Asian Journal of Psychiatry 114 (2025) 104762

Contents lists available at ScienceDirect

Asian Journal of Psychiatry

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

The stability paradox: Why high prediction accuracy does not guarantee reliable feature importance in psychiatric research

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

ARTICLE INFO

Keywords:
Feature selection stability
Supervised learning
Unsupervised learning
Psychiatric research
Machine learning reliability

ABSTRACT

This study examines the critical disconnect between prediction accuracy and feature importance reliability in machine learning applications for psychiatric research. Using a hikikomori dataset with 611 instances, we compared feature selection stability across supervised models (random forest, XGBoost, logistic regression), unsupervised methods (feature agglomeration, highly variable gene selection), and statistical approaches (Spearman correlation). Despite achieving highest classification accuracy (66.20 %), logistic regression exhibited significant instability in feature rankings when the top feature was removed. In contrast, unsupervised methods and statistical approaches demonstrated perfect stability in feature ranking orders. Our findings reveal that supervised models suffer from label-driven instability while unsupervised methods provide more consistent feature importance assessments, suggesting that psychiatric researchers should supplement high-accuracy supervised models with unsupervised approaches for reliable feature interpretation.

1. Introduction

Asian Journal of Psychiatry has published 168 articles on linear regression (18 in 2025), 104 articles on machine learning (29 in 2025), 124 articles on artificial intelligence (39 in 2025), 24 articles on random forest (9 in 2025), 427 articles on logistic regression (26 in 2025), and 18 articles on feature selection (5 in 2025), and 237 articles on dataset(s) (58 in 2025), indicating a rapid growing interest in data analysis with machine learning using datasets. However, due to lacking understanding of fundamental principles of supervised machine learning, researchers are not familiar with algorithm-induced errors while they are experts in their own domains.

Zhu et al. (2025) investigated factors associated with suicidal ideation (SI) in daily life among young people with mood disorders at risk of suicide, using ecological momentary assessment (EMA) to capture real-time contextual data. Employing multilevel logistic regression, they assessed concurrent, time-lagged, and adjusted associations between SI and environmental, interpersonal, and emotional factors.

However, researchers including Zhu et al. sometimes infer that high target-prediction performance (for example, a high R^2 or classification accuracy) ensures the reliability of model outputs such as coefficient estimates, odds ratios, confidence intervals and p-values. In supervised learning, models like logistic regression involve two distinct notions of performance: target prediction accuracy and feature-importance accuracy. While target prediction accuracy can be evaluated against ground-truth labels, feature importance lacks a direct ground truth for accuracy validation and reflects contributions to prediction rather than causal or "true" associations. Consequently, strong predictive performance does not guarantee that feature importances or the resulting substantive interpretations are reliable (Parr et al., 2024; Watson and Wright, 2021; Molnar et al., 2022; Lipton, 2018; Fisher et al., 2019; Lenhof et al., 2024; Mandler and Weigand, 2024; Potharlanka, Bhat, 2024; Wood et al., 2024).

Zhu et al. should also acknowledge that the credibility of inferential outputs depends on meeting model assumptions. Violations can bias estimates and distort inference, leading to erroneous conclusions. In particular, applying linear methods to fundamentally nonlinear relationships, or using parametric and semiparametric models such as logistic regression when their assumptions (for example, correct link function, linearity in the logit, independence, absence of severe multicollinearity, appropriate handling of time dependence and clustering, and well-specified random effects) are not satisfied, can compromise the validity of feature importance, odds ratios, confidence intervals, and p-values (Dey et al., 2025; Pinheiro-Guedes et al., 2024; Wang et al., 2023; Osborne, 2015; van Maanen et al., 2019; Work et al., 1989; Zulfadhli et al., 2024; Akturk et al., 2025; Rifada et al., 2022; Suliyanto et al., 2020; Wibowo et al., 2021; Steyerberg et al., 2011; Özkale, 2016). Careful diagnostics, sensitivity analyses, nonlinear or nonparametric

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.ajp.2025.104762
Received 9 October 2025; Received in revised form 31 October 2025; Accepted 3 November 2025
Available online 4 November 2025

alternatives, and robustness checks are essential to support interpretability alongside prediction.

Due to the absence of ground truth in feature importance calculations, this paper advocates for multifaceted approaches incorporating unsupervised models such as feature agglomeration (FA) and highly variable gene selection (HVGS), and followed by non-target-prediction nonlinear nonparametric methods such as Spearman's correlation with p-values instead of solely relying on parametric logistic regression. Due to model specific nature, supervised models like logistic regression must suffer from instability in feature ranking orders due to label-driven errors while FA, HVGS and Spearman exhibit stronger stability in feature ranking orders due to the absence of label-driven errors.

This paper examines the effectiveness of feature selection across supervised models, unsupervised models, and non-target-prediction methods using a public hikikomori dataset (Stavropoulos et al., 2019).

2. Methods

The dataset comprises 611 instances and 7 features, which has been converted into a binary classification problem (severe or not severe) (Stavropoulos et al., 2019). Our methodology involves a systematic evaluation of feature importance through multiple steps: first, assessing the top feature ranking orders from the complete feature set using diverse algorithms; then removing the highest-ranked feature to create a reduced dataset; and finally, re-evaluating the feature ranking orders in this reduced context. For feature selection and importance analysis, we employ a comprehensive suite of algorithms including random forest, XGBoost, logistic regression, feature agglomeration (FA), highly variable gene selection (HVGS), and Spearman's correlation. This approach allows us to identify the most influential predictors and understand how feature importance shifts when prominent features are excluded.

3. Results

Table 1 demonstrates the cross-validation accuracy and feature ranking order for six different models across both full and reduced feature sets. In the full feature set, logistic regression achieved the highest cross-validation mean accuracy of 0.6620, while XGBoost followed at 0.5908, and random forest obtained 0.5857. Four of the six models (random forest, FA, HVGS, and Spearman) identified IGD_Total as the most important feature in the full dataset. When IGD_Total was removed to create the reduced feature set, logistic regression maintained superior performance with a slightly improved accuracy of 0.6672, while all other models showed decreased performance. In the reduced dataset, feature importance varied considerably among models. Notably, the supervised models (random forest, XGBoost, and logistic regression) demonstrated substantial instability in feature ranking orders between the full and reduced sets, with significant reshuffling of feature importance after removing the top feature. In contrast, the unsupervised methods (FA and HVGS) and statistical approach (Spearman correlation) exhibited perfect stability in feature ranking orders, maintaining the exact same relative importance of features in the reduced set as they had in the full set, with features simply moving up one position after the removal of IGD_Total. This finding suggests that unsupervised feature selection methods may provide more consistent feature importance assessments when working with modified feature sets.

Table 1
cross-validation accuracy and feature ranking order.

model	CV mean (full set) Top 5 feature ranking order	CV mean (reduced set) Top 5 feature ranking order
random forest	0.5857 IGD_Total,Age_New,Hours, Gender_New,Living_condition	0.5094 Age_New,Hours,Gender_New, Living_condition,Country
XGBoost	0.5908 IGD_Total,Country, Living_condition,Age_New, Hours	0.5179 Country,Hours, Living_condition,Age_New, Gender_New
logistic regression	0.6620 Country,Living_condition, Gender_New,Hours,IGD_Total	0.6672 Gender_New,Hours,IGD_Total, Living_condition,Age_New
FA	0.5772 IGD_Total,Age_New,Hours, Gender_New,Country	0.5094 Age_New,Hours,Gender_New, Country,Living_condition
HVGS	0.5840 IGD_Total,Age_New,Hours, Gender_New,Living_condition	0.5026 Age_New,Hours,Gender_New, Living_condition,Country
Spearman	0.5857 IGD_Total,Hours,Country, Age_New,Gender_New	0.5094 h,Country,Age_New, Gender_New,Living_condition

For purposes of reproducibility and transparency, Python code, hikikomori.py is publicly available at GitHub (GitHub, 2025).

4. Discussion

Our investigation into feature selection methodologies reveals a crucial finding: high target prediction accuracy does not guarantee reliable feature importance rankings. This dichotomy is clearly demonstrated in our results, where logistic regression achieved the highest classification accuracy (0.6620 in the full dataset and 0.6672 in the reduced dataset) but exhibited significant instability in feature importance rankings when the dataset was modified by removing the top feature.

The supervised models (random forest, XGBoost, and logistic regression) all demonstrated considerable instability in their feature ranking orders between the full and reduced datasets. This instability stems from the label-driven nature of supervised learning, where the algorithms optimize for prediction accuracy based on the target variable rather than focusing on intrinsic data structure. When IGD_Total was removed, these models substantially reshuffled their feature importance assessments, suggesting that their feature rankings are highly dependent on the specific dataset configuration rather than reflecting stable underlying relationships.

In stark contrast, unsupervised methods (FA and HVGS) and the statistical approach (Spearman correlation) exhibited perfect stability in feature ranking orders. When IGD_Total was removed, the remaining features maintained precisely the same relative importance, simply moving up one position in the ranking. This remarkable stability can be attributed to the absence of label-driven errors in these methods, as they derive feature importance based on intrinsic data characteristics rather than prediction optimization for a specific target variable.

This finding has significant implications for psychiatric research relying on machine learning methods. Researchers often prioritize models with the highest predictive accuracy, potentially overlooking the reliability of feature importance interpretations. Our results suggest that while supervised models may excel at prediction tasks, their feature importance rankings should be interpreted with caution, particularly when making inferences about the relative significance of different factors in psychiatric conditions.

The stability exhibited by unsupervised methods and Spearman correlation highlights their value in providing consistent assessments of feature importance. These approaches are less susceptible to dataset-specific fluctuations and may offer more reliable insights into the underlying structure of psychiatric data. Therefore, we recommend incorporating these methods alongside supervised approaches when the research objective includes understanding the relative importance of different factors, rather than solely focusing on prediction accuracy.

The hikikomori dataset is the largest publicly available dataset in this domain. We replicated our analysis on MNIST (70,000 samples, 768 features) and a breast cancer omics dataset (705 samples, 1936 features). Supervised models showed unstable feature-importance rankings under leave-one-out approaches, whereas label-agnostic methods (FA, HVGS, Spearman) were more stable. Complex methods did not ensure higher prediction accuracy or stability; simpler approaches often

matched or exceeded them. We quantified ranking stability using Spearman's rank correlation and Kendall's Tau, and applied a leave-one-out stability test: select top n features from the full set (set1), remove the highest-ranked feature to form a reduced dataset, re-select top n-1 features from the reduced dataset (set2), and compare ranking orders between set1 and set2. Removing the highest feature yields the strongest stability stress test. This approach satisfies consistency and dose-response criteria for true association assessment. In our setting, Spearman outperformed Kendall for capturing stability. These results support the "stability paradox" across datasets. We will conduct a larger, preregistered multi-dataset study with standardized protocols and publicly release code and results.

Additionally, our findings reinforce the importance of conducting sensitivity analyses when interpreting feature importance in psychiatric research. The substantial reordering of features in supervised models after removing just one feature suggests that reported feature importances from these models may be highly contingent on specific dataset characteristics rather than reflecting generalizable relationships.

The application of machine learning in psychiatry, particularly in understanding schizophrenia, presents both significant opportunities and notable challenges. Tandon and Tandon (2019a) emphasize the critical need for standards and guidelines to ensure the responsible implementation of machine learning methods in psychiatric research and practice. While these approaches offer potential breakthroughs in untangling the complex heterogeneity of schizophrenia, researchers must navigate between realistic expectations and unwarranted hype (Tandon and Tandon, 2019b). The prospect of using machine learning to "cut the Gordian knot of schizophrenia" (Tandon and Tandon, 2018) is compelling, yet requires careful methodological considerations including appropriate validation, replication, and transparency in research design. A particular concern is the balance between sophisticated computational approaches and clinical utility, as machine learning models must ultimately translate into meaningful improvements in patient care rather than merely demonstrating statistical significance. These authors consistently advocate for interdisciplinary collaboration between data scientists, clinicians, and patients to develop machine learning applications that are both technically sound and clinically relevant to address the complex challenges in psychiatric disorders.

In conclusion, while supervised learning approaches like logistic regression may offer superior predictive performance, researchers should be mindful that this does not equate to reliable feature importance assessments. For robust interpretations of feature importance in psychiatric research, we recommend supplementing supervised models with unsupervised methods and statistical approaches that demonstrate greater stability in their feature rankings, thus providing a more comprehensive and reliable understanding of the factors influencing psychiatric conditions.

CRediT authorship contribution statement

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

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Code availability

Not applicable.

Funding

This research has no fund.

Declaration of Generative AI and AI-assisted technologies in the writing process

Not applicable.

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.

Acknowledgments

Not applicable.

Data availability

Not applicable.

References

Akturk, B., Beyaztas, U., Shang, H.L., et al., 2025. Robust functional logistic regression. Adv. Data Anal. Classif. 19, 121–145. https://doi.org/10.1007/s11634-023-00577-z.

Dey, D., Haque, M.S., Islam, M.M., Aishi, U.I., Shammy, S.S., Mayen, M.S.A., et al., 2025. The proper application of logistic regression model in complex survey data: a systematic review. BMC Med. Res. Methodol. 25, 15. https://doi.org/10.1186/s12874-024-02454-5.

Fisher, A., Rudin, C., Dominici, F., 2019. 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, 177.

GitHub. hikikomori,py. ⟨https://github.com/y-takefuji/hikikomori⟩.

Lenhof, K., Eckhart, L., Rolli, L.M., Lenhof, H.P., 2024. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. Brief. Bioinforma. 25 (5), bbae379. https://doi.org/10.1093/bib/bbae379.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16 (3), 31–57. https://doi.org/10.1145/3236386.3241340.

Mandler, H., Weigand, B., 2024. A review and benchmark of feature importance methods for neural networks. ACM Comput. Surv. 56 (12), 318. https://doi.org/10.1145/3679012.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., et al., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4.

Osborne, J., 2015. A practical guide to testing assumptions and cleaning data for logistic regression. A practical guide to testing assumptions and cleaning data for logistic regression. SAGE Publications, Ltd, pp. 84–130. https://doi.org/10.4135/9781483399041.n4.

Özkale, M.R., 2016. Iterative algorithms of biased estimation methods in binary logistic regression. Stat. Pap. 57, 991–1016. https://doi.org/10.1007/s00362-016-0780-9.

Parr, T., Hamrick, J., Wilson, J.D., 2024. Nonparametric feature impact and importance. Inf. Sci. 653, 119563. https://doi.org/10.1016/j.ins.2023.119563.

Pinheiro-Guedes, L., Martinho, C., Martins, M.R.O., 2024. Logistic regression: limitations in the estimation of measures of association with binary health outcomes. Acta Med. Port. 37 (10), 697–705. https://doi.org/10.20344/amp.21435.

Potharlanka, J.L., Bhat, M., N., 2024. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci. Rep. 14 (1), 2923. https://doi.org/10.1038/s41598-024-53141-w.

Rifada, M., Chamidah, N., Ningrum, R.A., 2022. Estimation of nonparametric ordinal logistic regression model using generalized additive models (GAM) method based on local scoring algorithm. AIP Conf. Proc. 2668 (1), 070013. https://doi.org/10.1063/5.0111771.

Stavropoulos, V., Anderson, E.E., Beard, C., Latifi, M.Q., Kuss, D., Griffiths, M., 2019. A preliminary cross-cultural study of hikikomori and Internet gaming disorder: the moderating effects of game-playing time and living with parents. Addict. Behav. Rep. 9, 100137. https://doi.org/10.1016/j.abrep.2018.10.001.

Steyerberg, E.W., Schemper, M., Harrell, F.E., 2011. Logistic regression modeling and the number of events per variable: selection bias dominates. J. Clin. Epidemiol. 64 (12), 1464. https://doi.org/10.1016/j.jclinepi.2011.06.016.

Suliyanto, Rifada, M., Tjahjono, E., 2020. Estimation of nonparametric binary logistic regression model with local likelihood logit estimation method (case study of diabetes mellitus patients at Surabaya Hajj General Hospital). AIP Conf. Proc. 2264 (1), 030007. https://doi.org/10.1063/5.0025807.

Tandon, N., Tandon, R., 2018. Will machine learning enable us to finally cut the Gordian knot of schizophrenia? Schizophr. Bull. 44 (5), 939–941. https://doi.org/10.1093/schbul/sby101.

Tandon, N., Tandon, R., 2019b. Using machine learning to explain the heterogeneity of schizophrenia: realizing the promise and avoiding the hype. Schizophr. Res. 214, 70–75. https://doi.org/10.1016/j.schres.2019.08.032.

Tandon, N., Tandon, R., 2019a. Machine learning in psychiatry: standards and guidelines. Asian J. Psychiatry 44, A1–A4. https://doi.org/10.1016/j.ajp.2019.09.009.

van Maanen, L., Katsimpokis, D., van Campen, A.D., 2019. Fast and slow errors: logistic regression to identify patterns in accuracy–response time relationships. Behav. Res. Methods 51, 2378–2389. https://doi.org/10.3758/s13428-018-1110-z.

Wang, T., Tang, W., Lin, Y., Su, W., 2023. Semi-supervised inference for nonparametric logistic regression. Stat. Med. 42 (15), 2573–2589. https://doi.org/10.1002/sim.9737.

Watson, D.S., Wright, M.N., 2021. Testing conditional independence in supervised learning algorithms. Mach. Learn. 110 (8), 2107–2129. https://doi.org/10.1007/s10994-021-06030-6.

Wibowo, W., Amelia, R., Octavia, F.A., Wilantari, R.N., 2021. Classification using nonparametric logistic regression for predicting working status. AIP Conf. Proc. 2329 (1), 060032. https://doi.org/10.1063/5.0043598.

Wood, D., Papamarkou, T., Benatan, M., et al., 2024. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. Data Min. Knowl. Discov. 38, 4184–4216. https://doi.org/10.1007/s10618-024-01070-7.

Work, J.W., Ferguson, J.G., Diamond, G.A., 1989. Limitations of a conventional logistic regression model based on left ventricular ejection fraction in predicting coronary events after myocardial infarction. Am. J. Cardiol. 64 (12), 702–707. https://doi.org/10.1016/0002-9149(89)90751-0.

Zhu, J., Hou, X., Tao, H., Lin, K., Zhou, L., Niu, L., 2025. What triggers suicidal ideation in daily life? A real-time study among young people with mood disorders at risk of suicide. Asian J. Psychiatry, 104718. https://doi.org/10.1016/j.ajp.2025.104718.

Zulfadhli, M., Budiantara, I.N., Ratnasari, V., 2024. Nonparametric regression estimator of multivariable Fourier Series for categorical data. MethodsX 13, 102983. https://doi.org/10.1016/j.mex.2024.102983.