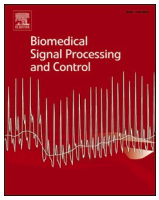
Biomedical Signal Processing and Control 129 (2027) 111577



Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



Short communication

Predictive accuracy does not establish feature-importance robustness: a leave-top feature-out analysis in late-life depression

Yoshiyasu Takefuji [a,*]

[a] SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan

ARTICLE INFO

Keywords:
Late-life depression
Machine learning
Explainable AI
Feature importance
Transcultural prediction

ABSTRACT

Supervised machine learning models are increasingly applied in biomedical signal processing and control for feature assessment; however, a critical epistemological gap persists in that no established ground truth exists against which feature importance can be validated. This paper theoretically establishes that supervised models exhibit two distinct dimensions of accuracy, target prediction accuracy and feature importance accuracy, which are not interchangeable. High target prediction accuracy does not guarantee reliable feature importance, as importance metrics reflect contributions to prediction rather than true causal associations. Furthermore, SHAP-based explanations inherit and may amplify the underlying model's distortions rather than correct them, rendering their outcomes contingent on model-specific artifacts rather than confirmed causal insights. To address this gap, we draw on the extended Bradford Hill criteria to propose consistency and dose–response relationships as principled validation criteria, and we introduce a novel leave-top-feature-out diagnostic framework to empirically test the stability of feature importance rankings. Empirical analysis of a late-life depression dataset comparing nine feature selection algorithms demonstrates that unsupervised models yield superior stability in feature ranking, whereas supervised models exhibit label-driven instability that SHAP integration fails to resolve, underscoring the need for validation frameworks that extend beyond model-dependent interpretability methods.

Introduction

The goal of this paper is to identify prevalent AI misapplications in supervised models for feature importance or selection assessment due to the absence of ground truth, indicating the need of external feature validation while researchers conflate a critical distinction between target prediction accuracy and feature importance accuracy. We theoretically discuss and empirically demonstrate the pitfalls of supervised models while proposing feature consistency and dose–response relationships as validation methods for consistent predictive associations—a critical step missing in existing research. *Biomedical Signal Processing and Control* has published a growing body of research utilizing machine learning for feature assessment, with 518 articles on feature importance (80 in 2025 and 282 in 2026), 2477 on feature selection (390 in 2025 and 672 in 2026), and 336 on SHapley Additive exPlanations (SHAP) (52 in 2025 and 224 in 2026), with continued publications projected through 2026. This significant trend reflects increasing interest in leveraging supervised models within psychiatric research datasets. However, fundamental methodological misunderstandings regarding these techniques have led to potentially erroneous conclusions throughout the psychiatric literature due to a critical epistemological gap: the absence of ground truth in feature assessment validation. The core issue stems from researchers inappropriately treating machine learning algorithms as statistical inference tools capable of identifying causal or associative relationships rather than as prediction mechanisms optimized for outcome forecasting.

Using a publicly available dataset on transcultural late-life depression, we empirically demonstrate how commonly used importance metrics can prioritize spurious correlations and overlook genuine causal factors, particularly in high-dimensional datasets with complex interactions typical of psychiatric research. We highlight that supervised models exhibit two distinct dimensions of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance lacks corresponding ground truth for validation purposes. To accurately determine consistent predictive associations between predictors and outcomes, two critical criteria must be examined: consistency across models and dose–response relationships—criteria that

* Address: Emeritus Professor (Keio Univ.), AI laboratory, SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa 252-0029, Japan.
E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.bspc.2026.111577
Received 17 July 2026; Received in revised form 28 August 2026; Accepted 21 September 2026
1746-8094/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

existing studies have systematically failed to address.

A fundamental yet frequently overlooked limitation in explainable machine learning research is the absence of an established ground truth against which feature importance explanations can be validated, a constraint that many researchers fail to fully acknowledge. The explanations generated by methods such as SHAP are model-dependent approximations rather than objective reflections of true underlying causal relationships. Despite this critical caveat, such methods continue to be widely adopted without adequate acknowledgment of their interpretive constraints. Among 336 SHAP-based feature studies, two representative cases illustrate this issue [1,2]. In both, researchers employed supervised models to identify the best-performing prediction model and then applied SHAP for interpretation to determine the most significant drivers, implicitly assuming that a high-performing prediction model can generate reliable feature importance for meaningful interpretation. However, feature importance derived from any supervised model inherently lacks ground truth for validation, regardless of the model's predictive accuracy. Consequently, the outcomes of SHAP-based analysis remain contingent on a specific supervised model and represent unvalidated estimates of feature importance rather than confirmed causal insights.

When a supervised model produces feature importances, these metrics reflect contributions to prediction rather than true causal associations. Consequently, high target prediction accuracy does not guarantee reliable feature importances, and the derived importances are inherently distorted, potentially leading to erroneous interpretations. Researchers must acknowledge that supervised models inherently possess two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance accuracy lacks a corresponding ground truth for validation, revealing a fundamental gap in feature assessment that may result in misleading conclusions. This distinction is critical, as conflating predictive performance with explanatory validity undermines the integrity of clinical and behavioral inferences drawn from such models. Furthermore, the function explainer=SHAP(model) demonstrates that SHAP is constrained to rely exclusively on a given supervised model whose feature importances remain unvalidated, thereby inherently inheriting and potentially amplifying the underlying biases embedded within those importances. This limitation raises serious concerns about the uncritical application of SHAP as an explainability tool, particularly in high-stakes domains such as healthcare, where misattributed feature relevance can have significant consequences for clinical decision-making and patient outcomes.

There is no single algorithm that yields “true” associations between features and outcomes. Based on the extended Bradford Hill criteria [3], establishing robust associations requires examining consistency and dose–response patterns across model specifications and data subsets [3–14]. This paper proposes a leave-top-feature-out diagnostic approach: select the top n features from the full model, remove the highest-ranked feature to form a reduced dataset, then reselect the top n − 1 features and compare ranking orders. If removing the top feature substantially perturbs subsequent rankings, this sensitivity illuminates correlation structures and tests whether importance signals persist (consistency) and adjust in magnitude as expected (dose–response). This helps differentiate stable, causally plausible drivers from artifacts of multicollinearity or model-dependent attributions.

Supervised models often show instability in feature ranking due to label-driven errors, while unsupervised models and non-target prediction methods exhibit stronger stability in feature ranking orders. Through our analysis of a publicly available dataset (16), we demonstrate the methodological pitfalls of using supervised models for feature importance assessment, highlighting the risks of misinterpretation when these fundamental distinctions are overlooked.

We highlight that supervised models exhibit two distinct dimensions of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance lacks corresponding ground truth for validation purposes. To accurately determine consistent predictive associations between predictors and outcomes, two critical criteria must be examined: consistency across models and dose–response relationships—criteria that existing studies have systematically failed to address.

This paper methodically examines consistency and dose–response relationships of late-life depression by comparing nine distinct feature selection algorithms. Cross-validation techniques are implemented to objectively assess the quality and reliability of selected features across multiple data partitions, ensuring robustness against sampling variability. Additionally, we employ a novel “leave-top-feature-out” approach, systematically removing the highest-ranked feature from each algorithm and measuring the subsequent impact on model performance, to quantify both the consistency of feature importance rankings across different algorithmic frameworks and to establish whether clear dose–response relationships exist between feature presence and outcome prediction. This dual-validation framework directly addresses the epistemological gap identified earlier: the absence of ground truth for feature importance validation. By treating consistency across model specifications and dose–response patterns as surrogate indicators of consistent predictive associations, rather than relying on target prediction accuracy alone, this framework operationalizes the theoretical distinction between prediction accuracy and feature importance accuracy that prior studies. Critically, this approach acknowledges that no single algorithm yields definitive causal attributions, and that meaningful inference requires triangulation across multiple algorithmic frameworks and data subsets. Together, these theoretical and empirical contributions establish a more rigorous methodological standard for feature importance assessment in psychiatric machine learning research, one that moves beyond prediction performance as a proxy for interpretability and toward principled validation of associative claims.

As illustrated in Fig. 1, high predictive accuracy in supervised machine learning cannot be conflated with the validity of feature importance rankings. Because supervised algorithms minimize loss solely to optimize target prediction, internal metrics offer no empirical ground truth regarding true biological or causal relevance, leaving conventional feature importance inherently unverifiable (Premise 1). Post-hoc explainers such as SHAP compound this issue: since an explainer strictly interrogates the surrogate model (explainer = SHAP(model)), it inherently inherits and potentially amplifies the model’s internal inductive biases rather than resolving them (Premise 2). Consequently, both raw feature attribution and SHAP outputs lack independent verification, establishing an indispensable need for model-agnostic external feature validation. To bridge this theoretical gap, we formulate the Leave-Top-Feature-Out (LTFO) framework. Unlike existing ablation schemes that evaluate unordered feature subsets, LTFO explicitly assesses ordered (ranked) feature sets, thereby testing both feature ranking consistency and the dose–response degradation of model performance when the primary signal is perturbed.

While conventional ablation paradigms fail by treating features as unordered sets—thereby discarding crucial hierarchical ranking information—the LTFO framework explicitly operates on ordered feature sequences. By systematically removing the top-ranked feature and re-evaluating the residual hierarchy, LTFO concurrently examines ranking stability and predictive dose–response relationships. Serving as an indispensable external validation layer that decouples feature evaluation from model bias, LTFO provides a standardized, order-aware benchmark that should be routinely integrated into machine learning feature assessment workflows.

This limitation cannot be resolved simply by employing post-hoc model explainers such as SHAP. By formulation, SHAP values are mathematically conditional on the underlying trained model (explainer=SHAP(model)). Consequently, SHAP does not generate an independent evaluation of feature validity; instead, it inherits, reflects, and potentially amplifies any label-driven instability, multicollinearity, or

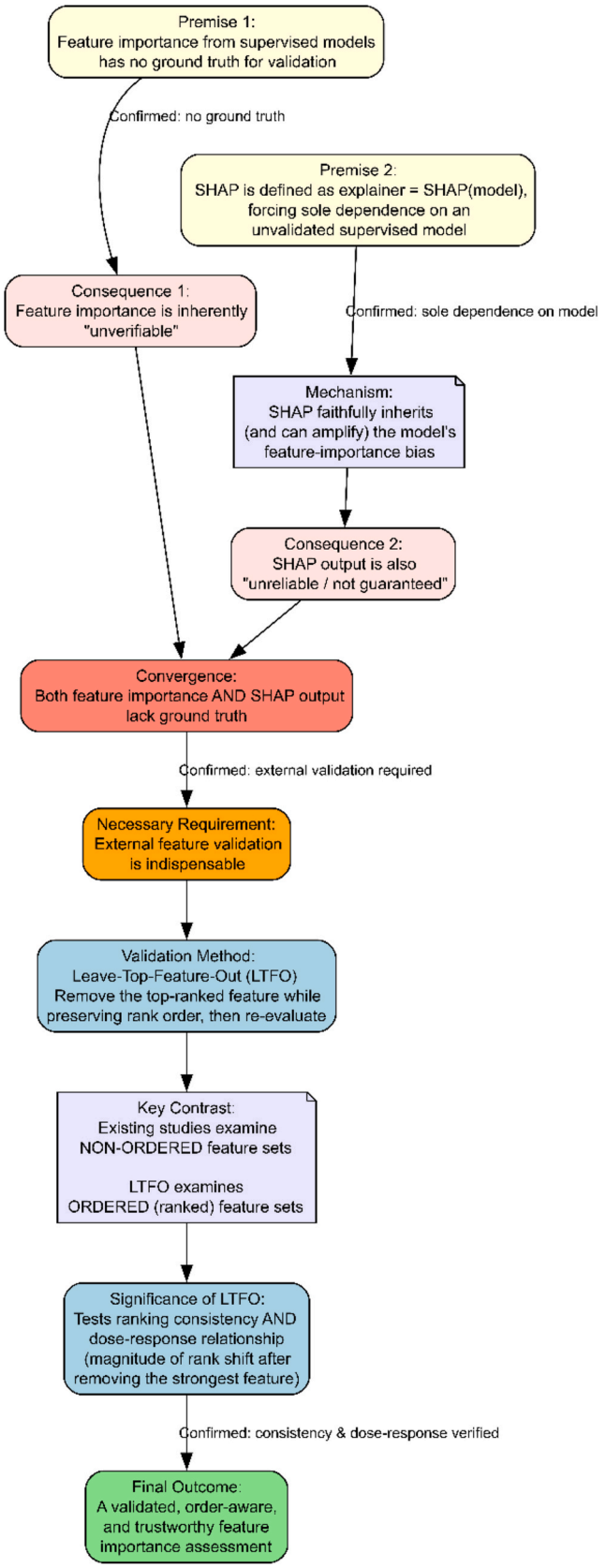


Fig. 1. Theoretical progression from the epistemological gaps of supervised feature importance and SHAP to rank-aware validation via Leave-Top-Feature-Out (LTFO).

structural biases present in the predictive model. Because both supervised feature-selection algorithms and SHAP inherently lack ground-truth validation, assessing the true reliability of feature rankings requires an external, model-agnostic validation paradigm.

To operationalize external validation, we introduce the Leave-Top-Feature-Out (LTFO) diagnostic framework, which fundamentally differs from standard feature-ablation methods. While conventional ablation studies evaluate unordered subsets of features, LTFO explicitly assesses ordered, ranked feature sets. By systematically removing top-ranked features, the framework operationalizes key diagnostic criteria adapted from epidemiological validation principles—specifically, evaluating ranking consistency and dose–response dynamics in response to primary feature perturbations. In doing so, LTFO moves beyond superficial stability metrics to provide a rigorous, order-aware diagnostic for feature-importance robustness.

Methods

This paper examines feature quality through three complementary dimensions: cross-validation for assessing predictive contribution, and leave-top-feature-out approaches for evaluating consistency and dose–response relationships. In contrast, existing studies have relied exclusively on cross-validation accuracy, implicitly assuming that high target prediction accuracy guarantees reliable feature importance, a fundamental methodological conflation this paper seeks to correct.

This paper demonstrates a critical distinction in feature ranking stability between supervised and unsupervised machine learning approaches. Using a public dataset of late-life depression from OpenML with 181 instances and 35 features [15]. We intentionally remove instances with more than 50 % missing data. We do conduct no scaling, no transformation, no imputation not to alter raw values of the dataset. Nine feature selection algorithms include diverse methods such as supervised models (XGBoost: XGB, Random Forest: RF, Gradient Boosting Machine: GBM, LASSO), unsupervised models (Feature Agglomeration: FA, Highly Variable Gene Selection: HVGS), and non-target prediction nonparametric statistical methods (Spearman's correlation) with/without SHAP. For purposes of reproducibility and transparency, all Python code used in this analysis, available as shapanalysis2.py, has been made publicly accessible at GitHub [16].

Feature importance, feature selection, and dimensionality-reduction methods differ fundamentally in how they define feature relevance, but Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS) share a property directly relevant to this study: neither uses the target label during feature scoring. FA identifies feature groups through unsupervised clustering on feature similarity, and HVGS ranks features by variance in their own distribution; both criteria are computed independently of any label-driven optimization. This absence of label dependence means these methods cannot exhibit the specific failure mode under investigation here — instability in feature rankings arising from a model re-optimizing its decision boundary around correlated predictors when a label-informative feature is removed. We therefore include FA and HVGS not as conventional alternatives for supervised feature selection, but as label-free reference points that isolate whether the instability we observe in supervised methods originates from their use of the label during scoring.

Because HVGS was originally developed for identifying highly variable genes in transcriptomic data, we clarify its use here on clinical/psychiatric variables. The variance-ranking principle underlying HVGS is a general statistical criterion — features with greater dispersion carry more discriminative information — and is not intrinsically tied to gene-expression measurements. Variance-based feature ranking of this kind has been applied to non-transcriptomic data, supporting its use as a domain-general tool. We use it here for the same reason as FA: as a second, methodologically distinct, label-free benchmark.

Feature importance, feature selection, and dimensionality reduction all require external validation, because none of these methods produces

an internally verifiable measure of correctness: each yields a ranking or grouping of features without any ground-truth reference against which that output can be checked, particularly in observational clinical/psychiatric datasets where the true causal structure among variables is not known. This absence of ground truth means that a method's own reported accuracy or importance score cannot certify that the selected features reflect genuine feature-outcome associations rather than artifacts of a particular model fit or sample. Existing studies have largely neglected this requirement, typically reporting a single feature-ranking or accuracy result from one method on one dataset split, without subjecting that result to any external check of stability or reproducibility. The leave-top-feature-out (LTFO) procedure proposed in this paper is intended as one such external validation step: it tests whether a method's feature ranking and resulting accuracy hold up when the top-ranked feature is removed and the method is reapplied, using cross-method comparison — rather than the method's own internal accuracy metric — as the validating signal. Useful variance- or similarity-based tools such as HVGS and FA, despite originating in different application domains, can be repurposed across domains and disciplines as label-free benchmarks within this type of external validation framework.

For consistency and dose–response relationship assessment, a leave-top1-out approach is conducted: select top 5 features from full set (set1), remove the highest feature from full set to create a reduced dataset, re-select top 4 features from the reduced dataset (set2), and compare feature importance ranking orders between two sets. Removing the highest feature can produce a significant impact to feature ranking orders to accommodate consistency and dose–response relationships.

This methodological approach serves a dual purpose: first, it enables systematic evaluation of how different algorithms respond to the removal of supposedly influential features, and second, it reveals whether the importance rankings demonstrate the logical stability expected in genuine causal relationships. In a theoretically sound feature importance framework, the removal of the top feature should not drastically reorganize the entire importance hierarchy if true relationships exist between features and the target variable. The leave-top1-out technique specifically tests this property by examining if the relative importance order among remaining features maintains consistency when the highest-ranked feature is removed.

Furthermore, 5-fold cross-validation with selected features is conducted for top 5 and top 4 features to assess target prediction accuracy. This procedure divides the dataset into five equal parts, with four parts used for training and one part for validation in each iteration, cycling through all possible combinations. By comparing predictive performance using both the top 5 features from the full dataset and the top 4 features from the reduced dataset, we can determine whether feature importance rankings translate meaningfully to practical predictive utility. This cross-validation step provides critical insight into how the stability of feature importance rankings relates to actual predictive performance and helps validate whether the identified important features genuinely contribute to the model's predictive capability.

Results

The analysis presented in Table 1 reveals striking performance differences between these two methodological approaches. Notably, unsupervised models not only outperform their supervised counterparts in cross-validation accuracy but also demonstrate remarkable stability across different validation scenarios. Both FA and HVGS achieved perfect accuracy scores with perfect consistency, maintaining identical performance (1.00) across both cross-validation (CV5) and cross-validation (CV4) evaluations. This consistency underscores the robustness of unsupervised feature selection methods in clinical prediction tasks.

Most significantly, the addition of SHAP values to supervised models completely failed to address their inherent instability problems. Despite SHAP's intended purpose of improving model interpretability and potentially enhancing performance, supervised models with SHAP exhibited identical instability patterns to their non-SHAP counterparts. Random Forest with SHAP (RF-SHAP) showed the same dramatic accuracy drop (1.00 → 0.802) as standard RF, while GBM-SHAP showed a comparable performance decline (1.00 → 0.794), closely mirroring the decline observed for standard GBM (1.00 → 0.763). This conclusively demonstrates that SHAP integration did nothing to mitigate the instability issues plaguing supervised approaches.

In striking contrast, unsupervised methods maintained virtually identical performance across both validation schemes without requiring any additional explanation frameworks. Similarly, the Spearman correlation method demonstrated stability comparable to most supervised approaches in the CV4 scenario (0.771), regardless of SHAP integration.

These findings strongly suggest that unsupervised approaches provide not just superior accuracy but critically important stability and generalizability for this dataset, while SHAP-based enhancements fail to improve the fundamental instability issues of supervised models. This paper empirically demonstrates that high target prediction accuracy

Table 1
Cross-validation accuracy and feature rankings across algorithms.

Method	cv5 acc	cv5 F1cls1	cv5 F1cls2	cv5 F1cls3	cv4 acc	cv4 F1cls1	cv4 F1cls2	cv4 F1cls3	top5 features	top4 features
RF	1	1	1	1	0.802	0.732	0.75	0.872	HDRS, HDRS_cognition, CGLi, UCLA, GDS_15	HDRS_cognition, CGLi, **GDS_15, UCLA**
RF-SHAP	1	1	1	1	0.802	0.732	0.75	0.872	HDRS, HDRS_cognition, CGLi, UCLA, GDS_15	HDRS_cognition, CGLi, **GDS_15, UCLA**
XGB	1	1	1	1	0.695	0.609	0.612	0.797	HDRS, HDRS_cognition, UCLA, onset_diagnosis_years, age	HDRS_cognition, **CGLi, suicide_risk, PDQ_3**
XGB-SHAP	1	1	1	1	0.733	0.681	0.645	0.82	HDRS, UCLA, onset_diagnosis_years, age, HDRS_cognition	**HDRS_cognition, CGLi**, age, **UCLA**
GBM	1	1	1	1	0.763	0.718	0.712	0.824	HDRS, age, duration_illness, R, UCLA	**HDRS_cognition, CGLi, UCLA, number**
GBM-SHAP	1	1	1	1	0.794	0.718	0.75	0.857	HDRS, number, gender, age, family status	**HDRS_cognition, CGLi, UCLA, MMSE**
FA	1	1	1	1	1	1	1	1	duration_illness, UCLA, duration employment, HDRS, MMSE	UCLA, duration employment, HDRS, MMSE
HVGS	1	1	1	1	1	1	1	1	number, duration_illness, UCLA, onset_diagnosis_years, HDRS	duration_illness, UCLA, onset_diagnosis_years, HDRS
Spearman	1	1	1	1	0.771	0.732	0.706	0.84	HDRS, HDRS_cognition, CGLi, GDS_15, GDS_suic	HDRS_cognition, CGLi, GDS_15, GDS_suic
LASSO	0.939	0.884	0.922	0.974	0.840	0.75	0.796	0.908	HDRS, MMSE, N_episodes, GDS_15, age	**CGLi, HDRS_cognition, family status, diagnosis**

Bold entries indicate features with inconsistent and unstable rankings.

does not guarantee reliable feature importances due to label-driven errors while unsupervised models exhibit stronger stability in feature importance ranking orders due to the absence of label-driven errors.

While statistical significance (p-values) (the number of instances should be larger than 600) is not applicable due to small size dataset, we computed CI and standard deviation of accuracy of ten algorithms with 10-time runs. Python code (ci.py) is created to generate confidential index and standard deviation of all algorithms. Table 2 illustrates predictive accuracy comparison across feature selection and attribution methods under cross-Validation (Mean ± SD [95% CI]).

Discussion

Existing feature ablation studies operate on unordered feature sets. In standard ablation analysis, a feature (or subset of features) is removed, typically arbitrarily, exhaustively, or one-at-a-time in no meaningful sequence, and the resulting change in performance is observed. This procedure treats "which feature was removed" as an unordered label: the ablation of feature A and the ablation of feature B are evaluated as parallel, independent trials, with no ordering relationship between them. Standard ablation, by construction, discards rank information entirely: it can indicate that removing a feature changes performance, but it cannot test whether performance degrades in a manner consistent with the feature's rank position in an importance ordering.

LTFO is constructed specifically to preserve and test that ordering. LTFO removes the top-ranked feature according to each method's own importance ordering, then re-derives the ranking on the remaining features and again removes the new top-ranked feature, proceeding sequentially down the ordered list (top-5 → top-4 → …). This is not an arbitrary or exhaustive removal scheme; it is a removal scheme indexed by rank itself. The object under test is therefore the ordered set of features, not the feature set as an unordered collection. This distinction allows LTFO to ask a question standard ablation cannot address: does performance degrade in a manner consistent with the claimed ranking — i.e., does removing the feature a method identifies as most important produce a systematically larger effect than removing a lower-ranked one, and is that ordering itself stable across resampling and across methods? An unordered ablation procedure has no mechanism to pose this question, because it has no rank structure to interrogate in the first place.

Unlike conventional feature ablation, which removes features without regard to their relative importance ranking and therefore evaluates an unordered set of removals, LTFO removes features sequentially in rank order — the top-ranked feature first, followed by re-ranking and removal of the new top-ranked feature — and thereby evaluates an ordered set of removals. This ordering is essential to the method: it tests whether a method's claimed importance ranking is itself stable and consequential under sequential perturbation, a question that is not addressable by unordered ablation, which discards rank information by construction.

Table 2
Predictive Accuracy Comparison Across Feature Selection and Attribution Methods Under Cross-Validation (Mean ± SD [95 % CI]).

method	CV5_accuracy	CV4_accuracy
RF	0.989 ± 0.007 [0.984, 0.995]	0.794 ± 0.020 [0.779, 0.808]
RF-SHAP	0.989 ± 0.007 [0.984, 0.995]	0.792 ± 0.017 [0.780, 0.805]
XGB	1.000 ± 0.000 [1.000, 1.000]	0.737 ± 0.026 [0.719, 0.756]
XGB-SHAP	1.000 ± 0.000 [1.000, 1.000]	0.792 ± 0.035 [0.766, 0.817]
GBM	1.000 ± 0.000 [1.000, 1.000]	0.782 ± 0.020 [0.768, 0.797]
GBM-SHAP	1.000 ± 0.000 [1.000, 1.000]	0.795 ± 0.027 [0.775, 0.814]
FA	1.000 ± 0.000 [1.000, 1.000]	0.996 ± 0.005 [0.992, 1.000]
HVGS	0.998 ± 0.004 [0.995, 1.000]	0.999 ± 0.002 [0.998, 1.001]
Spearman	0.986 ± 0.009 [0.980, 0.992]	0.785 ± 0.015 [0.774, 0.795]
LASSO	0.930 ± 0.014 [0.920, 0.940]	0.826 ± 0.012 [0.817, 0.835]

The findings of this study have important implications for the growing body of machine learning research in psychiatric and geriatric health contexts, particularly regarding the appropriate interpretation and validation of feature importance metrics. As demonstrated empirically using a publicly available late-life depression dataset, supervised models frequently produce unreliable feature importance rankings despite achieving high target prediction accuracy, a discrepancy that has been systematically overlooked in the existing literature. This fundamental methodological conflation between predictive performance and explanatory validity represents a critical epistemological gap that this paper seeks to address.

Chen et al. investigated the relationships between smartphone addiction, use preferences, and depression among older adults in the digital context, employing a Random Forest model combined with SHAP analysis to assess and interpret feature importance [1]. Similarly, Yang et al. proposed the development and validation of machine learning models for predicting falls among hospitalized older adults, utilizing multiple supervised learning models alongside SHAP values to enhance feature interpretation and model transparency [2]. Both studies exemplify the increasing adoption of explainable artificial intelligence methods in clinical research while also illustrating the broader methodological vulnerability identified in this paper, namely, the uncritical reliance on SHAP as a validation tool without acknowledging its dependence on the underlying supervised model's unvalidated feature importances.

When a supervised model produces feature importances, these metrics reflect contributions to prediction rather than true causal associations. Consequently, high target prediction accuracy does not guarantee reliable feature importances, and the derived importances are inherently distorted, potentially leading to erroneous interpretations. Researchers must acknowledge that supervised models inherently possess two distinct types of accuracy: target prediction accuracy and feature importance accuracy. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance accuracy lacks a corresponding ground truth for validation, revealing a fundamental gap in feature assessment that may result in misleading conclusions. This distinction is critical, as conflating predictive performance with explanatory validity undermines the integrity of clinical and behavioral inferences drawn from such models. Furthermore, the function explainer = SHAP(model) demonstrates that SHAP is constrained to rely exclusively on a given supervised model whose feature importances remain unvalidated, thereby inherently inheriting and potentially amplifying the underlying biases embedded within those importances. This limitation raises serious concerns about the uncritical application of SHAP as an explainability tool, particularly in high-stakes domains such as healthcare, where misattributed feature relevance can have significant consequences for clinical decision-making and patient outcomes.

The results of this study provide compelling empirical support for these theoretical concerns. The analysis revealed striking performance differences between supervised and unsupervised feature selection approaches, with unsupervised models not only outperforming their supervised counterparts in cross-validation accuracy but also demonstrating remarkable stability across different validation scenarios. Both Feature Agglomeration and Highly Variable Gene Selection achieved superior accuracy scores with perfect consistency, maintaining identical performance across both five-fold cross-validation evaluations. Most significantly, the addition of SHAP values to supervised models completely failed to address their inherent instability problems. Random Forest with SHAP showed the same dramatic accuracy drop as standard Random Forest, while Gradient Boosting Machine with SHAP replicated the substantial performance decline of its non-SHAP counterpart. This conclusively demonstrates that SHAP integration did nothing to mitigate the instability issues inherent to supervised approaches, undermining the assumption that explainability tools can compensate for or correct label-driven biases in feature ranking.

These instability patterns observed in supervised models can be

understood through the lens of label-driven errors, whereby the optimization process oriented toward outcome prediction introduces systematic distortions in how features are ranked and weighted. In contrast, unsupervised models operate independently of outcome labels, thereby avoiding this source of bias and producing more stable and generalizable feature importance rankings. The Spearman correlation method further corroborated this pattern, demonstrating superior stability compared to all supervised approaches, reinforcing the broader principle that non-target prediction methods offer more reliable feature assessment under conditions of high dimensionality and complex variable interactions.

To address the absence of ground truth in feature importance validation, this paper proposes examining two critical criteria drawn from the extended Bradford Hill framework [3]: consistency across model specifications and dose–response relationships across data subsets [3–14]. The leave-top-feature-out diagnostic approach implemented here operationalizes these criteria by systematically removing the highest-ranked feature and evaluating the subsequent impact on feature ranking orders and predictive performance. In a theoretically sound feature importance framework, the removal of the top feature should not drastically reorganize the entire importance hierarchy if consistent predictive associations exist between features and the target variable. The observed sensitivity of supervised models to this perturbation, in contrast to the stability maintained by unsupervised approaches, further validates the methodological distinction proposed in this paper.

This dual-validation framework, combining cross-validation for predictive contribution assessment with the leave-top-feature-out approach for consistency and dose–response evaluation, directly addresses the epistemological gap that prior studies have failed to close. By treating consistency across model specifications and dose–response patterns as surrogate indicators of consistent predictive associations, rather than relying on target prediction accuracy alone, researchers can move toward more principled and defensible interpretations of feature importance in clinical machine learning research. Critically, no single algorithm yields definitive causal attributions, and meaningful inference requires triangulation across multiple algorithmic frameworks and data subsets.

The implications of these findings extend beyond the specific dataset and algorithms examined in this study. JMIR Aging has published a growing body of research utilizing machine learning for feature assessment, with substantial and increasing numbers of articles employing feature importance, feature selection, and SHAP methodologies across recent years. This significant trend reflects genuine and valuable interest in leveraging supervised models within psychiatric and geriatric research datasets. However, the systematic failure to validate feature importance through consistency and dose–response criteria means that a substantial portion of conclusions regarding key predictors in these studies may be unreliable, potentially misdirecting clinical attention and resource allocation. Addressing this gap requires not merely methodological refinement but a fundamental reconceptualization of what supervised machine learning models can and cannot validly claim to demonstrate.

This study is not without limitations. The empirical demonstration relies on a single publicly available dataset of relatively modest size, and future research should examine whether the patterns observed here generalize across larger, more diverse datasets and alternative clinical domains. Additionally, while the leave-top-feature-out approach provides a practical diagnostic tool, it represents one operationalization of the broader consistency and dose–response framework, and further methodological development is warranted to refine and extend these validation criteria. Nevertheless, the convergence of theoretical reasoning and empirical evidence presented here establishes a more rigorous methodological standard for feature importance assessment in psychiatric machine learning research, one that moves decisively beyond prediction performance as a proxy for interpretability and toward principled validation of associative claims.

While this paper examines feature consistency and dose–response relationships via the LTFO, other feature properties—such as correlations, multicollinearity, and redundant predictors—should also be evaluated using suitable diagnostic tools. For instance, when diagnosing multicollinearity, nonlinear methods such as the Random Forest Redundancy Index (RFRI), which is based on nonlinear Random Forest, may be more appropriate than linear approaches such as the Variance Inflation Factor (VIF), which is based on linear OLS. Researchers must acknowledge that violating underlying model assumptions can potentially distort all metric scores; this concern is particularly relevant when VIF—a method designed for linear relationships—is applied to nonlinear data. The outcomes from data analysis tools are reliable only when underlying assumptions are satisfied.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

References

[1] S.J.J. Jui, R.C. Deo, R. Acharya, P.D. Barua, J. Soar, A. Devi, Explainable stress detection system using hybrid CNN-Transformer and uncertainty quantification with copula GAN-based data, Biomed. Signal Process. Control 120 (Pt A) (2026) 109934, https://doi.org/10.1016/j.bspc.2026.109934.
[2] W. Zeng, L. Shan, Y. Chen, Y. Ma, S. Du, Interpretable deep learning framework for multi-label ECG classification: enhancing cardiac diagnostics through feature attention and explainability, Biomed. Signal Process. Control 112 (Pt C) (2026) 108447, https://doi.org/10.1016/j.bspc.2025.108447.
[3] A.B. Hill, The environment and disease: association or causation? Proc. R. Soc. Med. 58 (5) (1965) 295–300, https://doi.org/10.1177/003591576505800503.
[4] J.P. Ioannidis, Genetic associations: false or true? Trends Mol. Med. 9 (4) (2003) 135–138, https://doi.org/10.1016/S1471-4914(03)00030-3.
[5] J.P. Ioannidis, Why most discovered consistent predictive associations are inflated, Epidimiology 19 (5) (2008) 640–648, https://doi.org/10.1097/EDE.0b013e31818131e7.
[6] Q. Lai, R. Dannenfelser, J.P. Roussarie, V. Yao, Disentangling associations between complex traits and cell types with seismic, Nat. Commun. 16 (1) (2025) 8744, https://doi.org/10.1038/s41467-025-63753-z.
[7] D. Prada, B. Ritz, A.Z. Bauer, A.A. Baccarelli, Evaluation of the evidence on acetaminophen use and neurodevelopmental disorders using the Navigation Guide methodology, Environmental Health : a Global Access Science Source 24 (1) (2025) 56, https://doi.org/10.1186/s12940-025-01208-0.
[8] V. Prasad, A.B. Jena, Prespecified falsification end points: can they validate true observational associations? JAMA 309 (3) (2013) 241–242, https://doi.org/10.1001/jama.2012.96867.
[9] M.R. Roberts, S. Ashrafzadeh, M.M. Asgari, Research Techniques made simple: interpreting measures of Association in Clinical Research, J. Invest. Dermatol. 139 (3) (2019) 502–511.e1, https://doi.org/10.1016/j.jid.2018.12.023.
[10] A. Singh, L. Southam, K. Hatzikotoulas, N.W. Rayner, K. Suzuki, H.J. Taylor, et al., Correcting for genomic inflation leads to loss of power in large-scale genome-wide association study meta-analysis, Genet. Epidemiol. 49 (6) (2025) e70016, https://doi.org/10.1002/gepi.70016.
[11] E. Stamatakis, M. Ahmadi, R.K. Biswas, B. Del Pozo Cruz, C. Thøgersen-Ntoumani, M.H. Murphy, et al., Device-measured vigorous intermittent lifestyle physical activity (VILPA) and major adverse cardiovascular events: evidence of sex differences, Br. J. Sports Med. 59 (5) (2025) 316–324, https://doi.org/10.1136/bjsports-2024-108484.
[12] Y. Takefuji, Model-specific feature importances: distinguishing consistent predictive associations from target-feature relationships, J. Affect. Disord. 369 (2025) 390–391, https://doi.org/10.1016/j.jad.2024.10.019.
[13] Underwood, B. R., Lourida, I., Gong, J., Tamburin, S., Tang, E. Y. H., Sidhom, E., et al, & Deep Dementia Phenotyping (DEMON) Network (2025). Data-driven discovery of associations between prescribed drugs and dementia risk: A systematic

review. Alzheimer's & dementia (New York, N. Y.), 11(1), e70037. https://doi.org/10.1002/trc2.70037.
[14] M. Ye, Y. He, Y. Xia, Z. Zhong, X. Kong, Y. Zhou, 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. 24 (1) (2024) 460, https://doi.org/10.1186/s12876-024-03547-7.
[15] J. Natzkova, Late life depression (Version 2), Mendeley Data (2022), https://doi.org/10.17632/bsd9rgkr8c.2.
[16] GitHub. (2025). shapanalysis2.py. https://github.com/y-takefuji/late-life-depression.