Received: 23 April 2026 | Revised: 21 August 2026 | Accepted: 25 August 2026

DOI: 10.1002/cend.70026

FULL PAPER

Ernst & Sohn
A Wiley Brand

A leave-top-feature-out framework for validating feature importance reliability in concrete compressive strength modeling

Yoshiyasu Takefuji

AI Laboratory, SciencePark Corporation, Kanagawa, Japan

Correspondence
Yoshiyasu Takefuji, AI Laboratory, SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan.
Email: takefuji@keio.jp

Abstract

Civil Engineering Design has seen a surge in AI-driven analyses of concrete compressive strength, yet this proliferation has obscured a methodological crisis in supervised machine learning for sustainable materials research. Existing studies conflate two accuracy dimensions: target prediction accuracy, validatable against known labels, and feature importance accuracy, which lacks validation benchmarks. Establishing true variable associations requires two criteria from causal inference: consistency (stable feature rankings across perturbations) and dose–response relationships (monotonic changes in model output upon feature removal), criteria prior studies have failed to meet. We introduce a leave-top-feature-out validation framework applied to a concrete compressive strength dataset, wherein top-ranked features are systematically removed and perturbations quantified. Evaluation of supervised models (Random Forest, XGBoost), unsupervised methods (Feature Agglomeration, Highly Variable Gene Selection), and nonparametric statistics (Spearman correlation) reveals that supervised approaches produce volatile rankings susceptible to label-driven biases, while unsupervised methods yield more consistent hierarchies, with SHapley Additive exPlanations (SHAP) explanations shown to amplify model biases. Our framework establishes the first theoretically grounded benchmark for distinguishing genuine from spurious associations in concrete strength prediction.

KEYWORDS

construction analytics, feature importance stability, prediction-causation disconnect, supervised learning limitations, variable relationship consistency

1 | INTRODUCTION

Civil Engineering Design has witnessed a remarkable surge in research addressing advanced association analytics, with 1 article on feature importance, 3 articles on feature selection (2 in 2025), 1 article on SHapley Additive exPlanations (SHAP), 18 articles on relationships between variables, 35 articles on correlation analysis, and 21 articles on concrete strength studies. This proliferation reflects the engineering informatics research community's intensifying interest in uncovering genuine relationships or true associations among variables.

© 2026 John Wiley & Sons Ltd.
All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

Despite this growing emphasis, the theoretical foundations and core assumptions underlying supervised machine learning remain inadequately understood, resulting in widespread misinterpretation of feature importance measures. A central theoretical concern is that supervised models exhibit dual accuracy dimensions: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be reliably validated against ground truth labels, feature importance assessment lacks corresponding validation benchmarks, signifying the theoretical absence of ground truth in feature assessment and creating a fundamental methodological gap with far-reaching consequences for scientific inference.

This absence of ground truth for feature importance calculations has led to potentially erroneous interpretations across the field. Many existing studies have fallen short due to insufficient validation protocols and fundamental misconceptions regarding the theoretical applicability of supervised models for feature assessment. To accurately assess true associations, two theoretically recommended components must be examined: consistency and dose–response relationships. This paper focuses on concrete strength analysis using publicly available data, theoretically discusses this pitfall, and empirically demonstrates distortions using cross-validation and consistency analysis of ordered feature sets, whereas existing studies failed by examining non-ordered sets of features, a distinction with profound theoretical implications for the validity of reported associations.

A prevalent and theoretically consequential misconception is that models with high predictive accuracy necessarily produce reliable feature importance rankings, a conflation that merits rigorous theoretical scrutiny. This paper offers three significant theoretical contributions. First, it develops theoretical and empirical arguments challenging the methodological validity of employing supervised models for feature importance assessment in the absence of established ground truth, thereby advancing the epistemological understanding of what supervised learning can and cannot reliably infer. Second, it illustrates these theoretical concerns using a public dataset by systematically comparing supervised approaches against more transparent statistical and unsupervised alternatives, providing empirical evidence that high predictive performance does not guarantee interpretive fidelity. Third, it proposes a novel validation framework called "leave-top-feature-out" (LTFO), grounded in the theoretical principles of consistency and dose–response criteria, to evaluate the stability and credibility of identified associations in a manner that is both reproducible and scientifically defensible. These contributions illuminate interpretability pitfalls that potentially undermine research integrity and practical applications while providing a theoretically coherent and systematic approach for more reliable assessment of true variable relationships in structural research.

While researchers excel in their specialized domains, they often lack awareness of algorithm-induced errors and the fundamental assumptions surrounding ground truth availability. Critically, prior studies have not adequately validated true variable associations. To accurately determine true associations, two key elements must be rigorously examined: feature ranking consistency and dose–response relationships between variables[1–11] while existing studies have fallen short. Many researchers mistakenly assume that high target prediction accuracy guarantees reliable feature-importance rankings, while overlooking that ground-truth values for feature importance are typically unavailable for validation. As a result, conclusions in existing studies have frequently been drawn without properly validating true associations based on consistency and dose–response evidence.

This paper identifies fundamental flaws in prevailing supervised models used for feature-importance assessment and empirically demonstrates the instability of feature-importance ranking orders using a publicly available dataset on concrete compressive strength. We introduce a novel LTFO approach that integrates both consistency and dose–response criteria by imposing a strong perturbation on the top-ranked feature and observing resultant shifts in the ranking order. We also recommend complementary methodologies to enhance the reliability of feature-importance inference.

Supervised learning models operate along two separate accuracy axes: their ability to predict target outcomes and their capacity to identify important features. The first dimension can be assessed using known labels, but the second lacks any definitive benchmark for validation—representing a critical methodological gap. Additionally, feature-importance measures capture predictive contribution rather than genuine causal influence. This fundamental difference accounts for why various supervised approaches produce inconsistent, algorithm-specific importance rankings. Without ground truth and due to susceptibility to label-induced biases, strong predictive performance does not ensure reliable importance measurements,[12–20] making feature-importance estimates from supervised models prone to misinterpretation.[21–30] Although these challenges have been thoroughly documented—across more than 300 peer-reviewed publications—many practitioners continue to be unaware of these limitations.

To probe true associations more rigorously, we propose a LTFO procedure. First, select the top n features from the full dataset (Set 1). Next, remove the highest-ranked feature from the full set to create a reduced dataset, then reselect the top $n-1$ features from this reduced dataset (Set 2). Finally, compare the ranking orders between set 1 and Set 2. If the model faithfully captures consistency and dose–response relationships, removing the top-ranked feature should induce substantial, interpretable shifts in the remaining rankings.[1–10] This procedure offers a more robust basis for assessing variable associations in structures research.

We further examine the consistency of variable associations across diverse analytical frameworks, including supervised machine learning models (Random Forest [RF], XGBoost [XGB]), unsupervised dimensionality-reduction techniques (Feature Agglomeration [FA], Highly Variable Gene Selection [HVGS]), and traditional statistics (Spearman's rank correlation). By assessing concordance across these methodologically distinct approaches, we establish a framework for validating associations that persist regardless of analytical paradigm. Our results show that high prediction accuracy does not ensure reliable feature importances and that importances derived from supervised models can be systematically skewed.

This study confronts the rapid growth of AI-based association analytics in construction by showing that high target prediction

accuracy does not validate feature-importance reliability or true associations. Using the concrete compressive strength dataset, we implement a LTFO protocol to jointly test consistency and dose–response relationships,[1–11] and we triangulate across supervised models (RF, XGB), unsupervised dimensionality reduction (FA, HVGS), and traditional statistics (Spearman's rank correlation). We find that supervised models lack ground truth for feature-importance accuracy and are vulnerable to label-driven errors,[12–20] producing unstable, model-dependent rankings that reflect predictive contribution rather than causal relevance, a limitation long documented but still underappreciated.[21–30]

Critically, the function of explain = SHAP(model) inherits the supervised model's importance structure; when that structure is biased or inconsistent, SHAP can faithfully propagate and even amplify those biases, leading to misleading explanations despite excellent prediction.[20,31–39] In contrast, unsupervised methods, especially FA, yield perfectly consistent rankings and sustain reasonable performance under LTFO perturbations, offering a more stable signal of association. Together, these results (i) empirically demonstrate the disconnect between prediction accuracy and importance reliability, (ii) document systematic skew in supervised-model and SHAP-derived importances, and (iii) provide a practical validation framework grounded in consistency, dose–response, perturbation testing, and cross-method concordance. We recommend pairing supervised prediction with complementary unsupervised and non–target-prediction analyses, applying LTFO to probe true associations, and interpreting SHAP as contingent on the underlying model rather than as ground truth, thereby mitigating algorithm-induced errors and avoiding misleading conclusions.

This paper establishes a critical distinction between target prediction accuracy and feature importance assessment, demonstrating that these represent fundamentally different analytical objectives. Exclusively relying on supervised models for determining feature importances, regardless of their predictive performance, is methodologically unsound. For credible feature importance assessment, researchers are advised to implement rigorous validation protocols that examine the derived importance rankings through multiple lenses. Specifically, true associations should demonstrate both consistency across different analytical approaches and clear dose–response relationships between features and outcomes, characteristics that cannot be assumed merely from a model's predictive success. This validation framework represents an essential safeguard against spurious associations that might otherwise mislead engineering decisions or research directions. We must acknowledge that there is no single algorithm to accurately calculate true associations.

The LTFO framework has been successfully validated across multiple domains and disciplines using diverse datasets, consistently demonstrating both its robustness and the limitations of SHAP-based explanations.[40–46]

2 | METHODS

This paper examines feature quality through two complementary lenses: cross-validation to assess target prediction accuracy and consistency of selected feature subsets, and a LTFO approach to evaluate the stability of ordered feature rankings across two derived feature sets.

This study leverages a publicly available dataset on concrete compressive strength sourced from OpenML (ID: 46917).[47] The dataset was selected for its completeness—containing no missing values—to ensure reproducibility and methodological transparency. It comprises 1030 samples described by eight physicochemical features (cement, slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age in days), with concrete compressive strength (MPa) as the continuous prediction target.

We employ a comprehensive analytical framework comprising five distinct feature importance assessment methods: two supervised learning algorithms—RF and XGB —two unsupervised dimensionality reduction techniques—FA and HVGS—and one traditional statistical approach, Spearman's rank correlation. To facilitate equitable comparison across methods, all algorithms are run with default hyperparameter settings, and no scaling or transformation is applied to the input features.

The experimental design implements a novel LTFO methodology to evaluate the stability of feature importance rankings. In the first step, each method ranks all available features and the top six are retained to form the initial reduced dataset, denoted Set 1. In the second step, the single highest-ranked feature is removed from the full dataset—not from Set 1—and the feature selection process is repeated on this reduced pool. The top five features identified in this step form the second reduced dataset, denoted Set 2. Both sets therefore contain six and five features, respectively, while being derived from inputs of differing cardinality. The central analytical question is whether the rankings within Set 2 remain consistent with those in Set 1: stable methods should reproduce largely the same ordered feature subset even after the leading feature has been excluded. For the unsupervised methods (FA and HVGS) and Spearman's correlation, RF serves as the downstream predictive model during cross-validation, allowing performance to be assessed on a common basis across all five approaches.

Supervised and unsupervised methods differ fundamentally in their use of label information. Supervised models learn a mapping $y = f(X)$, leveraging both the feature matrix (X) and the target labels (y) to determine feature importance, whereas unsupervised models operate exclusively on X, without any reference to y. In this paper, two unsupervised methods are employed: FA and HVGS, both of which rank features without access to the target labels.

FA is a hierarchical clustering-based technique that groups features according to their pairwise similarity (typically measured via correlation or Euclidean distance) and merges them into clusters that minimize within-cluster variance while maximizing between-cluster distance. The most significant features are identified as those contributing most to inter-cluster variance, that is, the features that are least redundant with others and thus carry the most distinct structural information. HVGS, originally developed for identifying informative genes in high-dimensional biological data, ranks features purely by their variance across observations, on the reasoning that features

exhibiting greater dispersion carry more information relevant to distinguishing between samples, independent of any outcome variable.

Because neither method incorporates the target label during the ranking process, both are structurally immune to label-driven bias, a source of instability to which supervised models are inherently susceptible. This makes FA and HVGS particularly well-suited as validation tools in engineering datasets: by identifying feature structure intrinsic to the physicochemical measurements themselves (e.g., cement, slag, water, aggregates), they provide an independent benchmark against which supervised, label-conditioned importance rankings can be cross-checked. Their use in this study is therefore not intended to replace supervised prediction, but to serve as a complementary, label-independent lens for assessing whether apparent feature importance reflects genuine data structure or merely an artifact of the modeling and labeling process.

For purposes of reproducibility and transparency, Python code (concrete.py) is publicly available on GitHub.[48]

3 | RESULTS

As detailed in Table 1, unsupervised feature ranking methods (FA, HVGS) and the correlation-based Spearman approach show markedly perfect higher stability in feature orderings across folds, whereas supervised models (RF, XGB) exhibit greater variability in their top features between CV6 and CV5. Despite this stability advantage, the unsupervised and correlation-based methods deliver competitive accuracy: FA (CV6 $R^2 = 0.8919$, CV5 $R^2 = 0.8831$), HVGS (CV6 $R^2 = 0.8900$, CV5 $R^2 = 0.8810$), and Spearman (CV6 $R^2 = 0.8928$, CV5 $R^2 = 0.8820$) are close to the supervised baselines.

Supervised models achieve the highest accuracies—XGB (CV6 $R^2 = 0.9341$, CV5 $R^2 = 0.9347$) and RF (CV6 $R^2 = 0.9138$, CV5 $R^2 = 0.9146$)—but their top feature rankings shift across folds (e.g., RF moves from Age first in CV6 to Cement first in CV5, and XGB elevates FlyAsh to first in CV5), underscoring less consistent feature importance.

In contrast, unsupervised rankings remain consistent: FA repeatedly prioritizes aggregates (FineAggregate, CoarseAggregate) alongside Cement, Water, and Superplasticizer, and HVGS consistently highlights Cement, BlastFurnaceSlag, and aggregate features. The Spearman method likewise preserves a stable core of Cement, Superplasticizer, Water, and aggregates across CV6 and CV5.

Overall, unsupervised (FA, HVGS) and Spearman methods outperform supervised models on stability of feature ordering while maintaining competitive predictive performance, making them strong choices when robust, interpretable feature ranking is critical.

4 | DISCUSSION

The results presented in this study provide compelling empirical evidence for a fundamental methodological concern that has long been theoretically acknowledged but rarely demonstrated with such clarity in pavement and structural engineering research: high target prediction accuracy and reliable feature importance assessment represent orthogonal analytical objectives that cannot be conflated without serious risk of methodological error. The concrete compressive strength dataset, precisely because of its familiarity and widespread use in the engineering informatics literature, serves as an ideal vehicle for exposing these pitfalls, as any distortions identified here are likely replicated silently and undetected across hundreds of published studies employing similar analytical pipelines.

The core finding is unambiguous. Supervised models, including both RF and XGB, achieved the highest predictive accuracies (CV6 $R^2 = 0.9138$ and 0.9341, respectively), yet exhibited the most pronounced instability in feature importance rankings under the LTFO perturbation. The shift from Age as the top-ranked feature in RF's CV6 configuration to Cement as the top-ranked feature in CV5, and XGB's elevation of FlyAsh to first position in CV5, are not minor numerical fluctuations. These are substantive reorderings that would lead practitioners to draw categorically different engineering conclusions depending on which model run they happened to consult. The observed instability is therefore not a statistical artifact; it is direct evidence that supervised importance scores reflect label-conditioned predictive contribution rather than true variable association.

In stark contrast, unsupervised dimensionality reduction approaches, specifically FA and HVGS, alongside Spearman's rank

TABLE 1 cross-validation accuracy and feature rankings across algorithms.

Method	CV6 R^2	CV5 R^2	Top 6 feature rankings of CV6	Top 5 feature rankings of CV5
RF	0.91379	0.914583	Age, Cement, Water, BlastFurnaceSlag, Superplasticizer, FineAggregate	Cement, Water, **FineAggregate**, Superplasticizer, **BlastFurnaceSlag**
XGB	0.934064	0.934655	Cement, Age, Superplasticizer, BlastFurnaceSlag, Water, FlyAsh	**FlyAsh, Age, Water, Superplasticizer, FineAggregate**
FA	0.891861	0.883096	FineAggregate, CoarseAggregate, Cement, Water, Superplasticizer, Age	CoarseAggregate, Cement, Water, Superplasticizer, Age
HVGS	0.889953	0.881023	Cement, BlastFurnaceSlag, FineAggregate, CoarseAggregate, FlyAsh, Age	BlastFurnaceSlag, FineAggregate, CoarseAggregate, FlyAsh, Age
Spearman	0.892818	0.881999	Age, Cement, Superplasticizer, Water, CoarseAggregate, FineAggregate	Cement, Superplasticizer, Water, CoarseAggregate, FineAggregate

Note: Bold entries indicate unstable features.

correlation, produced perfectly consistent feature orderings across both configurations while delivering competitive predictive performance with R^2 values ranging from 0.8810 to 0.8928. Methods that identify structure in the feature space without conditioning on outcome labels are not susceptible to the label-induced biases that systematically distort supervised importance measures, and they recover substantially the same ordered subsets under perturbation because they respond to genuine inter-feature relationships that persist regardless of which features are present in the pool.

The widespread adoption of SHAP-based explanations warrants particular scrutiny in light of these findings. The relationship captured by explain = SHAP(model) is not an independent assessment of variable importance; it is a post-hoc decomposition of the predictions generated by the underlying supervised model. When that model's importance structure is unstable or label-conditioned in ways that diverge from genuine associations, SHAP inherits, formalizes, and in some cases amplifies those distortions. The mathematical guarantees that SHAP provides apply to the consistency of the decomposition relative to the model's predictions, not to the correspondence between those predictions and true variable associations. The surge in SHAP applications documented across 33 articles in the *International Journal of Pavement Engineering* alone therefore warrants careful methodological scrutiny, as the apparent rigor conferred by Shapley value decomposition may obscure rather than resolve the underlying reliability problem.

When Spearman's correlation and FA agree on a feature ordering that remains stable under LTFO perturbation, the convergent evidence across methodologically independent approaches constitutes a substantially stronger basis for inferring true association than any single supervised model can provide regardless of its predictive accuracy. The fact that unsupervised preprocessing yields R^2 values within approximately 0.04–0.05 points of the best supervised models while providing dramatically more stable rankings suggests that the predictive penalty for adopting a more principled approach to feature assessment is modest and practically acceptable.

The absence of ground truth for feature importance assessment is not merely a technical inconvenience; it is a structural feature of the supervised learning paradigm that cannot be resolved through better models or larger datasets alone. The variables that a model identifies as important are important relative to predicting the target given the other features in the dataset, the training procedure employed, and the specific realization of the training sample, none of which are guaranteed to align with the physical mechanisms that actually govern the system under study. Acknowledging this limitation is a prerequisite for developing rigorous, multi-method validation frameworks that can provide warranted confidence in association claims despite the absence of ground truth.

The observed divergence in feature behavior is not arbitrary but reflects a structural difference in how each class of method resolves collinearity and label-conditioned variance among physicochemically related predictors. Age, Cement, and FlyAsh are known to be interrelated in their contribution to strength development: Cement content directly governs early hydration and strength gain, Age modulates the degree to which that hydration has progressed, and FlyAsh substitutes partially for Cement while contributing strength more slowly through pozzolanic reaction. Because supervised models optimize importance scores against the label y, whichever of these correlated variables best explains the residual variance in a given training fold—an artifact of sample composition and cross-validation partitioning—is assigned disproportionate importance, while its correlated counterparts are correspondingly suppressed. This explains why removing the top-ranked feature causes a different, correlated feature to abruptly assume the top rank rather than producing a gradual, interpretable redistribution of importance: the model is redistributing credit among interchangeable proxies for the same underlying physicochemical process, not revealing a stable causal hierarchy.

Unsupervised methods and Spearman's correlation, by contrast, are not exposed to this label-conditioned competition among correlated predictors. FA and HVGS rank features according to their intrinsic variance and clustering structure in the feature space alone, so a feature's relationship to physicochemically similar variables remains stable regardless of which specific training partition or label configuration is used. Consequently, the consistent rankings produced by these methods are more plausibly capturing durable structural relationships among the mix constituents themselves, rather than a fold-specific optimization artifact.

From an engineering standpoint, this distinction carries direct interpretive consequences. If a practitioner relies on a single supervised model run in which Age is ranked as the dominant feature, the natural design inference is that curing duration should be prioritized over adjustments to binder composition. If, in a different fold or model, Cement or FlyAsh assumes that top rank instead, the corresponding inference shifts toward binder proportioning as the primary design lever. Because both inferences cannot be simultaneously correct as stable causal claims, and because the identity of the top feature is shown here to be an artifact of sample partitioning rather than a robust physicochemical signal, engineers should treat single-run supervised importance rankings as insufficient evidence for prioritizing one design intervention over another. The consistent ordering recovered by unsupervised and correlation-based methods offers a more defensible starting point for such decisions, precisely because it is not contingent on which particular label-fitting exercise happened to be performed.

For engineering practice, we recommend a two-track approach that separates prediction from association assessment: supervised models (e.g., RF, XGB) should be retained for target prediction tasks, where their superior accuracy is directly verifiable and beneficial, whereas unsupervised methods (FA, HVGS) or Spearman's rank correlation, validated through the LTFO procedure, should serve as the primary basis for identifying influential variables that inform mix design, material substitution, and other decisions requiring trustworthy association claims. Although unsupervised approaches exhibit marginally lower predictive accuracy (R^2 between 0.8810 and 0.8928 vs. up to 0.9341 for supervised models), this modest cost is outweighed by their substantially greater ranking stability, making them the more defensible choice whenever feature importance—rather than prediction

alone—drives engineering decisions. When supervised and unsupervised rankings converge on the same top-ranked variables, practitioners can proceed with high confidence; when they diverge, this discrepancy should be treated as a signal warranting further investigation rather than being resolved by simply trusting the higher-accuracy model.

This study has several limitations. The analysis relies on a single dataset comprising 1030 samples with eight features, and generalizability to other datasets with different correlation structures, sample sizes, or target distributions cannot be assumed without further investigation. All methods were run with default hyperparameter settings, and carefully tuned supervised models might exhibit somewhat greater stability, though tuning for predictive accuracy is unlikely to resolve the fundamental label-conditioning problem. The LTFO protocol applies a single strong perturbation targeting the highest-ranked feature; alternative strategies involving lower-ranked or grouped feature removal may reveal additional facets of importance stability. Finally, the analysis addresses a single continuous outcome variable, and extension to multi-output or classification settings remains an open question.

Future work should prioritize replication of the LTFO framework across broader portfolios of engineering datasets to characterize the conditions under which supervised importance instability is most severe. Development of hybrid analytical frameworks that use unsupervised feature selection as a preprocessing filter before supervised prediction would preserve predictive advantages while grounding association claims in label-independent criteria. Systematic investigation of how importance instability relates to inter-feature collinearity would clarify whether the stability advantages of unsupervised methods reflect insensitivity to label information or different treatment of correlated feature groups. Application of the framework to factorial experimental datasets where true associations are known from controlled studies would provide external validation of the framework's discriminative capacity. Finally, developing quantitative metrics for dose–response coherence that can be integrated into automated validation pipelines would strengthen the practical utility of the proposed approach and reduce the risk that spurious associations propagate into published conclusions or engineering practice.

5 | CONCLUSIONS

This study establishes that target prediction accuracy and feature importance reliability constitute fundamentally distinct analytical objectives in concrete compressive strength modeling, with direct implications for how the civil engineering community should assess and trust AI-driven association analytics in sustainable and resilient infrastructure design. Using a LTFO validation framework applied to a publicly available concrete compressive strength dataset, we demonstrated that supervised models (RF, XGB), despite achieving superior predictive accuracy (R^2 up to 0.9341), produced highly unstable feature importance rankings under perturbation, with the top-ranked feature shifting entirely between cross-validation folds. In contrast, unsupervised methods (FA, HVGS) and Spearman's rank correlation yielded perfectly consistent rankings while sustaining competitive predictive performance (R^2 between 0.8810 and 0.8928), indicating that the cost of adopting more trustworthy feature assessment methods is modest.

These findings carry direct significance for sustainable and resilient infrastructure practice. Concrete mix design decisions informed by unreliable feature importance rankings risk misdirecting efforts toward variables that appear influential only as artifacts of a particular model or training sample, rather than reflecting genuine physicochemical drivers of strength. Such misattribution has tangible consequences: it can lead engineers to over- or under-specify cementitious materials, misjudge the true contribution of supplementary cementitious materials such as fly ash and blast furnace slag, and consequently undermine decarbonization efforts that depend on accurately identifying which constituents can be safely reduced without compromising structural performance. Because cement production remains a major contributor to global CO_2 emissions, correctly and stably identifying the true influence of substitute materials is essential for advancing low-carbon concrete formulations and for supporting resource-efficient use of raw materials.

Reliable feature association also underpins lifecycle performance and service continuity. Infrastructure decisions grounded in spurious or volatile associations may result in structures that underperform relative to design expectations, increasing the likelihood of premature maintenance, repair, or replacement, and thereby compromising both durability and adaptability over the structure's service life. By contrast, associations validated through the consistency and dose–response criteria proposed here offer a more credible foundation for long-term performance prediction, enabling engineers to design mixes and structural elements with greater confidence in their resilience under varying service conditions. This directly supports the goals of improved and de-risked infrastructure decision-making, as practitioners can distinguish genuine physicochemical relationships from artifacts of a particular supervised algorithm before committing to material specifications or design codes.

The demonstrated vulnerability of SHAP-based explanations to inherit and amplify the biases of underlying supervised models further underscores the need for caution when such tools are used to justify sustainability-related claims, for example, when SHAP importances are invoked to argue that a given supplementary material contributes little to strength and can therefore be increased for decarbonization purposes without adequate cross-validation using consistency and dose–response criteria. Our results suggest that such claims should be corroborated through unsupervised and correlation-based methods before being translated into material substitution strategies or design guidance.

We therefore recommend that civil engineering researchers and practitioners adopt a multi-method validation protocol, combining supervised prediction with unsupervised feature analysis and traditional statistical correlation, and apply the LTFO procedure whenever feature importance findings are intended to inform infrastructure design, material selection, or sustainability claims. Such a protocol offers a practical and reproducible safeguard against spurious

associations, strengthening the reliability of engineering decision-making and contributing to the broader goals of resource efficiency, decarbonization, and resilient, de-risked infrastructure development that lie at the core of sustainable civil engineering design. Future work extending this framework to additional datasets, alternative perturbation strategies, and classification or multi-output settings will further consolidate its role as a general-purpose validation tool for association claims in structural and materials research.

AUTHOR CONTRIBUTIONS

Yoshiyasu Takefuji completed this research and wrote the program and this article. According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1 287 415 scholars in life sciences, 22nd out of 805 705 in COVID-19, and 1st out of 109 919 in environmental sciences.

FUNDING INFORMATION

This research has no funding.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

DATA AVAILABILITY STATEMENT

The author has no permission to share datasets. The Python code is publicly available at GitHub.

ORCID

Yoshiyasu Takefuji https://orcid.org/0000-0002-1826-742X

REFERENCES

1. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640–8. https://doi.org/10.1097/EDE.0b013e31818131e7
2. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? Jama. 2013;309(3):241–2. https://doi.org/10.1001/jama.2012.96867
3. Ioannidis JPA. Genetic associations: false or true? Trends Mol Med. 2003;9(4):135–8. https://doi.org/10.1016/S1471-4914(03)00030-3
4. Roberts MR, Ashrafzadeh S, Asgari MM. Research techniques made simple: interpreting measures of association in clinical research. J Invest Dermatol. 2019;139(3):502–511.e1. https://doi.org/10.1016/j.jid.2018.12.023
5. Lai Q, Dannenfelser R, Roussarie JP, Yao V. Disentangling associations between complex traits and cell types with seismic. Nat Commun. 2025;16(1):8744. https://doi.org/10.1038/s41467-025-63753-z
6. Prada D, Ritz B, Bauer AZ, Baccarelli AA. Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology. Environ Health. 2025;24(1):56. https://doi.org/10.1186/s12940-025-01208-0
7. Stamatakis E, Ahmadi M, Biswas RK, Del Pozo Cruz B, Thøgersen-Ntoumani C, Murphy MH, et al. Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences. Br J Sports Med. 2025;59(5):316–24. https://doi.org/10.1136/bjsports-2024-108484
8. Ye M, He Y, Xia Y, Zhong Z, Kong X, Zhou Y, et al. Association between bowel movement frequency, stool consistency and MAFLD and advanced fibrosis in US adults: a cross-sectional study of NHANES 2005-2010. BMC Gastroenterol. 2024;24(1):460. https://doi.org/10.1186/s12876-024-03547-7
9. Underwood BR, Lourida I, Gong J, Tamburin S, Tang EYH, Sidhom E, et al. Data-driven discovery of associations between prescribed drugs and dementia risk: a systematic review. Alzheimers Dement. 2025;11(1):e70037. https://doi.org/10.1002/trc2.70037
10. Takefuji Y. Model-specific feature importances: distinguishing true associations from target-feature relationships. J Affect Disord. 2025;369:390–1. https://doi.org/10.1016/j.jad.2024.10.019
11. Singh A, Southam L, Hatzikotoulas K, Rayner NW, Suzuki K, Taylor HJ, et al. Correcting for genomic inflation leads to loss of power in large-scale genome-wide association study meta-analysis. Genet Epidemiol. 2025;49(6):e70016. https://doi.org/10.1002/gepi.70016
12. Parr T, Hamrick J, Wilson JD. Nonparametric feature impact and importance. Inform Sci. 2024;653:119563. https://doi.org/10.1016/j.ins.2023.119563
13. Watson DS, Wright MN. Testing conditional independence in supervised learning algorithms. Mach Learn. 2021;110(8):2107–29. https://doi.org/10.1007/s10994-021-06030-6
14. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31–57. https://doi.org/10.1145/3236386.3241340
15. Fisher A, Rudin C, Dominici F. 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. 2019;20:177.
16. Lenhof K, Eckhart L, Rolli LM, Lenhof HP. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. Brief Bioinform. 2024;25(5):bbae379. https://doi.org/10.1093/bib/bbae379
17. Mandler H, Weigand B. A review and benchmark of feature importance methods for neural networks. ACM Comput Surv. 2024;56(12):318. https://doi.org/10.1145/3679012
18. Potharlanka JL, Bhat MN. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci Rep. 2024;14(1):2923. https://doi.org/10.1038/s41598-024-53141-w
19. Wood D, Papamarkou T, Benatan M, Allmendinger R. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. Data Min Knowl Discov. 2024;38:4184–216. https://doi.org/10.1007/s10618-024-01070-7
20. Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger A, Goebel R, Fong R, Moon T, Müller K-R, Samek W, editors. xxAI – Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science. Volume 13200. Cham: Springer; 2022. p. 39–68. https://doi.org/10.1007/978-3-031-04083-2_4
21. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLoS Digit Health. 2023;2(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278
22. Ugirumurera J, Bensen EA, Severino J, Sanyal J. Addressing bias in bagging and boosting regression models. Sci Rep. 2024;14(1):18452. https://doi.org/10.1038/s41598-024-68907-5
23. Alaimo Di Loro P, Scacciatelli D, Tagliaferri G. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. Stat Methods Appl. 2023;32:237–70. https://doi.org/10.1007/s10260-022-00643-4
24. Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy. 2022;24(5):687. https://doi.org/10.3390/e24050687
25. Steiner PM, Kim Y. The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. J Causal Inference. 2016;4(2):20160009. https://doi.org/10.1515/jci-2016-0009
26. Wallace ML, Mentch L, Wheeler BJ, Tapia AL, Richards M, Zhou S, et al. Use and misuse of random forest variable importance metrics in

medicine: demonstrations through incident stroke prediction. BMC Med Res Methodol. 2023;23(1):144. https://doi.org/10.1186/s12874-023-01965-x
27. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007;8:25. https://doi.org/10.1186/1471-2105-8-25
28. Salles T, Rocha L, Gonçalves M. A bias-variance analysis of state-of-the-art random forest text classifiers. Adv Data Anal Classif. 2021;15:379–405. https://doi.org/10.1007/s11634-020-00409-4
29. Dunne R, Reguant R, Ramarao-Milne P, Szul P, Sng LMF, Lundberg M, et al. Thresholding Gini variable importance with a single-trained random forest: an empirical Bayes approach. Comput Struct Biotechnol J. 2023;21:4354–60. https://doi.org/10.1016/j.csbj.2023.08.033
30. Huti M, Lee T, Sawyer E, King AP. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. Clin Image Based Proced Fairness AI Med Imaging Ethical Philos Issues Med Imaging. 2023;14242:225–34. https://doi.org/10.1007/978-3-031-45249-9_22
31. Wu L. A review of the transition from Shapley values and SHAP values to RGE. Stat. 2025;59(5):1161–83. https://doi.org/10.1080/02331888.2025.2487853
32. Bilodeau B, Jaques N, Koh PW, Kim B. Impossibility theorems for feature attribution. Proc Natl Acad Sci U S A. 2024;121(2):e2304406120. https://doi.org/10.1073/pnas.2304406120
33. Huang X, Marques-Silva J. On the failings of Shapley values for explainability. Int J Approx Reason. 2024;171:109112. https://doi.org/10.1016/j.ijar.2023.109112
34. Hooshyar D, Yang Y. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access. 2024;12:137472–90. https://doi.org/10.1109/ACCESS.2024.3463948
35. Lones MA. Avoiding common machine learning pitfalls. Patterns. 2024;5(10):101046. https://doi.org/10.1016/j.patter.2024.101046
36. Kumar I, Scheidegger C, Venkatasubramanian S, Friedler S. Shapley residuals: Quantifying the limits of the Shapley value for explanations. Adv Neural Inf Process Syst. 2021;34:26598–608.
37. Létoffé O, Huang X, Marques-Silva J. Towards trustable SHAP scores. Proceedings of the AAAI Conference on Artificial Intelligence. 2025;39(17):18198–208. https://doi.org/10.1609/aaai.v39i17.34002
38. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci. 2024;17(11):e70056. https://doi.org/10.1111/cts.70056
39. Coupland H, Scheidwasser N, Katsiferis A, Davies M, Flaxman S, Hulvej Rod N, et al. Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health. 2025;25(1):1520. https://doi.org/10.1186/s12889-025-22705-4
40. Takefuji Y. Beyond XGBoost and SHAP: unveiling true feature importance. J Hazard Mater. 2025;488:137382. https://doi.org/10.1016/j.jhazmat.2025.137382
41. Takefuji Y. The reliability gap: why high predictive accuracy doesn't guarantee stable feature importance. Mar Pollut Bull. 2026;226:119398. https://doi.org/10.1016/j.marpolbul.2026.119398
42. Takefuji Y. Evaluating linear parametric Poisson regression vs non-parametric unsupervised learning in ulcerative colitis data. Gastroenterology. 2026;171(2):384–6. https://doi.org/10.1053/j.gastro.2026.04.011
43. Takefuji Y. When high accuracy misleads: stability limits of supervised feature importance in QSAR biodegradation. Chemosphere. 2026;396:144846. https://doi.org/10.1016/j.chemosphere.2026.144846
44. Takefuji Y. Beyond prediction: assessing stability in feature selection methods for materials science applications. Comput Mater Sci. 2026;267:114609. https://doi.org/10.1016/j.commatsci.2026.114609
45. Takefuji Y. Ensuring reliable feature importance in food chemistry AI. Food Chem. 2026;509:148515. https://doi.org/10.1016/j.foodchem.2026.148515
46. Takefuji Y. The absence of true association validation in supervised feature importance: evidence from battery capacity prediction for second-life applications. J Energy Storage. 2026;180:124146. https://doi.org/10.1016/j.est.2026.124146
47. Yeh I-C. Modeling of strength of high-performance concrete using artificial neural networks. Cem Concr Res. 1998;28(12):1797–808. https://doi.org/10.1016/S0008-8846(98)00165-3
48. GitHub. Concrete.py. https://github.com/y-takefuji/concrete

How to cite this article: Takefuji Y. A leave-top-feature-out framework for validating feature importance reliability in concrete compressive strength modeling. Civil Engineering Design. 2026. https://doi.org/10.1002/cend.70026