SUPPLEMENTARY DETAILS ON RELATED WORK

An analysis of variance (ANOVA) test is employed to evaluate the significance of interaction terms in an additive model that includes all possible pairwise interactions Fisher (1992); Mandel (1961). The test examines the variance within each group compared to the variance between the groups and determines whether any differences observed in the means of the groups are likely due to chance or to a significant factor. H-statistics measure the strength of pairwise interactions based on the concept of Partial Dependence Function (PDF), which represents the relationship between a target variable and a specific feature while holding other features constant. The H-statistics can be calculated by comparing the PDF of the target feature with the PDF of the target feature conditioned on the presence of another feature. The method has been extended to feature importance and feature interaction detection Friedman & Popescu (2008); Greenwell et al. (2018). Sorokina et al. Sorokina et al. (2008) proposed a grove-based method called Additive Groves (AG) that extends the concept of decision tree ensembles to incorporate additive structure in the model. The method involves training individual trees to capture additive effects of the features on the target variable and then combining them to create a powerful ensemble model. By incorporating additivity into the model, Additive Groves can capture more complex relationships among features. In interaction detection, lasso-based methods are also widely used due to that they can quickly select interactions. One can construct an additive model with many different interaction terms and let lasso shrink the coefficients of unimportant terms to zero Bien et al. (2013).

Recently, more neural network-based methods are proposed to detect feature interactions. Bayesian Neural Networks (BNN)Cui et al. (2020) is to evaluate pairs of features with significant second-order derivatives at the input. By analyzing the posterior distribution of the model parameters, BNN can identify which pairs of features have a significant interaction effect. Specifically, the method evaluates the pairwise interaction by computing the difference between the posterior distributions of the model predictions with and without the interaction term. The Neural Interaction Detection (NID) method detects statistical interactions between features by examining the weight matrices of feed-forward neural networks Tsang et al. (2021). Specifically, interactions are determined by finding a cutoff on the ranking using a special form of the generalized additive model. Deep Feature Interaction Maps detect interactions between two features by calculating the change in the attribution of one feature incurred by changing the value of the second Greenside et al. (2018). It calculates the interaction between pairs of features by comparing the attribution maps generated when each feature is varied individually to the maps generated when both features are varied simultaneously. Singh et. al. Singh et al. (2019) generalized Contextual Decomposition Greenwell et al. (2018) to explain interactions for feed-forward and convolutional architectures.

Most recently, several works have been proposed to attribute predictions to feature interactions The Shapley-Taylor Interaction Index measures the contribution of pairwise feature interactions in machine learning models by combining the Shapley value and the Taylor expansion to estimate the interaction effects. The Shapley value quantifies the marginal contribution of each feature, while the Taylor expansion approximates the prediction function considering main effects and pairwise interactions Grabisch & Roubens (1999); Sundararajan et al. (2020). Integrated Hessians (IH) Janizek et al. (2021) is a method used to assess feature interactions in deep neural networks. It involves computing the integrated Hessian matrix and integrating the Hessian matrix over the input data, IH captures both the local and global curvature information of the loss landscape by analyzing the eigenvalues and eigenvectors of the integrated Hessian matrix. The Archipelago Tsang et al. (2020) architecture is specifically designed for attributing feature interactions in machine learning models. It offers a framework based on mixed partial derivatives for identifying and quantifying the contributions of interactions between features. which consists of an interaction attribution method, ArchAttribute, and an interaction detector, ArchDetect.

These above methods show excellent accuracy in feature interaction predictions from a pre-specified model. However, as more researchers advocate exploring a set of equally good models, it is worthy discovering feature interactions in a model class, defined as the Rashomon set. The Rashomon set is named after Akira Kurosawa's film "Rashomon", and it refers to a collection of diverse but equally plausible models that achieve comparable performance on a given task or dataset. Fisher Fisher et al. (2019) introduces the concept of model class reliance (MCR) as a measure that captures the variability of Variable Importance (VI) values across a well-performing model class known as the Rashomon set. MCR provides a more comprehensive and nuanced understanding of feature

importance by considering the range of VI values obtained from multiple prediction models within the class. Following their work, Dong Dong & Rudin (2020) explores the cloud of variable importance, referred to as VIC for the set of all good models and provides concrete examples in linear regression and logistic regression. Another paper Li & Barnard (2022) proposed a post-hoc method, variance tolerance factor (VTF) to interpret a set of neural networks by greedy searching all possible neural networks with certain conditions.

REFERENCES

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Tianyu Cui, Pekka Marttinen, and Samuel Kaski. Learning global pairwise interactions with Bayesian neural networks. *ECAI*, 2020.

Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 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):1–81, 2019.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Springer, 1992.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, pp. 916–954, 2008.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28:547–565, 1999.

Peyton Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics*, 34(17): i629–i637, 2018.

Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *The Journal of Machine Learning Research*, 22(1):4687–4740, 2021.

Sichao Li and Amanda Barnard. Variance tolerance factors for interpreting neural networks. *arXiv preprint arXiv:2209.13858*, 2022.

John Mandel. Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, 56(296):878–888, 1961.

Chandan Singh, W. James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SkEqro0ctQ`.

Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pp. 1000–1007, 2008.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? Interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159, 2020.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting Statistical Interactions from Neural Network Weights. *ICLR*, 2021.