Contents lists available at ScienceDirect

Brain Behavior and Immunity

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

Letter to the Editor

Reevaluating feature importances in machine learning models for schizophrenia and bipolar disorder: The need for true associations

ARTICLE INFO

Keywords:
Machine learning
Feature importance
Biases
Schizophrenia
Bipolar disorder
Statistical methods
True associations

ABSTRACT

Skorobogatov et al. developed supervised machine learning models to predict diagnoses and illness states in schizophrenia and bipolar disorder. However, their reliance on bootstrap forests and generalized regressions introduces significant biases in feature importance assessments. This paper highlights the critical distinction between feature importances generated by machine learning and actual associations, which are often model-specific and context-dependent. We underscore the limitations of biased feature importances and advocate for the use of robust statistical methods, such as Chi-squared tests and Spearman's correlation, to reveal true associations. Reassessing findings with these methods will enable more accurate interpretations and reinforce the importance of understanding the limitations inherent in machine learning methodologies.

Skorobogatov et al. developed supervised machine learning models to predict diagnoses and illness states in schizophrenia and bipolar disorder (Skorobogatov et al., 2024). While they employed bootstrap forests and generalized regressions for feature selection, their methods exhibit significant inherent biases associated with machine learning models (Barton-Henry et al., 2021; Chen et al., 2023; Ma et al., 2024; Watanabe et al., 2021), leading to incorrect conclusions. It is crucial for researchers, including Skorobogatov et al., to recognize the distinction between feature importances derived from machine learning and actual associations. The nature of these feature importances is model-specific, meaning that different models yield varying results, while true associations can be accurately determined through robust statistical methods such as Chi-squared tests and Spearman's correlation, both accompanied by p-values (Murakami et al., 2024; Yaseen et al., 2023; Carter et al., 2023).

This paper underscores the limitations posed by biased feature importances generated through machine learning and emphasizes the necessity of relying on genuine associations for drawing reliable conclusions. Consequently, it is essential for Skorobogatov et al. to reevaluate their findings using true associations, which will facilitate a more accurate interpretation of their results. Importantly, this paper does not aim to discredit machine learning; rather, it acknowledges that while machine learning primarily seeks to predict outcomes accurately, feature importances are intended to represent associations between the target variable and features. However, these associations can be skewed by model-specific biases. This paper shows why machine learning models generate biases feature importances and advocates for the use of true associations between the target and features.

Machine learning models, including techniques such as bootstrap forests and generalized regressions, can induce biases in feature importance for several reasons (Barton-Henry et al., 2021; Chen et al., 2023; Ma et al., 2024; Watanabe et al., 2021).

First, the model-specific nature of different machine learning algorithms contributes to variability in feature importance assessments. Each model interprets data in its own way, and feature importance is calculated based on how each model utilizes individual features to make predictions. Consequently, the importance assigned to features can vary significantly across modeling techniques, meaning that feature importances are context-dependent and not universally applicable.

Overfitting is another significant factor. Many machine learning models, especially complex ones, can overfit to the training data, capturing noise rather than the underlying signal. When overfitting occurs, the derived feature importances may reflect spurious relationships that do not generalize to new data, leading to misleading conclusions about true associations. This highlights the inherent risks of relying solely on machine learning for feature significance without additional validation.

Moreover, machine learning models often assess the strength and significance of features based on correlations rather than causal relationships. As a result, features that are correlated with the response variable may be assigned high importance even if they do not have a genuine causal impact. This limitation underscores the need to distinguish between correlation and causation in feature importance evaluations.

Feature interactions also complicate the assessment of importance. Many models fail to account adequately for interactions between features, leading to inflated or deflated importance scores based on how features interact within the model. For instance, in bootstrap forests, the importance of a particular feature may vary based on its interaction with other features, combined with the randomness inherent in the bootstrapping process. Data imbalance can further skew feature importances. In scenarios where classes are imbalanced, machine learning models can produce biased feature importances. Features that are more prevalent in the majority class may be overrepresented, overshadowing those that could be more critical for minority classes, leading to an inaccurate assessment of their importance.

Another consideration is data preprocessing. The methods used for preprocessing—such as normalization, scaling, or encoding categorical

https://doi.org/10.1016/j.bbi.2024.11.036
Received 30 October 2024; Accepted 28 November 2024
Available online 29 November 2024
0889-1591/© 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

variables—can influence the importance assigned to various features. Inconsistent preprocessing techniques may result in different importance values across models or runs. Additionally, feature importances can lack robustness. They can be sensitive to minor variations in the dataset, such as outliers or noise. A model might assign drastically different importance scores if the training data is slightly altered, raising concerns about the stability and reliability of those features.

Finally, randomness in ensemble methods like bootstrap forests can affect feature importance calculations. In these methods, importance is often based on the degree to which removing specific features decreases model accuracy. However, this evaluation can be heavily influenced by the random sampling of training data, leading to inconsistent and potentially biased feature importance scores.

In summary, while bootstrap forests and generalized regressions are powerful tools for prediction, researchers must approach the interpretation of feature importances with caution. Relying solely on these values without an understanding of their inherent biases can result in incorrect conclusions regarding the relationships between features and outcome variables. To accurately identify true associations, it is crucial to complement machine learning approaches with robust statistical methodologies (Murakami et al., 2024; Yaseen et al., 2023; Carter et al., 2023).

Funding

This research has no fund.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Code availability

Not applicable.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

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

References

Barton-Henry, K., Wenz, L., Levermann, A., 2021. Decay radius of climate decision for solar panels in the city of Fresno, USA. Sci. Reports 11 (1), 8571. https://doi.org/10.1038/s41598-021-87714-w.

Carter, K.A., Fischer, M.D., Petrova, M.I., Balkus, J.E., 2023. Epidemiologic evidence on the role of lactobacillus iners in sexually transmitted infections and bacterial vaginosis: a series of systematic reviews and meta-analyses. Sexually Transmitted Diseases 50 (4), 224–235. https://doi.org/10.1097/OLQ.0000000000001744.

Chen, J., Ooi, L.Q.R., Tan, T.W.K., Zhang, S., Li, J., Asplund, C.L., Eickhoff, S.B., Bzdok, D., Holmes, A.J., Yeo, B.T.T., 2023. Relationship between prediction accuracy and feature importance reliability: an empirical and theoretical study. NeuroImage 274, 120115. https://doi.org/10.1016/j.neuroimage.2023.120115.

Ma, W., Wu, H., Chen, Y., Xu, H., Jiang, J., Du, B., Wan, M., Ma, X., Chen, X., Lin, L., Su, X., Bao, X., Shen, Y., Xu, N., Ruan, J., Jiang, H., Ding, Y., 2024. New techniques to identify the tissue of origin for cancer of unknown primary in the era of precision medicine: progress and challenges. Briefings Bioinform. 25 (2), bbae028. https://doi.org/10.1093/bib/bbae028.

Murakami, K., Shinozaki, N., Okuhara, T., McCaffrey, T.A., Livingstone, M.B.E., 2024. Prevalence and correlates of dietary and nutrition information seeking through various web-based and offline media sources among Japanese adults: web-based cross-sectional study. JMIR Public Health Surveillance 10, e54805.

Skorobogatov, K., De Picker, L., Wu, C.-L., Foiselle, M., Richard, J.-R., Boukouaci, W., Bouassida, J., Laukens, K., Meysman, P., le Corvoisier, P., Barau, C., Morrens, M., Tamouza, R., Leboyer, M., 2024. Immune-based machine learning prediction of diagnosis and illness state in schizophrenia and bipolar disorder. Brain, Behavior, Immunity 122, 422–432. https://doi.org/10.1016/j.bbi.2024.08.013.

Watanabe, E., Noyama, S., Kiyono, K., Inoue, H., Atarashi, H., Okumura, K., Yamashita, T., Lip, G.Y.H., Kodani, E., Origasa, H., 2021. Comparison among random forest, logistic regression, and existing clinical risk scores for predicting outcomes in patients with atrial fibrillation: a report from the J-RHYTHM registry. Clin. Cardiol. 44 (9), 1305–1315. https://doi.org/10.1002/clc.23688.

Yaseen, N.R., Barnes, C.L.K., Sun, L., Takeda, A., Rice, J.P., 2023. Genetics of vegetarianism: a genome-wide association study. PloS One 18 (10). https://doi.org/10.1371/journal.pone.0291305.

Yoshiyasu Takefuji
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku,
Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.