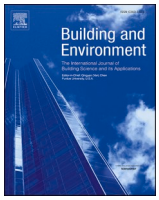
Building and Environment 284 (2025) 113493

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv



Feature importance in building machine learning: Beyond model-dependent interpretations

ARTICLE INFO

Keywords:
Building performance
Optimal indoor air quality
Feature importance
Machine learning interpretability
Statistical validation

Letter to the editor

The recent paper by Godasiaei et al. in Building and Environment, "Integrating experimental analysis and machine learning for enhancing energy efficiency and indoor air quality in educational buildings," makes a valuable contribution to the management of energy efficiency and indoor air quality [1]. Their study addressed the challenge of balancing energy consumption with indoor air quality (IAQ) through experimental analysis integrated with advanced machine learning (ML) techniques. They evaluated the potential of ML models, including Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN), for enhancing energy efficiency and indoor air quality. However, the reliance on complex machine learning (ML) models and the interpretation of feature influence using SHAP warrants further discussion.

Godasiaei et al. employed RNN, LSTM, GRU, and CNN models using a dataset of over 35,000 records to predict IAQ and HVAC energy efficiency optimization in real-time. They reported robust performance, with the GRU model achieving an R2 value of 0.973 and a Mean Absolute Error (MAE) of 0.291. The LSTM model also showed strong performance with an R2 of 0.925 and an MAE of 0.309. Overall, the predictive models achieved over 92 % accuracy, enabling precise real-time HVAC control. Beyond evaluating model performance, a key aspect of their work involved interpretability analysis using SHAP values, revealing influential parameters such as CO2 levels, outdoor air temperature, and HCHO. While this approach selects the best-performing model based on prediction accuracy, it implicitly assumes that the interpretability of this chosen model is also superior or more reliable. Nevertheless, high predictive accuracy alone doesn't guarantee reliable feature importance, a limitation broadly acknowledged in over 300 studies [2,3]. Further details are available in the supplementary material.

The challenge of definitively determining feature importance in complex machine learning models is significant. While recurrent models such as RNN, LSTM, and GRU are powerful predictive tools, their intricate architectures often create learned representations closely tied to specific training data. This means high prediction accuracy doesn't automatically validate the reliability of derived importances. It is crucial to distinguish between target prediction accuracy and feature importance accuracy (or reliability); the former does not inherently guarantee the latter. Specifically for the GRU model, its recurrent nature and internal gating mechanisms capture complex temporal dependencies. However, these very mechanisms introduce significant biases into its feature importance assessments. The primary bias in GRUs stems from the intricate propagation of information across sequential time steps and through multiple, non-linear internal transformations [4,5]. This interwoven flow means individual feature influence is distributed, ambiguous, and highly context-dependent, making it exceedingly difficult to pinpoint precise contributions. Moreover, issues like the vanishing gradient problem introduce a critical temporal bias, implicitly downplaying the importance of significant features from earlier time steps as their influence diminishes during training. This algorithmic predisposition to distort feature contributions poses substantial methodological hurdles for accurate interpretation.

This concern extends to interpretation methods like SHAP. While these values are intended to elucidate feature influence, they are intrinsically tied to the model they interpret, potentially reflecting or amplifying its inherent characteristics. The function 'explain = SHAP (model)' underscores this dependency. Consequently, when such methods are applied to the unreliable GRU model, their explanations fundamentally reflect the GRU's abstract temporal logic and internal states, rather than clear, direct causal relationships in the system. As they rely on the model's output for explanations, their vulnerability to model nuances demands careful interpretation. Despite SHAP's widespread adoption, its fundamental dependency means it acts as a mirror, reflecting and potentially amplifying the biases of the model it explains. Therefore, relying solely on these values derived from complex ML models to determine true feature importance is problematic, as current techniques cannot fully eliminate inherent bias [6,7].

Validating feature influence is difficult due to the lack of ground truth, as different models yield varying insights based on their methodologies. This issue worsens when interpretation relies mainly on a single 'best-performing' model selected for prediction accuracy, potentially overlooking alternative perspectives. In Godasiaei et al.'s study,

DOI of original article: https://doi.org/10.1016/j.buildenv.2025.113494.

https://doi.org/10.1016/j.buildenv.2025.113493
Received 29 May 2025; Accepted 27 July 2025
Available online 28 July 2025

the multivariate nature of environmental sensor data and variable interdependencies further obscure true relationships. Thus, model-dependent interpretations hinder confident identification of actual determinants of energy efficiency and IAQ, as feature influences may reflect model-specific biases and complex sensor interactions. These factors show interpretability measures are sensitive to model configurations, affecting reliability.

To effectively transcend these limitations, a robust analytical paradigm is imperative, one that intrinsically links data characteristics with the statistical fabric of variable relationships and stringent validation. True mastery in modeling and interpretation demands an acute grasp of the core environmental and building processes driving energy efficiency and IAQ. It is paramount to unearth complex associations, particularly through non-parametric methods. Moreover, establishing the statistical significance of findings remains crucial to avert specious conclusions. Instead of solely relying on complex ML models and their embedded interpretability tools like SHAP to identify key drivers and understand model behavior, we champion a synergistic framework. This framework marries the predictive might of machine learning with impartial and rigorous statistical methodologies. Such methods include Spearman's rho and Kendall's tau, exceptionally suited for dissecting monotonic relationships [8]. For plumbing the depths of more intricate dependencies, including non-monotonic interactions, non-parametric avenues like Mutual Information and Total Correlation offer profound insights [9,10]. Elevating these foundational statistical principles, in conjunction with ML and domain expertise, will profoundly enhance the veracity and trustworthiness of feature influence and model behavior assessments within building and environmental engineering contexts.

In conclusion, Godasiaei et al.'s study provides valuable models and identifies features relevant to energy efficiency and IAQ optimization through ML and SHAP analysis. While insightful, interpreting these features as definitive drivers in complex environmental systems requires careful consideration of methodological limitations and validation challenges. To better understand complex building performance outcomes like energy efficiency and IAQ, we must move beyond model-dependent interpretations. A more robust approach integrates machine learning's predictive power with complementary, rigorous statistical methods, offering a more reliable foundation for interpreting feature influence and system behavior.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Souichi Oka: Writing – original draft. **Takuma Yamazaki:** Investigation. **Yoshiyasu Takefuji:** Supervision, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.buildenv.2025.113493.

Data availability

No data was used for the research described in the article.

References

[1] S.H. Godasiaei, O.A. Ejohwomu, H. Zhong, D. Booker, Integrating experimental analysis and machine learning for enhancing energy efficiency and indoor air quality in educational buildings, Build Environ. 276 (2025) 112874, https://doi.org/10.1016/j.buildenv.2025.112874.
[2] Z.C. Lipton, The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery, ACM Queue 16 (3) (2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[3] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 177, https://doi.org/10.48550/arXiv.1801.01489.
[4] W. Tang, Q. Zhang, Y. Chen, X. Liu, H. Wang, W. Huang, An intelligent airflow perception model for metal mines based on CNN-LSTM architecture, Process Saf. Environ. Prot. 187 (2024) 1234–1247, https://doi.org/10.1016/j.psep.2024.05.044.
[5] K. Thanjavur, D.T. Hristopulos, A. Babul, K.M. Yi, N. Virji-Babul, Deep Learning recurrent neural network for concussion classification in adolescents using raw electroencephalography signals: toward a minimal number of sensors, Front. Hum. Neurosci. 15 (2021) 734501, https://doi.org/10.3389/fnhum.2021.734501.
[6] B. Bilodeau, N. Jaques, P.W. Koh, B. Kim, Impossibility theorems for feature attribution, in: Proc. Natl. Acad. Sci. 121, e2304406120, 2024, https://doi.org/10.1073/pnas.2304406120.
[7] X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int. J. Approx Reason 171 (2024) 109112, https://doi.org/10.1016/j.ijar.2023.109112.
[8] K. Okoye, S. Hosseini, Correlation tests in R: pearson Cor, Kendall's tau, and Spearman's rho, in: K. Okoye, S. Hosseini (Eds.), R Programming: Statistical Data Analysis in Research, Springer Nature, 2024, pp. 247–277, https://doi.org/10.1007/978-981-97-3385-9_12.
[9] J.D. Gibson, Entropy and mutual information, in: information Theoretic Principles for Agent learning, Synth. Lect. Eng. Sci. Technol. (2025), https://doi.org/10.1007/978-3-031-65388-9_2.
[10] S. Tserkis, S.M. Assad, P.K. Lam, P. Narang, Quantifying total correlations in quantum systems through the Pearson correlation coefficient, Phys. Lett. A 543 (2025) 130432, https://doi.org/10.1016/j.physleta.2025.130432.

Souichi Oka[a,*], Takuma Yamazaki[a], Yoshiyasu Takefuji[b]
[a] *Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan*
[b] *Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan*

[*] Corresponding author.
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), tyamazaki@sciencepark.co.jp (T. Yamazaki), takefuji@keio.jp (Y. Takefuji).