Scripta Materialia 268 (2025) 116857

Contents lists available at ScienceDirect

Scripta Materialia

journal homepage: www.journals.elsevier.com/scripta-materialia

Comments on "Toward prediction and insight of porosity formation in laser welding: A physics-informed deep learning framework"

ARTICLE INFO

Keywords:
Laser beam welding
Porosity prediction
Physics-informed deep learning
SHapley Additive exPlanations, statistical validation

ABSTRACT

Meng et al. (2025) introduce a physics-informed deep learning (PIDL) framework for predicting porosity in aluminum alloy laser welding. Their PIDL model, assessed via SHAP, exhibited superior predictive performance over conventional deep learning models, demonstrated by a 41% reduction in mean square error (MSE). However, feature importances derived from SHAP may be biased, potentially misrepresenting the genuine physical influences on porosity formation. High predictive accuracy does not automatically ensure the reliability of feature importance metrics. This letter underscores the critical need for rigorous statistical validation for reliable feature importance assessments. Integrating robust statistical methods like Spearman's rho, Goodman-Kruskal's gamma, Kendall's tau, and Somers' delta with machine learning enhances the credibility of insights in materials science and manufacturing. Future research should focus on combining ML with robust statistical analysis to improve feature importance reliability and deepen understanding of underlying physical mechanisms.

Meng et al.'s 2025 study, "Toward prediction and insight of porosity formation in laser welding: A physics-informed deep learning framework," presents a PIDL framework for predicting porosity levels during laser beam welding of aluminum alloys [1]. Their research evaluated the PIDL output through SHapley Additive exPlanations (SHAP), highlighting the hierarchical importance of physical variables such as keyhole ratio and downward flow. They demonstrated the PIDL model's superior predictive performance compared to a conventional deep learning model using only welding parameters, achieving a 41% reduction in mean square error (MSE). However, the prioritized feature importances derived from this PIDL model and SHAP may be influenced by bias, implying that the resulting hierarchy may not accurately reflect the true physical influence on porosity formation. We suggest incorporating robust statistical validation methods to address this potential issue and enhance the reliability of these findings.

While Meng et al. (2025) have made a significant contribution to advancing the understanding and prediction of porosity formation in laser welding, this letter raises critical concerns regarding the interpretation of feature importances derived from their methodology. Although they clearly present their assessment of predictive performance using metrics like MSE and maximum relative error, it is vital to distinguish between achieving a successful prediction and ensuring the reliability of the feature importance metrics derived from machine learning (ML) models [2,3]. Accurately predicting an outcome is distinct from gaining a dependable understanding of which input variables genuinely drive that outcome or the extent of their influence. As widely recognized in the ML community and supported by numerous publications, high predictive accuracy does not automatically ensure the validity or interpretability of feature importance metrics derived from a model. Recognizing this difference is crucial when using model-based explanations to infer conclusions about the underlying physical mechanisms. According to established consensus, supported by over 100 peer-reviewed articles, this distinction is critical [4–8]. A detailed discussion and supporting references are provided in the supplementary material.

Meng et al. employed a physics-informed deep learning (PIDL) framework utilizing a deep learning network (DNN) for prediction. While DNNs are powerful tools for this purpose, their complex architectures often result in learned representations that are closely tied to the specifics of the training data. The key point of discussion lies in the caution required when interpreting the results of such a DNN using SHAP values to determine feature importance, as achieving high prediction accuracy does not automatically validate the reliability of these derived importances [9]. This is because the function 'explain = SHAP (model)' fundamentally links SHAP to the underlying model, meaning SHAP values consequently inherit and may even exacerbate any biases present in the DNN's learned representations or its predictions [10–14]. Despite SHAP's widespread adoption, often attributed to its game-theoretic principles, and the availability of other interpretation methods like Permutation Importance or LIME, it is critically important to recognize that SHAP's fundamental dependency means it acts as a mirror, reflecting and potentially amplifying the biases of the model it explains. Therefore, relying on SHAP values derived from a DNN to definitively determine true feature importance is problematic, as current techniques cannot fully eliminate the bias inherited through this interpretation process.

A fundamental challenge in validating feature importance stems from the absence of true ground truth values. This lack means that validating the reliability of feature importance derived from complex models is inherently difficult. Consequently, different models, employing distinct methodologies, inevitably yield model-specific biases and varying feature rankings. A common and significant pitfall observed in many studies is the tendency to substitute prediction accuracy as a proxy for the reliability of feature importance assessment, often treating a

https://doi.org/10.1016/j.scriptamat.2025.116857
Received 24 April 2025; Accepted 29 June 2025

feature's contribution to prediction performance as if it were its true physical significance. However, because ML models are optimized primarily to maximize prediction accuracy, they may, in pursuit of even slight improvements, overfit or leverage patterns that distort the genuine feature importance ranking, leading to the differing outputs seen across various models.

To mitigate these limitations, a rigorous approach is essential, focusing on the nature of data distribution, the statistical relationships between variables, and statistical validation. Effective modeling strategies depend on a thorough understanding of data distribution. Exploring complex variable relationships, particularly through non-parametric methods, is crucial. Furthermore, ensuring the statistical significance of results via hypothesis testing and p-value analysis is vital to prevent spurious conclusions. Robust statistical methodologies comprehensively address these critical considerations. Instead of relying solely on machine learning models and SHAP for feature selection, we advocate for a synergistic integration with unbiased, resilient statistical methods, notably Spearman's rho and Kendall's tau accompanied by p-value analysis [15,16]. These non-parametric methods are particularly effective in capturing monotonic relationships. Other suitable non-parametric methods include Total correlation Effective transfer entropy, effective for complex relationships like non-monotonic collinearity and interactions [17–20]. Emphasizing these statistical foundations will significantly enhance the trustworthiness and credibility of feature importance evaluations in materials science and manufacturing process analysis.

In conclusion, while physics-informed deep learning techniques are effective for prediction, the inherent biases in the underlying deep learning models and their SHAP interpretations necessitate caution in relying solely on them for feature importance assessment. To address these limitations in analyzing porosity formation in laser welding, it is essential to integrate robust statistical methods and rigorous validation. This combined approach is vital for obtaining accurate and reliable insights. Future research should focus on developing innovative methodologies that leverage both machine learning and statistical analysis to enhance the reliability of feature importance assessments.

CRediT authorship contribution statement

Souichi Oka: Conceptualization, Writing – original draft. **Yoshiyasu Takefuji:** Writing – review & editing, Project administration, Supervision.

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.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scriptamat.2025.116857.

References

[1] X. Meng, et al., Toward prediction and insight of porosity formation in laser welding: A physics-informed deep learning framework, Acta Mater 286 (2025) 120740, https://doi.org/10.1016/j.actamat.2025.120740.
[2] A.M. Musolf, et al., What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics, Hum. Genet. 141 (2022) 1515–1528, https://doi.org/10.1007/s00439-021-02402-z.
[3] Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, ACM Queue 16 (2018) 31–57, https://doi.org/10.1145/3236386.3241340.
[4] K. Lenhof, et al., Trust me if you can: A survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer, Brief. Bioinform. 25 (2024) bbae379, https://doi.org/10.1093/bib/bbae379.
[5] H. Mandler, B. Weigand, A review and benchmark of feature importance methods for neural networks, ACM Comput. Surv. 56 (2024) 318, https://doi.org/10.1145/3679012.
[6] J.L. Potharlanka, M.N. Bhat, Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms, Sci. Rep. 14 (2024) 2923, https://doi.org/10.1038/s41598-024-53141-w.
[7] D. Wood, et al., Model-agnostic variable importance for predictive uncertainty: an entropy-based approach, Data Min. Knowl. Discov. 38 (2024) 4184–4216, https://doi.org/10.1007/s10618-024-01070-7.
[8] P.M. Steiner, Y. Kim, The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases, J. Causal Inference 4 (2016) 20160009, https://doi.org/10.1515/jci-2016-0009.
[9] A. Fisher, et al., 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 (2019) 177, https://doi.org/10.48550/arXiv.1801.01489.
[10] B. Bilodeau, et al., Impossibility theorems for feature attribution, PNAS 121 (2024) e2304406120, https://doi.org/10.1073/pnas.2304406120.
[11] X. Huang, J. Marques-Silva, On the failings of Shapley values for explainability, Int. J. Approx. Reason. 171 (2024) 109112, https://doi.org/10.1016/j.ijar.2023.109112.
[12] I. Kumar, et al., Shapley residuals: quantifying the limits of the Shapley value for explanations, NeurIPS 34 (2021) 26598–26608, https://doi.org/10.48550/arXiv.2106.03583.
[13] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (2024) 101046, https://doi.org/10.1016/j.patter.2024.101046.
[14] C. Molnar, et al., General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller, W. Samek (Eds.), xxAI - Beyond Explainable AI, Springer, 2022, p. 4, https://doi.org/10.1007/978-3-031-04083-2_4.
[15] H. Yu, A.D. Hutson, A robust Spearman correlation coefficient permutation test, Commun. Stat. Theory Methods 53 (2024) 2141–2153, https://doi.org/10.1080/03610926.2022.2121144.
[16] K. Okoye, S. Hosseini, Correlation tests in R: Pearson Cor, Kendall's tau, and Spearman's rho, in: K. Okoye, S. Hosseini (Eds.), R programming: Statistical data analysis in research, Springer Nature, 2024, pp. 247–277, https://doi.org/10.1007/978-981-97-3385-9_12.
[17] S. Tserkis, et al., Quantifying total correlations in quantum systems through the Pearson correlation coefficient, Phys. Lett. A 543 (2025) 130432, https://doi.org/10.1016/j.physleta.2025.130432.
[18] Q. Li, et al., Functional connectivity via total correlation: analytical results in visual areas, Neurocomputing 571 (2024) 127143, https://doi.org/10.1016/j.neucom.2023.127143.
[19] N.A. Caserini, P. Pagnottoni, Effective transfer entropy to measure information flows in credit markets, Stat. Methods Appl. 31 (2022) 729–757, https://doi.org/10.1007/s10260-021-00614-1.
[20] N. Umeki, et al., Evaluation of information flows in the RAS-MAPK system using transfer entropy measurements, eLife 14 (2025) e104432, https://doi.org/10.7554/eLife.104432.

Souichi Oka[a,*], Yoshiyasu Takefuji[b]
[a] *Science Park Corporation, 3-24-9 Iriya-Nishi Zama-shi, Kanagawa 252-0029, Japan*
[b] *Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan*

[*] Corresponding author.
E-mail addresses: souichi.oka@sciencepark.co.jp (S. Oka), takefuji@keio.jp (Y. Takefuji).