

CHORD: A Novel XAI Framework for Guiding Counterfactuals via Higher-Order Sensitivity Analysis

Zohaib Shahid
Institute for Digital Technologies
Loughborough University London
London, UK
z.shahid@lboro.ac.uk

Yogachandran Rahulamathavan
Institute for Digital Technologies
Loughborough University London
London, UK
y.rahulamathavan@lboro.ac.uk

Safak Dogan
Institute for Digital Technologies
Loughborough University London
London, UK
s.dogan@lboro.ac.uk

Abstract—Building trustworthy AI systems requires robust and mechanically transparent explainability methods. Although counterfactual explanations effectively provide ‘what-if’ scenarios, their inability to reveal the underlying decision mechanisms in neural networks often limits their reliability. This paper introduces Counterfactuals via Higher-Order Relational Derivatives (CHORD), a novel framework that bridges the gap between sensitivity analysis and counterfactual generation. Unlike traditional first-order methods, CHORD utilizes higher-order partial derivatives to explicitly quantify feature interactions, mapping the complex non-linear dependencies that drive the outputs of a neural network. We demonstrate on the UCI Adult Income and UCI Bank Marketing datasets how this interaction-aware analysis can guide the generation of counterfactuals by targeting groups of coupled features, rather than isolated inputs. It was observed that using CHORD, if the value of the most impactful feature (based on the highest second-order partial derivative) is increased, a lower value of the most impactful feature (from the first-order sensitivity analysis, based on partial derivatives) is required to change the output. Crucially, our experiments show that prioritizing these interactions moves beyond simply finding a valid recourse to understanding the structural mechanisms that enable the switching of predictions. This synthesis results in explanations that are actionable, mechanistically interpretable, and mathematically grounded in the curvature of the network, thereby significantly enhancing AI transparency.

Index Terms—Explainable AI, Sensitivity Analysis, Feature Interactions, Higher-order Partial Derivatives, Counterfactual Examples

I. INTRODUCTION

The increasing deployment of complex machine learning (ML) models, particularly deep neural networks, in high-stakes decision-making has amplified the need for robust interpretability and transparency. Traditional explainable AI (XAI) methods often focus on post hoc explanations, providing insight into the decisions of a neural network after they are made [1]. A prominent and psychologically intuitive class of these methods is counterfactual explanation, which addresses the “what-if” question: “What would have to change in the input for the prediction of a model to change to a desired outcome?” [2], [3]. These explanations are particularly pow-

erful because they provide actionable recourse for individuals affected by algorithmic decisions, such as a loan applicant who must learn what to change in their profile to get approval [1].

Although counterfactuals offer a compelling form of explanation, their generation and analysis present significant challenges. The core objective of this research is to explore the application of higher-order partial derivatives, or sensitivity analysis, to systematically study counterfactual examples within neural networks by using a novel framework named **CHORD** shown in Figure 1. This approach seeks to move beyond simply identifying a counterfactual instance to understanding the mechanisms that govern its creation. By analyzing the partial derivatives of the output of the model with respect to its inputs and their interactions, we can quantify the influence of each input feature on the final prediction. This quantitative measure of impact can then be leveraged within a counterfactual framework to investigate how targeted changes in the most sensitive inputs propagate through the network to alter the final decision. The ultimate goal is to establish a robust method for evaluating and generating counterfactuals that are not only valid but also causally informative and robust to perturbations. Also, this study will help us to understand the non-linear behaviour of neural networks in a little more depth. By integrating sensitivity analysis, we can identify which features are truly “critical” for the prediction of a neural network and, conversely, which features can be altered without altering the outcome [4]. This synthesis will provide a powerful tool for developing more reliable, transparent, and trustworthy AI systems.

Concerning CHORD, first, a neural network (M1) is trained (given any tabular dataset), and the first-order partial derivatives (Jacobian matrices) for all rows in the test dataset are generated. The test dataset is used because it is unseen data and can help in better study of the explainability of the neural network. Now, the framework will go from one row to another and pick the features with the highest positive and negative first-order partial derivatives (most impactful features). Note that the highest negative partial derivative is the one which

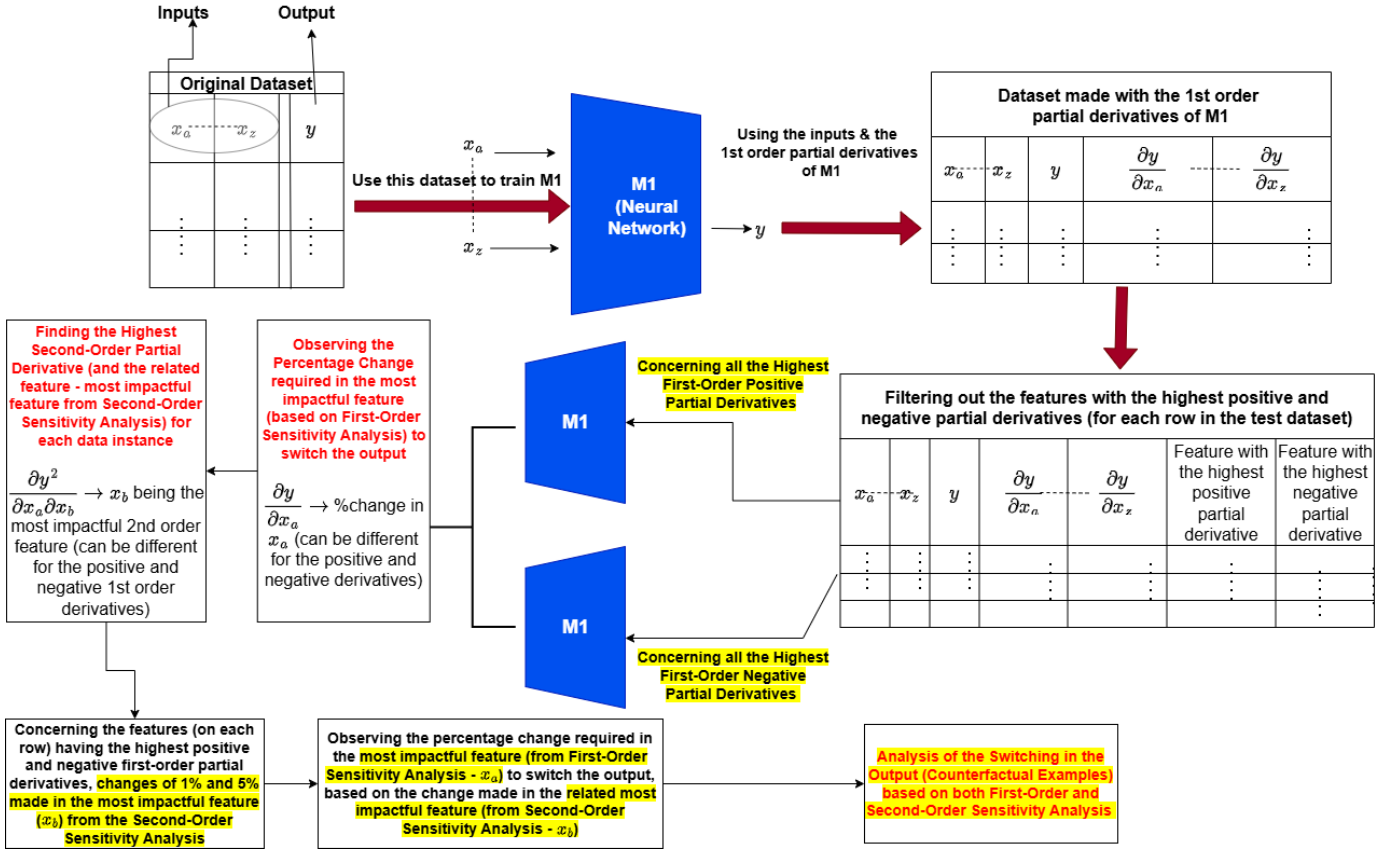


Fig. 1: High-level view of the proposed CHORD framework

has the highest absolute value (with the negative sign). Both positive and negative partial derivatives have been considered, so the impact of the direction of the partial derivative can also be studied.

The logic behind the study of first-order sensitivity analysis is that concerning a feature with the highest positive partial derivative, the output should switch in a directly proportional manner with respect to the change in the value of the feature. For a feature with the highest negative partial derivative, the switching of the output should be in an inversely proportional manner with respect to the change in the value of the feature. The required percentage change in the most impactful feature (required for switching the output) will be noted for each data point after passing the relevant datasets through M1 again.

The feature interactions are studied by finding the most impactful features (through second-order partial derivatives or Hessians) and changing them (per row of the dataset) by 1% and 5% (as per the experiments in Section III). Then, by keeping that feature at the percentage change of 1% or 5%, it is observed whether the output can be switched at a lower percentage change of the most impactful feature (based on the first-order partial derivative) as compared to first-order sensitivity analysis. The experiments around the CHORD framework are discussed in detail in Section III while the literature review, around sensitivity analysis and counterfactual examples, is done in Section II. The paper is concluded with

some future research directions in Section IV.

II. LITERATURE REVIEW

The field of XAI is dedicated to developing methods to understand and interpret complex ML models, particularly opaque “black-box” systems [4]. Post hoc explanations, which provide insight after a model has made a prediction, are a central focus of this research [1]. Counterfactual explanations (CFEs) are a particularly intuitive form of post hoc explanation rooted in human causal reasoning, providing ‘what-if’ scenarios that describe the minimal changes to an input that would alter a model’s prediction. The foundational gradient-based recourse proposed by Wachter et al. [5] and the work by the authors of [2] framed this problem as an optimization task, utilizing gradient descent to minimize the distance between the original and counterfactual instance while ensuring that the output of the model flipped to the desired class.

However, this approach (mostly distance-minimizing) has been criticized for not adequately considering other crucial properties. Subsequent research has expanded on this foundational work by incorporating additional desirable properties into the optimization process. These include diversity (providing multiple distinct counterfactuals), proximity (ensuring counterfactuals are close to the original instance), sparsity (making changes to as few features as possible), and feasibility or actionability (ensuring changes are practical and adhere to

real-world constraints) [1], [6]. For instance, a counterfactual that suggests changing one’s race to get a loan is not actionable, while changing one’s income might be [3], [7].

The work of [1] proposed the Diverse Counterfactual Explanations (DiCE) framework, which explicitly optimizes for diversity in addition to proximity, arguing that a single counterfactual is often insufficient for human understanding. Similarly, the study on collective counterfactual explanations by [2] introduced a novel mathematical optimization framework to find explanations for a group of instances, demonstrating how this approach can identify critical features of an entire dataset. These advancements highlight a shift in the field toward generating richer and more context-aware explanations.

A. The Role of Sensitivity Analysis and Derivatives

While foundational CFE optimization methods rely on first-order gradients to find the shortest path across a decision boundary [2], [3], [5], the application of higher-order partial derivatives, or sensitivity analysis, provides a direct and quantitative method for understanding the relationship between inputs and outputs in a differentiable model. In the context of neural networks, the gradient of the output with respect to an input feature can be used as a measure of the importance of that feature, indicating how a small change in that input affects the prediction of the model [8]. This is a foundational concept in adversarial examples, where tiny, imperceptible perturbations are crafted to cause a misclassification [4].

This same principle can be inverted and applied to counterfactual analysis. Unlike traditional first-order methods, the proposed CHORD framework aims to not only study and connect higher-order sensitivity analysis and counterfactual examples efficiently but also contributes to mechanistic interpretability (by proper emphasis on the study of the nonlinearity of neural networks). It calculates the first-order and second-order partial derivatives of the output of a neural network with respect to its input features. In this way, the sensitivity of the prediction to each feature (and interactions) can be quantified. In a counterfactual scenario, where the objective is to flip a prediction, these derivatives can be used to identify the most efficient *route* across the decision boundary. Specifically, it can be determined which input feature, if changed minimally, would have the greatest impact on the output score. This approach goes beyond purely distance-based or random methods by directly leveraging the internal mathematical structure of the neural network.

Furthermore, analyzing the sensitivity of a model in counterfactual space has implications for other areas of XAI. For example, in mechanistic interpretability, where the goal is to reverse-engineer the internal workings of a neural network, sensitivity analysis can be used to explore how changes to specific features or “circuits” within the network propagate and influence the final output. The paper by [9] introduces a framework for analyzing the sensitivity of counterfactuals to parametric assumptions, a problem analogous to the use of derivatives to explore the local behavior of a function (responsible for generating the output of a model). Their work

demonstrates the importance of moving beyond single-point estimates to analyze the robustness of model outputs under a range of conditions, a principle that is directly relevant to using derivatives to explore the sensitivity of a neural network. By integrating these quantitative methods, this research aims to provide a more rigorous and reliable basis for both generating and validating counterfactual explanations.

III. COUNTERFACTUALS VIA HIGHER-ORDER RELATIONAL DERIVATIVES (CHORD)

To find the connection between higher-order sensitivity analysis and counterfactual examples, the testing is initiated with first-order partial derivatives. It is well-known that first-order partial derivatives calculate the impact of a single input feature on the output. The CHORD framework has been described in detail in Section I. The UCI adult income [10] dataset, which aims to predict whether a person will make less than or equal to \$50K per year ($y = 0$) or above \$50K per year ($y = 1$), is a multivariate dataset with 30,162 instances (cleaned dataset) and 14 features. The neural network (M1) for the UCI Adult Income dataset had 4 hidden layers (with 10 neurons each and having ReLU activation) and 1 output (sigmoid activation). The UCI bank marketing dataset contains data for marketing campaigns based on phone calls, and the target was to assess whether a client would subscribe to a term deposit (*yes* or $y = 1$) or not (*no* or $y = 0$). The related neural network (M1) had 5 hidden layers (with 10 neurons each and having ReLU activation) and 1 output (sigmoid activation).

As shown in Table I, the performance metrics for the neural networks are quite moderate (concerning the chosen architectures designed to gain good performance, overall).

TABLE I: Performance Metrics of the chosen Datasets

Dataset	Accuracy of M1	Recall of M1	F1 score of M1
UCI adult income dataset	80.7%	76.3%	0.662
UCI Bank Marketing dataset	84.1%	55.3%	0.441

Some conditions were put in place to properly study the counterfactual examples, based on first-order sensitivity analysis and second-order sensitivity analysis. The combination of the first-order sensitivity analysis and counterfactuals was studied by increasing/decreasing the concerned feature (with the highest positive/negative first-order partial derivative) up/down to 3 times the magnitude of the original feature to observe the switching of the output. Any impactful features with a value of zero have not been included in the experiments, as they could not be increased or decreased. In addition, any rows where no change in the output could be achieved (within the limits of the experiments) have been excluded. Therefore, only those rows where a change in the output (counterfactuals) did occur via the experiments have been considered. Also, concerning second-order sensitivity analysis, any feature interaction of a feature with itself has been excluded, as that

TABLE II: Most Impactful Features, concerning both First-Order Positive and Negative Partial Derivatives

Dataset	Top Feature (1st Order analysis based on $+ve$ partial derivatives)	Top Feature (2nd Order analysis based on $+ve$ 1st order partial derivatives)	Top Feature (1st Order analysis based on $-ve$ partial derivatives)	Top Feature (2nd Order analysis based on $-ve$ 1st order partial derivatives)
UCI adult income dataset	capital_gain	education	sex	capital_gain
UCI Bank Marketing dataset	euribor3m	pdays	pdays	poutcome

TABLE III: Average Values of the Most Impactful Feature (from 1st Order Sensitivity Analysis) required to Switch the Output for the UCI Adult Income Dataset Before and After Changing the value of the Most Impactful Feature (from 2nd Order Sensitivity Analysis) by 1% and 5%

Type of 1st Order Partial Derivative	Avg Value before applying 2nd Order Sensitivity Analysis	Avg Value after applying 2nd Order Sensitivity Analysis (1 % change)	Avg Value after applying 2nd Order Sensitivity Analysis (5 % change)
$+ve$	21%	17.6%	17.5%
$-ve$	15.3%	12.3%	12.4%

would have greatly expanded the experiments. This will be researched in the future.

Then, according to Figure 1, these experiments were extended to include second-order partial derivatives, as discussed in Subsection III-A. According to Table II, concerning the interaction of the features for the adult income dataset, based on the first-order and second-order partial derivatives, it can be said that `capital_gain` and `education` have a synergistic relationship. This means that the higher the level of education, the more responsive the predicted income is to capital gains. Second, the dampening effect of both `sex` and `capital_gain` shows that the neural network relies less on the `sex` or gender feature to differentiate between outcomes when capital gains are already high. Concerning the bank marketing dataset, a reinforcing relationship between `euribor3m` (three-month Euribor rate, related to the loaning strength of banks) and `pdays` (number of days since last contact with a customer) shows that this `euribor3m` has a positive influence on the output if the customer has not been contacted recently. `poutcome` stands for the outcome of the previous calling campaign. The relationship between `pdays` and `poutcome` shows that if the outcome of a previous calling campaign was successful, the customer will most likely sign up for the subscription, regardless of the number of days

TABLE IV: Average Values of the Most Impactful Feature (from 1st Order Sensitivity Analysis) required to Switch the Output for the UCI Bank Marketing Dataset Before and After Changing the value of the Most Impactful Feature (from 2nd Order Sensitivity Analysis) by 1% and 5%

Type of 1st Order Partial Derivative	Avg Value before applying 2nd Order Sensitivity Analysis	Avg Value after applying 2nd Order Sensitivity Analysis (1 % change)	Avg Value after applying 2nd Order Sensitivity Analysis (5 % change)
$+ve$	12.2%	8.6%	8.64%
$-ve$	28.3%	24.9%	25.6%

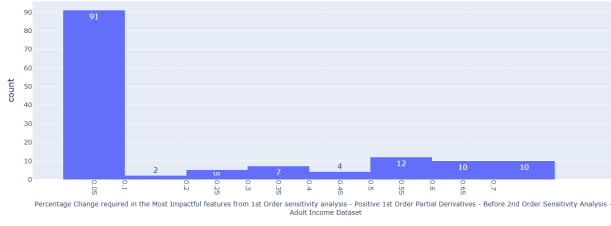
since the last contact.

According to Tables III and IV, it can be seen that with the change in the most impactful feature (from the second-order sensitivity analysis), the switch in the output is achieved at a lower average value of the most important feature from the first-order sensitivity analysis as compared to not applying the second-order sensitivity analysis. This shows the significance of the feature (from the second-order sensitivity analysis) in affecting the output while it interacts with the feature (from the first-order sensitivity analysis). Hence, it can be observed that a better understanding of counterfactuals is grounded in a proper understanding of the feature interactions based on higher-order partial derivatives. A number of histograms (Figure 2) have been included in subsection III-A to elaborate this further.

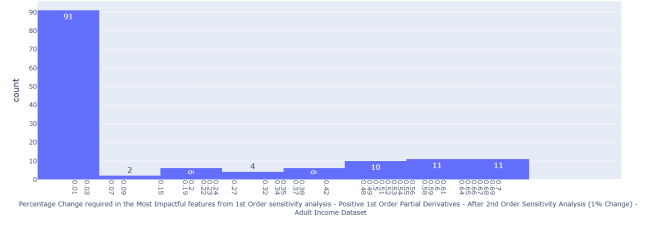
A. Further Experiments on the Second-Order Sensitivity Analysis

For the mentioned datasets, the experiments were extended towards second-order sensitivity analysis or feature interactions. Hessian matrices were generated against every row of the dataset. Hence, a 14×14 matrix was generated for each row of the UCI Adult Income Dataset, as the total number of features is 14. For the UCI bank marketing dataset, the generated hessian matrices for each row were of the dimensions, 19×19 . In every row of each dataset, changes were made in the features with the highest value of the second-order partial derivative against the most impactful feature (having the highest value of the first-order partial derivative found via first-order sensitivity analysis). For this experiment, the feature (with the highest second-order partial derivative) was changed by 1% and 5% to observe whether the feature (with the highest first-order partial derivative) could switch the output at a lower percentage change in itself.

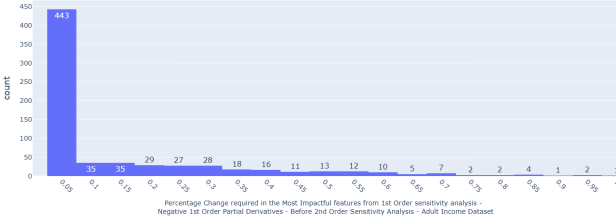
Figure 2 shows the histograms of the percentage change in the most impactful feature (first-order sensitivity analysis) required for switching the output when a change of 1% was made in the most impactful feature (based on second-order sensitivity analysis) for every row of the data. Concerning the positive first-order positive partial derivatives, in Figures 2e and 2f (bank marketing dataset) and Figures 2a and 2b (adult income dataset), it can be observed that there is a clear left shift in the values of percentage change in the most impactful feature, which shows that a small amount of change



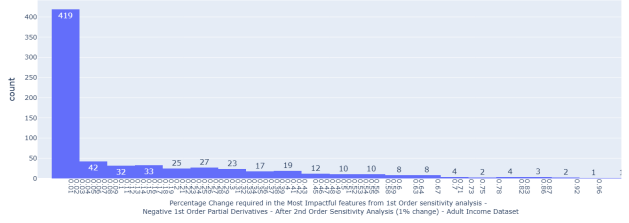
(a) Histogram of Percentage Change in the most Impactful Feature before Second-Order Sensitivity Analysis (Adult Income Dataset) - $+ve$ 1st Order Partial Derivatives



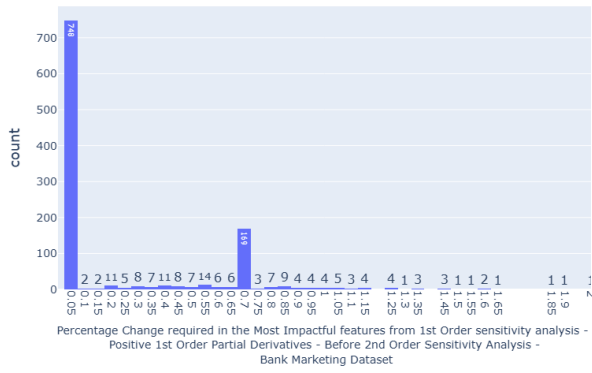
(b) Histogram of Percentage Change in the most Impactful Feature after Second-Order Sensitivity Analysis (Adult Income Dataset) - Change of 1% - $+ve$ 1st Order Partial Derivatives



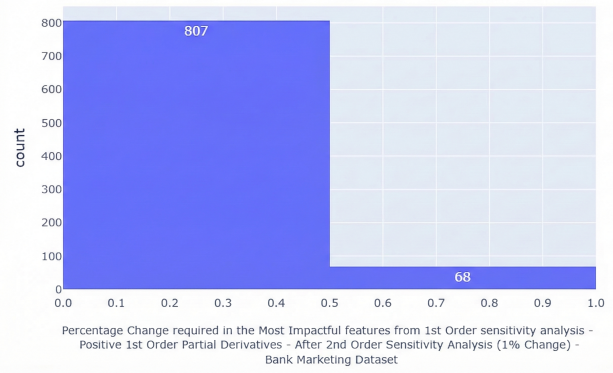
(c) Histogram of Percentage Change in the most Impactful Feature before Second-Order Sensitivity Analysis (Adult Income Dataset) - $-ve$ 1st Order Partial Derivatives



(d) Histogram of Percentage Change in the most Impactful Feature after Second-Order Sensitivity Analysis (Adult Income Dataset) - Change of 1% - $-ve$ 1st Order Partial Derivatives



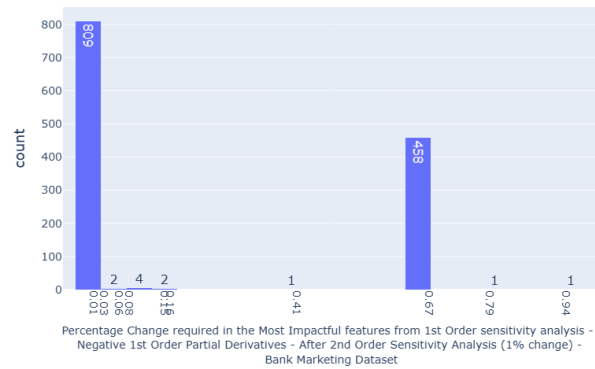
(e) Histogram of Percentage Change in the most Impactful Feature before Second-Order Sensitivity Analysis (Bank Marketing Dataset) - $+ve$ 1st Order Partial Derivatives



(f) Histogram of Percentage Change in the most Impactful Feature after Second-Order Sensitivity Analysis (Bank Marketing Dataset) - Change of 1% - $+ve$ 1st Order Partial Derivatives



(g) Histogram of Percentage Change in the most Impactful Feature before Second-Order Sensitivity Analysis (Bank Marketing Dataset) - $-ve$ 1st Order Partial Derivatives



(h) Histogram of Percentage Change in the most Impactful Feature after Second-Order Sensitivity Analysis (Bank Marketing Dataset) - Change of 1% - $-ve$ 1st Order Partial Derivatives

Fig. 2: Results of the 2nd Order Sensitivity Analysis

in the feature (with the highest first-order partial derivative) was required to switch the output (given a small increase in the most impactful feature from the second-order sensitivity analysis).

The same trend can be seen for the negative first-order partial derivatives (Figures 2c and 2d (adult income dataset) and Figures 2g and 2h (bank marketing dataset)). The summarized results for the change of 5% are present in Tables III and IV and reflect almost the same trend. In some cases like in Figure 2f, a clear left shift in the bars might not be visible but it can be seen that many values of percentage change in the most impactful feature (from first-order sensitivity analysis) have shifted to lower values for switching the output. Hence, the bars have grown a bit higher and wider. This clearly indicates that understanding feature interactions and their connection to counterfactual examples (through comparative analysis) can lead to a better understanding of the granularity of the functioning of neural networks. This also shows a directed approach where only a small number of features need to be changed together to get a desired output.

IV. CONCLUSION AND FUTURE WORKS

In this study, a novel XAI framework named Counterfactuals for Higher-Order Relational Derivatives (CHORD) is presented and discussed. By applying CHORD to the UCI Adult Income and Bank Marketing datasets, it was found that as sensitivity analysis produces a measurable causal analysis of counterfactual examples, better actionable insights are produced. One of the clear conclusions is that with a small increase in the most impactful feature (based on sensitivity analysis from second-order partial derivatives), a lower value of the most impactful feature (based on first-order partial derivatives) is required to switch the output. However, further experiments are required to understand how CHORD can be developed further for high-dimensional or multimodal models and investigating counterfactual feasibility under real-world constraints (like immutable attributes, correlated socio-economic variables, etc.). In addition, this research can be steered towards understanding adversarial attacks in neural networks. CHORD will be further developed to help understand mechanistic interpretability, especially in the study of polysematicity.

ACKNOWLEDGMENT

According to the rules of IEEE publications, it is acknowledged that Google Gemini was used to generate some text in this paper. The citation for Google Gemini has been added to the sections where any included content was generated by it.

REFERENCES

- [1] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)*. ACM, 2020, pp. 607–617.
- [2] E. Carrizosa, J. Ramírez-Ayerbe, and D. Romero Morales, "Generating collective counterfactual explanations in score-based classification via mathematical optimization," *Expert Systems With Applications*, vol. 238, p. 121954, 2024.
- [3] Google, "Gemini (Version 2.5 Pro)," <https://gemini.google.com/>, 2025, accessed: Dec. 16, 2025.
- [4] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. AAAI Press, 2021, pp. 11 575–11 585.
- [5] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
- [6] J. Ji, Z. Li, S. Xu, M. Xiong, J. Tan, Y. Ge, H. Wang, and Y. Zhang, "Counterfactual collaborative reasoning," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. ACM, 2023, pp. 249–257.
- [7] G. Cornacchia, V. W. Anelli, F. Narducci, A. Ragone, and E. Di Sciascio, "Counterfactual reasoning for bias evaluation and detection in a fairness under unawareness setting," *arXiv preprint arXiv:2302.08204*, 2023.
- [8] Z. Shahid, Y. Rahulamathavan, and S. Dogan, "Second glance: A novel explainable ai to understand feature interactions in neural networks using higher-order partial derivatives," in *xAI (Late-breaking Work, Demos, Doctoral Consortium)*, 2024, pp. 193–200.
- [9] T. Christensen and B. Connault, "Counterfactual sensitivity and robustness," *Econometrica*, vol. 91, no. 1, pp. 263–298, 2023.
- [10] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: <https://doi.org/10.24432/C5XW20>.