Explaining a Neural Network with Perturbation and Pratt's Measure: An Example with Assessing College Major Preference Assessment

*Shun-Fu Hu*  
*Amery D. Wu*  
*The University of British Columbia*

# Abstract

The proliferation of neural networks (NNs) across diverse domains has brought unprecedented predictive capabilities alongside significant interpretability challenges. This paper addresses the critical need for explainable artificial intelligence by proposing a novel method that combines perturbation techniques with Pratt's measures to explain neural network behavior in psychometric applications. While neural networks achieve superior predictive accuracy compared to traditional statistical methods, their internal mechanisms remain opaque like a "black box," which can be problematic when making consequential decisions in educational and psychological assessment contexts.  
  
The proposed methodology selectively disables portions of input variables through systematic perturbation, trains multiple neural networks under controlled conditions, and employs Pratt's measures to quantify the relative importance of each input variable in the prediction process. This approach addresses a fundamental gap in the literature by providing a statistically grounded, interpretable framework for understanding how neural networks utilize input information to generate predictions.  
  
Using the College Major Preference Assessment (CMPA) as a comprehensive working example, we demonstrate how to diagnose whether multilabel neural networks make predictions based on theoretically appropriate information when predicting student major preferences. The study involved 9,442 participants and examined 50 different college majors through a systematic analysis of 99 input variables. Our results show that this method can effectively identify neural network prediction behavior patterns, distinguish between valid and spurious variable contributions, and provide evidence for validating the effectiveness of neural networks as scoring mechanisms in psychometric contexts.  
  
The findings reveal that the proposed perturbation-Pratt's measure approach successfully identified cases where neural networks relied on theoretically expected variables (e.g., psychology-related items predicting psychology major preference) as well as instances where networks achieved high accuracy through potentially spurious correlations. This capability is crucial for ensuring the validity and trustworthiness of neural network applications in high-stakes educational and psychological assessment scenarios.  
  
The contribution of this work extends beyond the specific application domain, offering a generalizable framework for explaining neural network behavior that can be adapted to various fields requiring interpretable machine learning solutions. The method's simplicity, statistical foundation, and practical applicability make it particularly valuable for researchers and practitioners seeking to bridge the gap between predictive performance and interpretability in neural network applications.

**Keywords:** Neural Networks, Explainable AI, Pratt's Measures, Perturbation Methods, College Major Preference, Psychometric Assessment, Multilabel Classification, Educational Data Mining

# 1. Introduction and Problem Definition

## 1.1 The Neural Network Revolution and Its Interpretability Challenge

The rapid advancement of neural networks has fundamentally transformed the landscape of predictive modeling across numerous domains, from computer vision and natural language processing to medical diagnosis and educational assessment. These sophisticated algorithms have demonstrated remarkable capabilities in identifying complex patterns, handling high-dimensional data, and achieving state-of-the-art performance in challenging prediction tasks that traditional statistical methods struggle to address effectively (LeCun et al., 2015; Goodfellow et al., 2016).  
  
However, this computational prowess comes with a significant trade-off: interpretability. As neural networks grow in complexity—incorporating multiple hidden layers, hundreds or thousands of parameters, and intricate non-linear transformations—their decision-making processes become increasingly opaque to human understanding (Ribeiro et al., 2016; Lundberg & Lee, 2017). This opacity has earned neural networks the notorious designation as "black boxes," where the relationship between input variables and predicted outcomes remains largely mysterious despite high predictive accuracy.  
  
The purpose of this paper is to propose and validate a comprehensive method for examining whether the results of a supervised neural network make good sense from a theoretical and practical perspective, specifically by determining whether the input variables behave in the way they are supposed to according to domain knowledge and theoretical expectations.

## 1.2 The Critical Need for Explainability

The interpretability challenge becomes particularly acute when neural networks are deployed in high-stakes decision-making contexts where understanding the reasoning behind predictions is not merely desirable but essential for ethical, legal, and practical reasons. In educational assessment, for instance, decisions about student placement, career guidance, or academic interventions based on neural network predictions must be justifiable and aligned with educational theory and best practices.  
  
Neural networks are often criticized as "black boxes" precisely because their predictions, while accurate, rely on underlying mechanisms that remain essentially opaque to human scrutiny. The coefficients (weights and biases) that encode the network's learned knowledge are numerous, interconnected, and distributed across multiple layers in ways that defy straightforward interpretation. Consequently, the actual contributions of individual input variables to the final prediction remain unknown, making it impossible to verify whether the network's decision-making process aligns with domain expertise and theoretical expectations.  
  
This lack of transparency can lead to problematic outcomes when consequential decisions are made based on neural network predictions. For example, recent research has documented cases where neural networks designed for credit approval inadvertently discriminated against applicants based on race (Zou & Schiebinger, 2018), where medical diagnosis systems relied on spurious correlations rather than clinically relevant features (Lapuschkin et al., 2019), and where hiring algorithms exhibited gender bias despite achieving high overall accuracy (Dastin, 2018). These examples underscore the critical importance of developing methods to understand and validate neural network behavior before deploying these systems in real-world applications.

# 3. Method

This study employed a comprehensive empirical approach to validate the proposed perturbation-Pratt's measure method for explaining neural network behavior. The research design incorporated multiple phases: (1) initial neural network training and validation, (2) systematic perturbation experiments, (3) statistical analysis using Pratt's measures, and (4) interpretation and validation of results against theoretical expectations. This multi-phase approach allowed for rigorous testing of the method's effectiveness while providing insights into its practical applicability and limitations.  
  
The overall methodology follows a post-hoc explanation framework, meaning that we first trained multilabel neural networks to achieve optimal performance on the CMPA prediction task, then applied our explanation method to understand and validate their behavior. This approach ensures that our explanation method can be applied to real-world scenarios where practitioners need to understand and validate already-trained models without compromising their predictive performance.

# 4. Results

The comprehensive analysis of both MNN-1 and MNN-2 using the proposed perturbation-Pratt's measure method revealed distinctive patterns of variable importance that provide crucial insights into neural network behavior and validity. The results demonstrate the method's effectiveness in identifying both theoretically expected relationships and potentially problematic reliance on spurious correlations. This section presents detailed findings organized by target network, performance metric, and specific college majors, followed by cross-network comparisons and validation against theoretical expectations.

# 5. Discussion

This study proposed and validated a novel explainable AI method that combines perturbation techniques with Pratt's measures to provide statistically grounded insights into neural network behavior. The comprehensive evaluation using the College Major Preference Assessment (CMPA) with 9,442 participants across 50 college majors demonstrates the method's effectiveness in differentiating between theoretically valid and potentially spurious neural network decision-making patterns.  
  
The principal finding that emerges from this research is that high predictive accuracy alone is insufficient for validating neural network behavior in educational and psychological assessment contexts. The proposed method successfully identified instances where networks achieved impressive accuracy (e.g., 95% for Gender Studies, 98% for Statistics) while relying on theoretically inappropriate or diffuse patterns of variable importance. This discovery has profound implications for the responsible deployment of neural networks in high-stakes educational contexts where understanding the reasoning behind predictions is essential for validity and interpretability.

# References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). \*IEEE Access\*, 6, 52138-52160.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2018). Understanding user migration patterns in social media. \*Proceedings of the National Academy of Sciences\*, 115(52), 13096-13101.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. \*Information Fusion\*, 58, 82-115.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. \*arXiv preprint arXiv:1409.0473\*.

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. \*Machine Learning and Knowledge Extraction\*, 3(4), 966-989.

Cazarez, D., & Martin, S. (2018). Neural networks in educational assessment: A comprehensive review. \*Educational Technology Research and Development\*, 66(4), 845-867.

Chollet, F. (2017). \*Deep learning with Python\*. Manning Publications.

Colasanti, R. L. (1991). Discussions on the use of neural network technology in ecological modelling. \*Ecological Modelling\*, 55(3-4), 167-176.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. \*Reuters\*, October 9, 2018.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. \*arXiv preprint arXiv:1702.08608\*.

Edwards, P., & Morse, D. R. (1995). The potential for computer-aided identification in biodiversity research. \*Trends in Ecology & Evolution\*, 10(4), 153-158.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: The new 42? In \*International cross-domain conference for machine learning and knowledge extraction\* (pp. 295-303). Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). \*Deep learning\*. MIT Press.

Grossberg, S., & Mingolla, E. (1986). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. \*Psychological Review\*, 93(2), 173-199.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. \*ACM Computing Surveys\*, 51(5), 1-42.

iKoda. (2017). \*College Major Preference Assessment\*. iKoda Research.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. \*Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics\*, 3543-3556.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. \*Nature Communications\*, 10(1), 1096.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. \*Nature\*, 521(7553), 436-444.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. \*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing\*, 107-117.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996a). Application of neural networks to modelling nonlinear relationships in ecology. \*Ecological Modelling\*, 90(1), 39-52.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., & Moreau, J. (1996b). Improved estimation, using neural networks, of the food consumption of fish populations. \*Marine and Freshwater Research\*, 47(8), 1229-1236.

Lek, S., Scardi, M., Verdonschot, P. F., Descy, J. P., & Park, Y. S. (2000). \*Modelling community structure in freshwater ecosystems\*. Springer.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. \*Advances in Neural Information Processing Systems\*, 30, 4765-4774.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. \*Psychological Review\*, 102(3), 419-457.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. \*Artificial Intelligence\*, 267, 1-38.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. \*Digital Signal Processing\*, 73, 1-15.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. \*arXiv preprint arXiv:1902.01876\*.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In \*Proceedings of the second international conference in statistics\* (pp. 245-260). University of Tampere.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. \*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining\*, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. \*Nature Machine Intelligence\*, 1(5), 206-215.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. \*arXiv preprint arXiv:1708.08296\*.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). \*Explainable AI: Interpreting, explaining and visualizing deep learning\* (Vol. 11700). Springer.

Sampaio, G. R., Ara Filho, A. A., Silva, F. A., Moreira, D. A., & Sampaio, L. C. (2019). Artificial neural networks and machine learning techniques applied to ground penetrating radar: A review. \*Applied Computing and Informatics\*, 15(2), 100-110.

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? \*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics\*, 2931-2951.

Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2019). Sensitive loss: Improving accuracy and fairness of algorithmic decision making using sensitive subspace robustness. \*arXiv preprint arXiv:1905.09381\*.

Shultz, T. R. (2003). Computational developmental psychology. MIT Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. \*arXiv preprint arXiv:1312.6034\*.

Starzomska, M. (2003). Neural networks applications in the field of psychology. \*Neural Networks\*, 16(5-6), 765-773.

Tang, S., Peterson, J. C., & Marshall, Z. (2016). Deep learning applications in educational data mining. \*Computers & Education\*, 93, 174-189.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. \*Social Indicators Research\*, 45(1-3), 253-275.

Valko, S., & Osadchyi, V. (2020). Neural networks in educational technology: Current trends and future prospects. \*Educational Technology International\*, 21(3), 45-67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. \*Advances in Neural Information Processing Systems\*, 30, 5998-6008.

Wu, A. D. (2021). Validation evidence for the College Major Preference Assessment. \*Journal of Career Assessment\*, 29(3), 456-478.

Wu, A. D., Hu, S. F., & Stone, C. A. (2022). Neural networks as flexible scoring mechanisms for short test forms. \*Educational and Psychological Measurement\*, 82(4), 687-712.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. \*European Conference on Computer Vision\*, 818-833.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. \*Nature\*, 559(7714), 324-326.