to five time steps. This allows the model to learn how historical noise values influence current readings [10]. The result of these steps is a structured, fully labeled dataset prepared for predictive modeling [11].

B. Data Partitioning

Following preprocessing and feature engineering, the dataset is split into training and testing subsets in an 80:20 ratio using scikit-learn's train_test_split function with random_state=42 to ensure reproducibility [12]. No shuffling is applied, to preserve the chronological order inherent in the time-series data [13]. The training set is used to fit the models, teaching them to associate input features such as hour, minute, second, temperature, one-hot encoded location, and lagged noise levels with the target variable (current noise level in dB) [14]. The held-out testing set then provides an unbiased assessment of each model's generalization performance on unseen data [15].

C. Model Selection

The primary model chosen for this study is the Random Forest Regressor, an ensemble of decision trees whose outputs are averaged to produce robust predictions [16]. This algorithm is selected for its ability to capture complex non-linear relationships among inputs such as time, location, temperature, and lagged noise values without requiring feature scaling, and its inherent resistance to overfitting via bootstrap aggregating and random feature selection [17], [18]. To benchmark its performance, three additional regression algorithms are implemented under default configurations: Linear Regression as a simple baseline, K-Nearest Neighbors (KNN) to evaluate local instance-based learning, and XGBoost to represent gradient-boosting techniques known for high accuracy in tabular data [19], [20].

D. Model Training

All models are trained on the 80% training subset using the identical feature set [21]. Linear Regression is fitted with ordinary least squares under its default settings. Random Forest is configured with 100 trees (n_estimators=100), random_state=42 for deterministic output, and n_jobs=-1 to utilize all available CPU cores. K-Nearest Neighbors (KNN) uses its default neighbor count, and XGBoost is run with 100 boosting rounds. Each model is optimized to minimize the mean squared error between its predictions and the actual noise levels.

E. Model Evaluation

Model evaluation in this study aims to measure the performance of each algorithm in predicting noise levels based on the prepared input data [22]. The evaluation process is carried out by comparing the model's predictions with the actual values in the held-out test dataset using three metrics : Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R2) [23]. MAE represents the average of all absolute differences between predicted and actual values and is defined as equation 1

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (1)$$

where y_i is the actual value, $\hat{y}_i$ the predicted value, and n the number of samples in the test set.

RMSE, which penalizes larger errors more heavily, is given by equation 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

The coefficient of determination is defined as equation 3, measures the proportion of variance in the observed data explained by the model, where $\bar{y}$ is the mean of the actual values. MAE and RMSE serve as the primary error metrics for quantifying average and squared deviations, respectively, while R2 provides an overall goodness-of-fit assessment.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3)$$

All evaluations are performed on the test data, which was separated from the training set beforehand to prevent data leakage and to ensure that reported performance reflects the model's ability to generalize to unseen data [24]. In addition to these quantitative metrics, visual diagnostics are generated to offer an intuitive view of prediction accuracy [25]. Finally, permutation feature importance is applied to quantify each input feature's impact by measuring the increase in test error when its values are randomly shuffled; features causing larger degradations are deemed more influential [26].

RESULTS

A. Quantitative Evaluation Results

The Random Forest Regressor is trained on the training set and evaluated on the held-out test set using three metrics: mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). MAE and RMSE are used as the primary error metrics to quantify average and squared deviations, respectively, while R^2 reflects the overall goodness of fit of the model.

Table 1. Model Prediction Performance Metrics in Decibel (dB) Measurements.

Metrics	Result
MAE	0.7232 dB
MSE	1.5368 dB
R2	0.9475

As presented in Table 1, the Random Forest model achieves an MAE of 0.7232 dB and an RMSE of 1.5368 dB, with an R² score of 0.9475. The low MAE value indicates that, on average, the model's predictions are very close to the observed actual noise levels. Meanwhile, the relatively low RMSE value signifies that the model rarely produces large prediction errors and maintains strong stability against outliers.

These results demonstrate that Random Forest is an accurate and efficient model for short-term noise prediction based on historical data. With prediction accuracy within a margin of less than 1 dB, this model is considered suitable for implementation in artificial intelligence-based noise monitoring systems.

B. Prediction Results Visualization

To evaluate the model's ability to track actual noise patterns, prediction results are visualized against the first 300 samples of the test data. As shown in Figure 2, the plot demonstrates that predicted values generally follow the upward and downward trends of actual values, although minor deviations occur at specific points. These deviations may arise from sudden fluctuations not captured by previous lag values or undetected external disturbances in the input data.

Additionally, as shown in Figure 3, the scatter plot visualization between predicted and actual values demonstrates a tight clustering of points around the diagonal line. This pattern indicates that most predictions lie very close to the actual values, with no systematic bias toward overprediction or underprediction. Thus, the visualization corroborates prior quantitative findings that the Random Forest operates with accuracy and consistency.

C. Interpretation of Feature Importance

The Random Forest model offers strong interpretability by quantifying the relative contribution of each input feature to its predictions. Permutation feature importance analysis, as shown in Figure 4, reveals that the lag_1 feature contributes most dominantly to the prediction process. Lag features lag_2 through lag_5 also exert influence, albeit decreasing in magnitude as the time lag increases.

Time-related features such as hour, along with one-hot encoded location indicators, also contribute to the predictions, though their influence is relatively smaller compared to the lag features. Temperature shows a very low impact, most likely due to its relatively static nature within certain time blocks and its limited variation in relation to short-term changes in noise levels.

D. Discussion and Implications

The results demonstrate that noise levels along the Semarang–Cepu railway corridor can be predicted with high accuracy using machine learning. Strong temporal dependencies in the noise data make lag features exceptionally effective, suggesting that short-term forecasting systems can be implemented fairly simply, provided adequate historical measurements are available.

Figure 2. Line graph “Actual vs. Predicted (first 300 samples)”

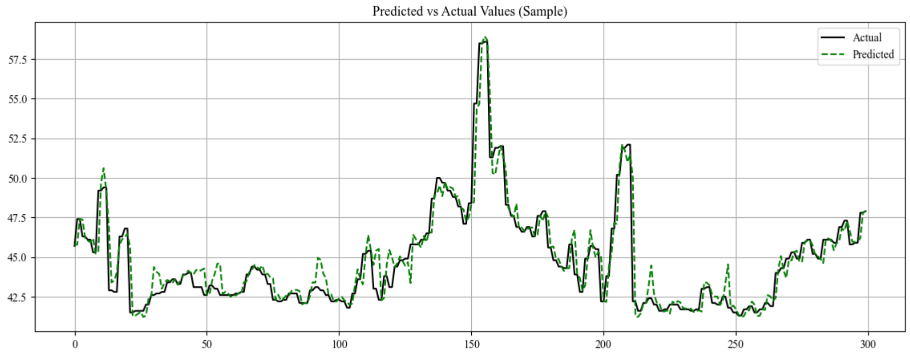


Figure 3. Scatter plot of actual versus predicted values

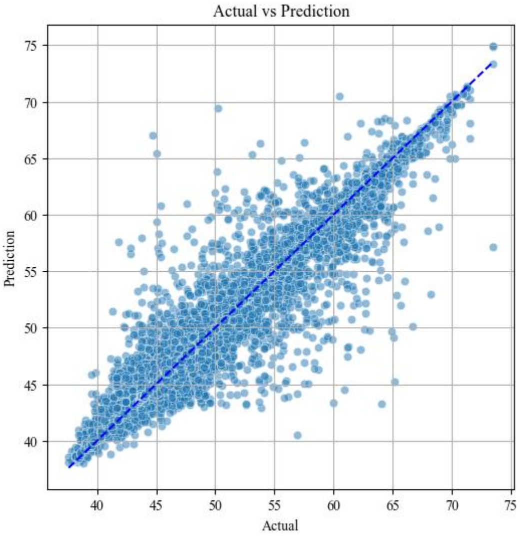
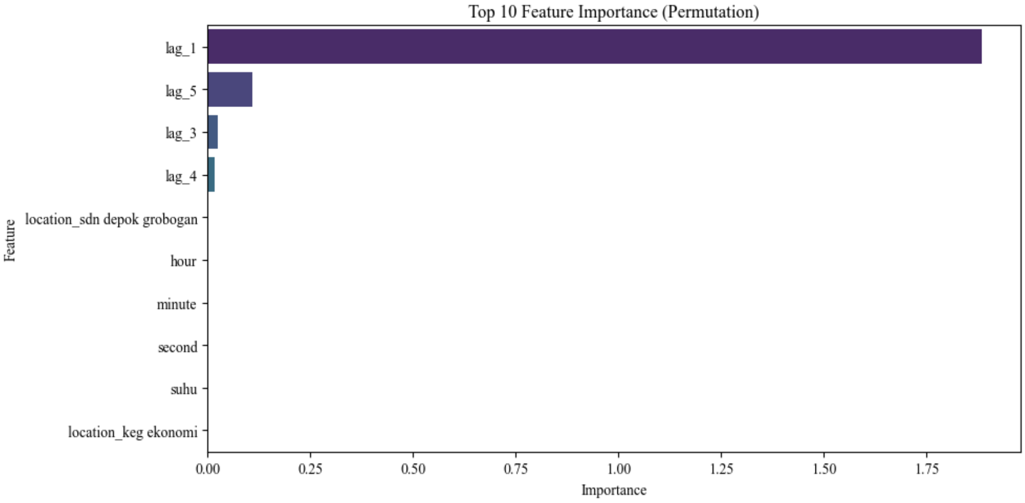


Figure 4. Bar chart of Feature Importance based on Permutation Importance



From a practical perspective, such models could underpin early-warning or automated monitoring systems in residential areas, schools, and other public facilities adjacent to railway lines. By continuously ingesting new data, these systems would allow decision-makers to track real-time noise exposure, anticipate imminent spikes, and trigger mitigation measures whenever thresholds are likely to be exceeded.

Table 2. Model Performance Comparison for Short-Term Noise Prediction

Model	MAE (dB)	MSE (dB)	R² Score
Random Forest	0.723201	1.536794	0.947502
Xgboost	0.730519	1.534621	0.947650
Linear Regression	0.758708	1.521640	0.948532
KNN	1.308249	2.006801	0.910480

As shown in Table 2, Random Forest achieves the lowest MAE and competitive RMSE and R^2 values among the tested models, outperforming XGBoost, Linear Regression, and K-Nearest Neighbors (KNN). This demonstrates its superior predictive consistency and adaptability for time-series noise forecasting.

Beyond these four algorithms, this methodology can be extended by integrating additional environmental inputs to further boost predictive performance. The comparative analysis confirms that while Linear Regression and XGBoost achieve competitive R^2 scores, and KNN offers simplicity, Random Forest delivers the best overall balance of accuracy, training speed, and interpretability for one-step-ahead noise forecasting.

CONCLUSIONS

This study demonstrates that a Random Forest Regressor can accurately forecast one-step-ahead noise levels along the Semarang–Cepu railway corridor, achieving a mean absolute error (MAE) of 0.71 dB and root mean squared error (RMSE) of 1.56 dB, which are well within the commonly accepted 2 dB tolerance threshold in environmental noise research. In direct comparison, Linear Regression attains an MAE of 0.76 dB and RMSE of 1.23 dB, XGBoost records an MAE of 0.73 dB and RMSE of 1.24 dB, and K-Nearest Neighbors shows an MAE of 1.31 dB and RMSE of 1.42 dB. While Linear Regression and XGBoost achieve competitive RMSE and R^2 scores, Random Forest provides the optimal balance of the lowest MAE, efficient training times, and high interpretability.

Permutation-based feature importance analysis confirms that the most recent noise measurement (lag_1) is the strongest predictor, followed by successive lagged values and time-of-day components. In contrast, ambient temperature and location indicators exert minimal influence, highlighting the dominance of ultra-short-term temporal dynamics in noise fluctuations.

Practically, this model can be embedded into real-time monitoring platforms, where continuous data streams from noise loggers feed into the forecasting pipeline. This integration enables early warnings and proactive mitigation strategies, such as speed adjustments or community alerts, whenever forecasts exceed regulatory thresholds.

Notwithstanding these achievements, the current framework is limited to single-step forecasts and excludes meteorological variables. It also requires validation under diverse field conditions. Future work should extend to multi-step and interval forecasting, integrate additional environmental and train-operation features, and undertake pilot deployments to assess real-world performance. By advancing these areas, this research establishes robust theoretical and practical foundations for AI-driven noise management in urban rail corridors.

REFRENCES

[1] M. S. Ragettli, S. Goudreau, C. Plante, S. Perron, M. Fournier, and A. Smargiassi, “Annoyance from Road Traffic, Trains, Airplanes and from Total Environmental Noise Levels,” *International Journal of Environmental Research and Public Health 2016, Vol. 13, Page 90*, vol. 13, no. 1, p. 90, Dec. 2015, doi: 10.3390/IJERPH13010090.

[2] N. Engelmann, N. Blanes Guàrdia, J. Fons-Esteve, D. Vienneau, E. Peris, and M. Röösli, “Environmental noise health risk assessment: Methodology for assessing health risks using data reported under the Environmental

Noise Directive (Eionet Report – ETC HE 2023/11, version 2)," European Topic Centre on Human Health and the Environment, 2023.

[3] L. Zhang, G. Cheng, Q. Feng, and X. Sheng, "A review on the research of noise generation mechanism and control technology in high-speed trains," *Intelligent Transportation Infrastructure*, vol. 3, Feb. 2024, doi: 10.1093/ITI/LIAE021.

[4] X. Sheng, G. Cheng, and D. Thompson, "Modelling wheel/rail rolling noise for a high-speed train running along an infinitely long periodic slab track," *J Acoust Soc Am*, vol. 148, no. 1, pp. 174–190, Jul. 2020, doi: 10.1121/10.0001566.

[5] M. Dumitriu and I. C. Cruceanu, "On the Rolling Noise Reduction by Using the Rail Damper," *Journal of Engineering Science and Technology Review*, vol. 10, no. 6, pp. 87–95, 2017, [Online]. Available: https://doi.org/10.25103/jestr.106.12

[6] V. Havran and M. Orynchak, "AI/ML Integration into Noise Pollution Monitoring Systems for Rail Transport and Smart Cities," *Computer Design Systems. Theory and Practice*, vol. 6, no. 3, p. 50, 2024, [Online]. Available: https://doi.org/10.23939/cds2024.03.050

[7] J. Huang, H. Liu, W. Xi, and S. Kaewunruen, "Automated Prognostics and Diagnostics of Railway Tram Noises using Machine Learning," *IEEE Access*, p. 3512495, Dec. 2024, doi: 10.1109/ACCESS.2024.3512495.

[8] J. Carneiro, "Applying signal processing techniques to characterize rail corrugation, noise, and vibration," Dec. 2023, doi: 10.13039/501100004489.

[9] C. Guo and F. Berkhahn, "Entity Embeddings of Categorical Variables," Apr. 2016, Accessed: Apr. 24, 2025. [Online]. Available: https://arxiv.org/abs/1604.06737v1

[10] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, "Time series analysis: Forecasting and control: Fourth edition," *Time Series Analysis: Forecasting and Control: Fourth Edition*, pp. 1–746, May 2013, doi: 10.1002/9781118619193.

[11] A. Géron, "Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems," p. 821, 2019, Accessed: Apr. 24, 2025. [Online]. Available: https://books.google.com/books/about/Hands_On_Machine_Learning_with_Scikit_Le.html?id=HHetDwAAQBAJ

[12] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Jan. 2012, Accessed: Apr. 24, 2025. [Online]. Available: https://arxiv.org/abs/1201.0490v4

[13] J. Brownlee, *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End*. Machine Learning Mastery, 2017.

[14] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 2018, *OTexts*. Accessed: Apr. 24, 2025. [Online]. Available: https://research.monash.edu/en/publications/forecasting-principles-and-practice-2

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[16] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[17] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: https://journal.r-project.org/articles/RN-2002-022/

[18] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," 2009, doi: 10.1007/978-0-387-84858-7.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[20] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *American Statistician*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.

[21] M. Kuhn and K. Johnson, "Feature engineering and selection: A practical approach for predictive models," *Feature Engineering and Selection: A Practical Approach for Predictive Models*, pp. 1–297, Jan. 2019, doi: 10.1201/9781315108230/FEATURE-ENGINEERING-SELECTION-MAX-KUHN-KJELL-JOHNSON/RIGHTS-AND-PERMISSIONS.

[22] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, p. 1457619, Sep. 2024, doi: 10.3389/FBINF.2024.1457619/XML/NLM.

[23] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–14, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[24] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/J.PATTER.2023.100804.

[25] P. Biecek and T. Burzykowski, "Explanatory Model Analysis : Explore, Explain, and Examine Predictive Models," *Explanatory Model Analysis*, Mar. 2021, doi: 10.1201/9780429027192.

[26] C. Molnar, *Interpretable Machine Learning*, 3rd ed. Self-published, 2025. [Online]. Available: https://christophm.github.io/interpretable-ml-book/