Contents lists available at ScienceDirect

International Journal of Cardiology

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

Enhancing heart disease feature analysis with spearman's correlation with *p*-values

ARTICLE INFO

Keywords:
Heart disease prediction
Machine learning
Feature importance
Statistical validation
Causation vs. correlation
Spearman's correlation

Zhang et al. enhanced heart disease prediction using machine learning models (XGB, RFC, DTC, KNNC, LRC) [1]. They evaluated the feature importance of each model through cross-validation. However, machine learning target predictions can be validated against known ground truth values to assess accuracy, while feature importances derived from models lack such definitive references for validation [2].

The lack of ground truth values in feature importances derived from machine learning models results in each model employing different methodologies for these importances. Consequently, each model generates unique feature importances, potentially reflecting inherent biases. More than 100 peer-reviewed articles have identified significant biases associated with feature importances in machine learning models.

While there are various bias mitigation methods available, none can entirely eliminate biases in feature importance assessments. This paper advocates for employing bias-free nonlinear and nonparametric robust statistical techniques, such as Spearman's correlation and Kendall's tau, both of which are accompanied by *p*-values. Unlike traditional feature importance metrics, which typically range from 0 to 1 and may reflect biased strengths, these statistical methods provide a range from −1 to 1, indicating both the strength and the direction of the relationships between variables. By utilizing Spearman's and Kendall's methods, researchers can gain more nuanced insights into the associations between features and the target variable, enhancing the interpretability and reliability of their analyses.

Although cross-validation offers insights into feature importance, these values should be interpreted with caution, as they can differ between models and data, and do not necessarily establish causal relationships. Cross-validation that involves splitting and shuffling data without altering the model is effective for assessing target accuracy, but it is not effective for accurately evaluating feature importance due to the absence of ground truth values.

Cross-validation is a technique that evaluates a model's accuracy by repeatedly splitting and shuffling the data into training and testing subsets without altering the model itself. Cross-validation is data specific nature. This approach helps in understanding how well the model generalizes to unseen data while maintaining its integrity without altering the model itself. Feature importance indicates the contribution of each feature to a model's predictions, helping identify the most influential variables.

The strength of statistical methods lies in their ability to provide interpretable insights grounded in well-defined principles. Specifically, statistical methods can yield better correlations due to three critical elements. First, traditional statistical approaches often operate under specific assumptions about data distributions, which can enhance their reliability in modeling relationships between variables. Second, these methods emphasize the examination of relationships between variables, allowing for a clearer interpretation of how independent variables impact a dependent variable. Third, the use of *p*-values for hypothesis testing provides a framework for validating findings, ensuring that observed relationships are statistically significant rather than due to chance. In contrast, while machine learning excels at handling nonlinear relationships, it can sometimes provide insights that lack transparency, leading to the "black box" criticism. As statistical methods continue to evolve, we may see a convergence where both approaches complement each other, leveraging their strengths to improve our understanding of complex data relationships.

Due to the lack of ground truth values in feature importance calculations, robust statistical methods such as Spearman's correlation and Kendall's tau—both of which yield accompanying *p*-values—are grounded in three essential components: data distribution, the statistical relationships between variables, and the evaluation of statistical significance through *p*-values. These methods facilitate accurate assessments of associations between the target variable and features, ensuring a reliable analysis. Although primarily employed for exploring features related to heart disease, these techniques are versatile and applicable across diverse feature analyses in various fields.

This paper recommends using statistical methods such as Spearman's correlation with *p*-values or Kendall's tau with p-values to verify true feature-outcome relationships. Machine learning algorithms can identify statistical correlations between variables, but these correlations do not inherently prove a causal relationship. A correlation suggests variables move together, but does not explain why or how they are related [3].

Spearman's correlation assesses the strength and direction of a

https://doi.org/10.1016/j.ijcard.2025.133207
Received 19 December 2024; Received in revised form 24 January 2025; Accepted 25 March 2025
Available online 30 March 2025

monotonic relationship between two ranked variables. It is calculated by determining the differences between the ranked values, then applying the formula:

$$\rho = 1 - \left(6^{*} \sum \mathrm{di}^{2}\right) \Big/ \left(\mathrm{n}^{*}\left(\mathrm{n}^{2}-1\right)\right).$$

where di is the difference between ranks and n is the number of observations. The associated *p*-value indicates the significance of the correlation, helping to determine if the observed relationship is statistically valid. The range of Spearman's correlation is from −1 to 1.

P-values represent the probability of obtaining results as extreme as those observed, assuming the null hypothesis is true. A low *p*-value (typically $\leq$0.05) suggests strong evidence against the null hypothesis, leading to its rejection, while a high p-value indicates weak evidence, signifying insufficient data to reject the null hypothesis. When associated with Spearman's correlation, *p*-values provide a measure of statistical significance, ensuring that identified relationships are not merely coincidental. To effectively incorporate Spearman's correlation into the analytical framework, one should calculate both the Spearman correlation coefficient and its corresponding *p*-value. A low p-value indicates that the observed correlation is statistically significant, confirming a meaningful association between the variables. This approach enhances the validity and reliability of the findings, allowing researchers to distinguish between genuine relationships and random correlations.

Moreover, although combining top-performing models with optimization algorithms may improve prediction accuracy, complex models may also capture misleading non-linear relationships that simpler models might overlook. [4]

Several factors hinder genuine associations in machine learning feature selection. Firstly, while machine learning algorithms typically identify correlations between features and the target variable, correlation does not necessarily mean one variable cause another. Secondly, feature importance is model-specific; different models like decision trees or random forests may assign varying importance scores to the same features.

To enhance the accuracy of heart disease research analysis, a more effective approach is to use reliable statistical methods such as Spearman's correlation with *p*-values, which can accurately reveal the true associations between targets and features [5]. These methods are based on hypothesis testing and can determine whether observed associations are statistically significant or coincidental. Therefore, to strengthen their conclusions, Spearman's correlation should be incorporated into their analyses to achieve more accurate and effective results.

The SciPy library currently includes a function for calculating Spearman's correlation, along with its associated *p*-value. This functionality enhances our ability to assess both the strength and significance of ranked relationships between variables effectively. Calculate correlation scores and sort them for feature selection:

$$\mathrm{spearman\_corr}, \mathrm{p\_value} = \mathrm{scipy.stats.spearmanr}(\mathrm{variable\_A}, \mathrm{variable\_B}).$$

To obtain Python code for calculating Spearman's correlation along with *p*-values, you can use the following query with generative AI: "Assume that data.csv contains the target variable (y) and independent variables (x1, x2,…, xn). Provide Python code to compute and sort Spearman's correlation coefficients with their corresponding p-values."

The design of a study plays a critical role in determining the ability to infer causal relationships. While machine learning and statistical analysis methods have their limitations regarding causality, it is essential to recognize that observational and cross-sectional studies, as well as diagnostic experiments, inherently lack the capacity to establish causal inferences due to their design. In contrast, cohort studies are more suited for this purpose, as they allow for a temporal sequence to be established and can better account for confounding variables. Ultimately, a combination of robust study design and appropriate analytical techniques is crucial for accurate causality analysis.

SHAP (SHapley Additive exPlanations) is a powerful tool for explainable AI, yet its reliance on the underlying model means that it can inherit and potentially amplify biases present in feature importance assessments. Given the function of `explain = SHAP(model)`, the biases of the model can distort the interpretation of feature contributions. In light of these limitations, this paper advocates for employing robust statistical methods such as Spearman's correlation and Kendall's tau. These methods provide a more reliable framework for understanding the relationships between features and target variables, free from the biases that may be introduced by the models themselves. By utilizing these robust techniques, researchers can achieve a clearer and more accurate understanding of feature importance.

To prevent confusing correlation with causation in machine learning analyses, it's essential to incorporate domain knowledge and causal inference methods. For example, one could use techniques such as Directed Acyclic Graphs (DAGs) to map out relationships between variables, helping to identify potential confounders. Additionally, conducting controlled experiments or utilizing methods like propensity score matching can provide insights into causal relationships. By emphasizing these approaches alongside statistical correlations, researchers can better distinguish between mere associations and true causal links in their analyses.

When Spearman's correlation coefficients approach −1 or 1, caution is necessary, particularly if outliers have the potential to distort the rankings. In these instances, visualizing the data through scatter plots or box plots can help assess the impact of outliers. To ascertain whether the relationship is genuinely monotonic or influenced by anomalies, researchers can employ robust regression analyses or nonparametric tests to examine consistency across different data subsets. This ensures that the identified correlations reflect authentic relationships rather than being skewed by outlier effects. Additionally, Kendall's tau is advantageous in mitigating the influence of outliers, providing a more stable measure of association.

According to Lakens [6], the sample size justification can be calculated using the following formula:

$$\mathrm{n} = (\mathrm{z}/\mathrm{M})^{2}{}_{*}\mathrm{p}_{*}(1-\mathrm{p}).$$

where (n) indicates the sample size, (z) represents the z-value, (M) is the margin of error, and (p) is the estimated proportion.

To find the z-value, you can use a standard normal distribution table (z-table) or a calculator. For example, with a 95 % confidence level, the cumulative area in the middle of the standard normal distribution is 0.95, leaving 0.05 (or 5 %) in the tails, split equally on both sides. Each tail has an area of 0.025 (or 2.5 %). To find the z-value corresponding to a cumulative area of 0.975 (since $0.95 + 0.025 = 0.975$), you can use a z-table or calculator. The z-value for a cumulative area of 0.975 is approximately 1.96. Therefore, the z-value for a 95 % confidence level is 1.96. The minimum required sample size for a 95 % confidence interval with a margin of error of 0.04, assuming an estimated proportion of 0.5, is approximately 601.

Authors' contribution

Haoqian Pan investigated and wrote this article, Yoshiyasu Takefuji supervised and wrote this article.

CRediT authorship contribution statement

Haoqian Pan: Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Yoshiyasu Takefuji:** Writing – review & editing, Writing – original draft, Validation, Conceptualization.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The author has no conflict of interest.

Data availability

Not applicable.

References

[1] H. Zhang, R. Mu, Refining heart disease prediction accuracy using hybrid machine learning techniques with novel metaheuristic algorithms, Int. J. Cardiol. 416 (2024) 132506, https://doi.org/10.1016/j.ijcard.2024.132506.

[2] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16), Association for Computing Machinery, 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[3] T.P. Pagano, R.B. Loureiro, F.V.N. Lisboa, R.M. Peixoto, G.A.S. Guimaraes, G.O. R. Cruz, M.M. Araujo, L.L. Santos, M.A.S. Cruz, E.L.S. Oliveira, I. Winkler, E.G. S. Nascimento, Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, Big Data Cognit. Comput. 7 (1) (2023) 15, https://doi.org/10.3390/bdcc7010015.

[4] J.W. Gichoya, K. Thomas, L.A. Celi, N. Safdar, I. Banerjee, J.D. Banja, L. Seyyed-Kalantari, H. Trivedi, S. Purkayastha, AI pitfalls and what not to do: mitigating bias in AI, Br. J. Radiol. 96 (1150) (2023) 20230023, https://doi.org/10.1259/bjr.20230023.

[5] J. Jiang, X. Zhang, Z. Yuan, Feature selection for classification with spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets, Expert Syst. Appl. 249 (Pt B) (2024) 123633, https://doi.org/10.1016/j.eswa.2024.123633.

[6] Lakens D. Collabra, Psychology 8 (2022) 1, https://doi.org/10.1525/collabra.33267.

Haoqian Pan, Yoshiyasu Takefuji*
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

* Corresponding author.
E-mail addresses: g2450009@stu.musashino-u.ac.jp (H. Pan), takefuji@keio.jp (Y. Takefuji).