

Interaction Networks for Cardiometabolic Risk Models: Auditing Learned Model Structure and Clinical Plausibility Across Chronic Disease Domains

Amber Fatima¹ and Zohaib Shahid²

Abstract—Cardiometabolic risk models are increasingly deployed in clinical settings. The issue is that most explainability methods treat input features as independent and offer counterfactual recourse without checking whether the suggested change is clinically achievable. We extend CHORD, a higher-order sensitivity analysis framework, with two additions, which are, a population-level interaction network, built by aggregating the Hessian of a scalar decision margin across a test population into a signed, weighted graph of reinforcing and dampening feature couplings and plausibility-constrained recourse. This recourse bounds counterfactual sweeps to clinically defensible ranges and reports the resulting plausibility gap. Hence, it defines the share of recourse lost once implausible changes are excluded. Applied to cardiovascular and diabetes risk classification datasets, the method recovers clinically coherent interaction structure, and exposes plausibility gaps between 6% and 83%, indicating that a substantial share of gradient-based recourse is not patient-actionable. Compared against the Contrastive Explanation Method (CEM) and Integrated Hessians, our proposed method produces more sparser recourse markedly, changing one feature against roughly three to seven for CEM. Although CEM flips more predictions, at most 8% of its instances received a clinically plausible counterfactual, against 29% to 75% for ours. Our method’s interaction rankings also correlate positively with Integrated Hessians (Spearman ρ from 0.37 to 0.90). Extending the interaction-network component to NHANES survey data, to study patients receiving GLP-1 receptor agonist therapy, is outlined as a next step.

I. INTRODUCTION

Cardiovascular disease (CVD) remains the leading cause of death worldwide. It is driven by obesity, type 2 diabetes, and disordered glucose and insulin regulation which co-occur and reinforce one another rather than acting in isolation [1]. Explainable AI (XAI) methods, and counterfactual explanations in particular, offer actionable recourse by expressing what would need to change for a prediction to change, but most treat features as independent contributors and miss the interaction structure central to cardiometabolic risk [2].

Our proposed method constructs a signed interaction network from higher-order sensitivity information and constrains counterfactual recourse to clinically plausible feature values (Figure 1). Throughout, the recovered interactions describe properties of a trained model rather than causal or physiological claims. The counterfactuals describe model behaviour rather than treatment recommendations.

*This work was not supported by any organization

¹ Amber Fatima is with the Department of Public Health and Social Care, QA Higher Education, London, UK. amber.fatima@qa.com

² Zohaib Shahid is with the Institute for Digital Technologies, Loughborough University London, UK. z.shahid@lboro.ac.uk

II. LITERATURE REVIEW

A. Interaction Detection and Counterfactual Explanation

Feature-attribution methods such as SHAP [3], and gradient-based methods such as Integrated Gradients [4], [5], assign an importance value to each input but do not reveal the interaction structure. SHAP interaction values [6] extend Shapley attribution to feature pairs. Neural Interaction Detection [7] recovers interactions directly from a network’s learned weights while graph-based approaches represent features as nodes and interactions as edges. All of these are diagnostic as they describe which features interact but do not generate recourse or enforce plausibility.

Counterfactual methods address recourse directly, from Wachter et al.’s distance-minimising formulation [8] to frameworks like DiCE which add diversity, proximity, sparsity, and actionability [9]. In clinical use, recourse must further satisfy causal consistency and clinical safety. A counterfactual suggesting an implausible glucose value, or modifying an immutable attribute such as age, is invalid regardless of its mathematical validity. No prior method, to the best of our knowledge, combines interaction-aware structure with plausibility-constrained counterfactual recourse.

III. PROPOSED METHOD

The proposed method extends the CHORD framework of [5], [10], which uses first-order and second-order partial derivatives to explain and generate counterfactuals for a trained neural network, into two directions which operate on a shared computational core. Figure 1 summarises the resulting framework. First-order sensitivity, restricted to mutable features the patient could plausibly modify, drives a plausibility-constrained recourse/counterfactual generation procedure (as described in Section III-C). Second-order sensitivity, aggregated across the population rather than examined row by row (per-patient), drives the construction of a signed feature interaction network. This describes the structure, the trained model (neural network) has learned (Section III-D). Both elements are computed from a single forward and backward pass per test instance.

A. Model and Decision Margin

Let $f_\theta : \mathbb{R}^p \rightarrow \mathbb{R}^K$ denote the trained network M1, mapping an input $\mathbf{x} = (x_1, \dots, x_p)$ to a vector of class logits $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \dots, z_K(\mathbf{x}))$, with predicted class, $c(\mathbf{x})$, as shown by $c(\mathbf{x}) = \arg \max_{k \in \{1, \dots, K\}} z_k(\mathbf{x})$. Carrying on from [10], we define the scalar *general decision margin* as

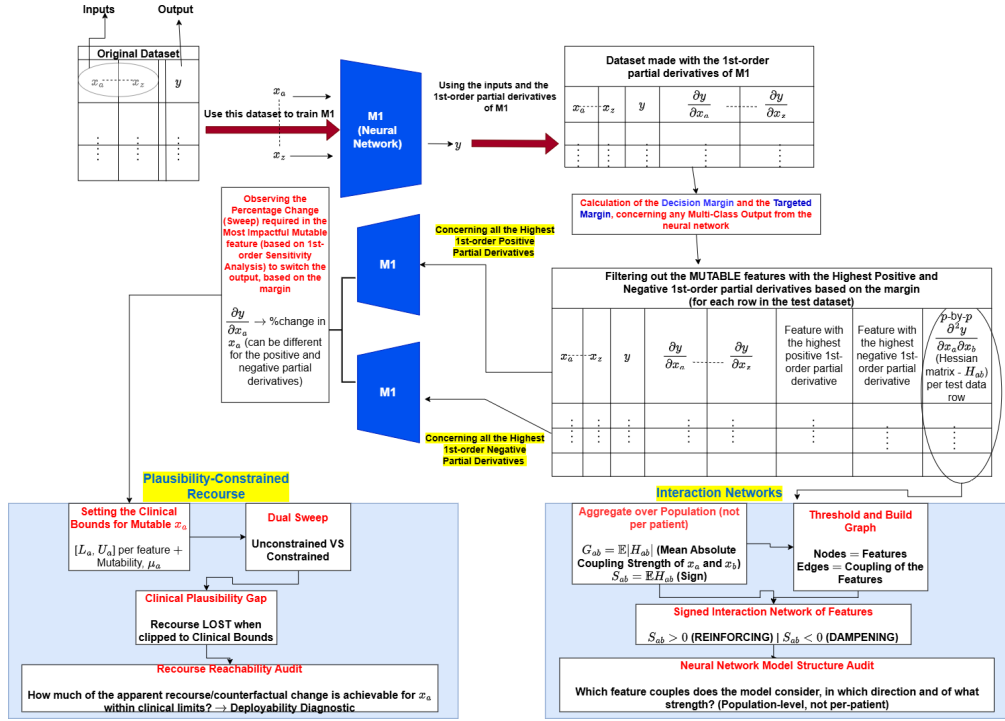


Fig. 1: Overview of the proposed method

the gap between the predicted class logit and its nearest competitor, as given by Equation (1).

$$m(\mathbf{x}) = z_{c(\mathbf{x})}(\mathbf{x}) - \max_{k \neq c(\mathbf{x})} z_k(\mathbf{x}), \quad (1)$$

and a *targeted margin* against a fixed class of clinical interest t , as given by $m_t(\mathbf{x}) = z_{c(\mathbf{x})}(\mathbf{x}) - z_t(\mathbf{x})$, for $\mathbf{x}$ with $c(\mathbf{x}) \neq t$. Both quantities are scalar regardless of K , which is what allows the first- and second-order machinery (for calculating the partial derivatives) described ahead to operate unchanged on any number of classes.

Equations (1) and the equation for targeted margin are stated here for generality, since they are used mainly for the multi-class formulation (classes > 2) in the underlying framework. However, for a binary outcome, $K = 2$ and the only class other than $c(\mathbf{x})$ being necessarily the competitor, considered in Equation (1), so $m(\mathbf{x}) \equiv m_t(\mathbf{x})$ for any choice of $t \neq c(\mathbf{x})$. Hence, the two definitions coincide and collapse into a single quantity. The cardiovascular (hypertension) and diabetes risk datasets [11] used in the present study (Section IV) both carry binary outcomes, and we report results, in terms of the general margin $m(\mathbf{x})$, throughout.

B. First-Order Sensitivity and Mutable Feature Selection

For each test instance, the Jacobian of the margin is obtained by automatic differentiation, as given by $\nabla_{\mathbf{x}} m(\mathbf{x}) = \left(\frac{\partial m}{\partial x_1}, \dots, \frac{\partial m}{\partial x_p} \right)$. Each feature $a \in \{1, \dots, p\}$ is assigned a mutability indicator $\mu_a \in \{0, 1\}$, with $\mu_a = 0$ marking a feature that is not actionable by a patient (for example, age or a diagnostic finding not under patient control). It is, therefore, excluded from consideration as a **primary driver or the most impactful feature from the first-order sensitivity**

analysis. Filtering by mutability is applied before feature selection, not after. While writing $\mathcal{E} = \{a : \mu_a = 1\}$ for the eligible set, the most impactful mutable drivers are, as defined by Equation (2).

$$a^+ = \arg \max_{a \in \mathcal{E}, \partial m / \partial x_a > 0} \frac{\partial m}{\partial x_a}, \quad a^- = \arg \min_{a \in \mathcal{E}, \partial m / \partial x_a < 0} \frac{\partial m}{\partial x_a}. \quad (2)$$

Because $m(\mathbf{x}) \geq 0$ with $m(\mathbf{x}) = 0$ on the decision boundary, a feature with $\partial m / \partial x_a > 0$ must be decreased to force a flip in the output (counterfactual generation), and a feature with $\partial m / \partial x_a < 0$ must be increased. Both directions are tracked separately throughout.

C. Plausibility-Constrained Counterfactual Recourse

For a selected driver, $a \in \{a^+, a^-\}$, let $x_a^{(0)}$ denote its observed value. A multiplicative sweep (small percentage changes in the observed value) produces a sequence of candidate values, as depicted in Equation (3).

$$x_a^{(k)} = \begin{cases} x_a^{(0)} (1 - k\delta), & \partial m / \partial x_a > 0, \\ x_a^{(0)} (1 + k\delta), & \partial m / \partial x_a < 0 \end{cases} \quad (3)$$

for $k = 1, 2, \dots, \lfloor \gamma / \delta \rfloor$, a step size δ , and a maximum multiplicative extent γ . The *unconstrained* recourse is the smallest $k\delta$ at which $c(\mathbf{x}^{(k)}) \neq c(\mathbf{x}^{(0)})$ (flipping of the class), with all other coordinates being fixed at $\mathbf{x}^{(0)}$.

Additionally, each mutable feature is assigned a clinically plausible interval, $[L_a, U_a]$, derived from clinical or definitional bounds rather than from the observed range of the data. The *constrained* recourse repeats the sweep of Equation (3) subject to the conditions shown by Equation (4).

$$x_a^{(k)} \in [L_a, U_a] \quad \text{for all } k \text{ up to the flip,} \quad (4)$$

It reports no counterfactual for that instance and the primary driver, if a class flip would first require $x_a^{(k)}$ to leave the range of $[L_a, U_a]$. Let $\mathcal{U}_a$ be the set of test instances for which the unconstrained sweep on a primary driver, x_a , finds a flip in the output, and let $\mathcal{C}_a \subseteq \mathcal{U}_a$ be the subset for which the constrained sweep also finds one. The *plausibility gap* for driver a is, calculated as per Equation (5).

$$\Delta_a = 1 - \frac{|\mathcal{C}_a|}{|\mathcal{U}_a|}, \quad (5)$$

and the overall plausibility gap aggregates over all primary drivers according to $\Delta = 1 - (\sum_a |\mathcal{C}_a|) / (\sum_a |\mathcal{U}_a|)$. Δ_a quantifies the proportion of recourse which a first-order gradient-based counterfactual would report as valid but that is not achievable without leaving the clinically plausible range for that feature. This serves as a diagnostic for deployability for the underlying model, rather than as a property to be minimised.

D. Population-Level Interaction Networks

Independently of the recourse procedure described before, the full Hessian of the margin is computed for every test instance, as given by $H_{ab}(\mathbf{x}) = \frac{\partial^2 m(\mathbf{x})}{\partial x_a \partial x_b}$, $a, b \in \{1, \dots, p\}$. This retains the entire $p \times p$ matrix rather than a single dominant entry. This second-order analysis requires f_θ to use a smooth and twice-differentiable activation such as $\tanh$. A piecewise-linear (ReLU) network has an identically zero Hessian, which would make the entire construction as meaningless. Rather than interpreting $H(\mathbf{x})$ instance by instance, we aggregate it over the N instances of the test population into two population-level matrices, as shown by Equation (6), and generate a feature interaction graph/network.

$$G_{ab} = \frac{1}{N} \sum_{i=1}^N |H_{ab}(\mathbf{x}_i)|, \quad S_{ab} = \frac{1}{N} \sum_{i=1}^N H_{ab}(\mathbf{x}_i) \quad (6)$$

Here, the diagonal is excluded from both matrices since a feature's self-curvature is not an interaction ($G_{aa} = S_{aa} = 0$). G_{ab} gives the mean absolute coupling strength between x_a and x_b and determines whether an edge is retained. S_{ab} gives the mean signed coupling and determines the direction of that edge, as given by Equation (7). Here, $H_{ab}(\mathbf{x})$ measures how x_a 's marginal influence on the margin changes as x_b varies. A *reinforcing* pair ($S_{ab} > 0$) has contributions which compound [5], so increasing x_b strengthens the effect of x_a on $m(\mathbf{x})$. In contrast, a *dampening* pair ($S_{ab} < 0$) has contributions which offset one another, so increasing x_b weakens the effect of x_a .

$$\text{edge}(a, b) \text{ is } \begin{cases} \text{reinforcing,} & S_{ab} > 0, \\ \text{dampening,} & S_{ab} < 0 \end{cases} \quad (7)$$

is present only if $G_{ab} \geq \tau$, for a threshold τ which, in our experiments, is a percentile of the off-diagonal entries of

G . The resulting weighted and signed graph, with features as nodes and thresholded couplings as edges, constitutes the population-level interaction network. It describes which feature pairs, the trained model has learned to treat jointly, in which direction, and with what strength. This is explicitly a property of the model over the population, rather than of any single patient.

IV. EXPERIMENTS AND RESULTS

The proposed method, described in Section III was evaluated on two binary clinical classification tasks, namely, the cardiovascular (hypertension) and diabetes risk [11]. The cardiovascular dataset is a UCI heart-disease-derived feature set of 26,083 records described by 13 clinical and electrocardiographic attributes, with a binary target which indicates the hypertension-related risk [5]. The target is mildly imbalanced, with 14,274 positive and 11,809 negative cases (54.7% against 45.3%). The diabetes dataset is a Behavioural Risk Factor Surveillance System (BRFSS)-derived indicator set of 70,692 records described by 17 behavioural and clinical attributes, with a binary diabetes target.

A. Preprocessing and Mutability

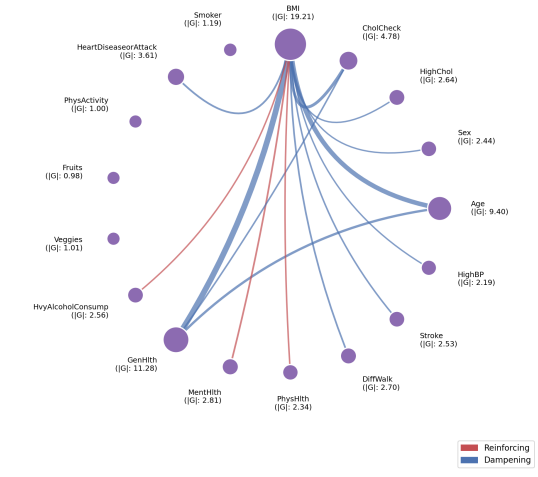
All features were scaled to $[0.05, 1.0]$ with a per-feature min-max transform so that every feature remains strictly positive and can be moved by the multiplicative sweep of Equation (3). Each dataset was partitioned into an 80/20 stratified train and test split with a fixed random seed. This gave 20,846 training and 5,212 test instances for the cardiovascular data, and 56,553 training and 14,139 test instances for the diabetes data.

A feature is eligible as a primary sweep driver only if it is genuinely continuous and mutable in the sense of Section III-B. Four features satisfy both conditions in each dataset, namely, resting blood pressure (`trestbps`), serum cholesterol (`chol`), maximum heart rate (`thalach`), and ST depression (`oldpeak`) for the cardiovascular model, and body mass index (BMI), mental health days (`MentHlth`), physical health days (`PhysHlth`), and general health rating (`GenHlth`) for the diabetes model. All remaining features like demographic attributes or diagnostic findings such as `ca`, `thal`, and prior heart disease, are immutable and excluded from consideration, though they remain eligible to appear as a coupled feature in the interaction network of Section IV-C. Table I lists the clinical bounds assigned to each eligible feature.

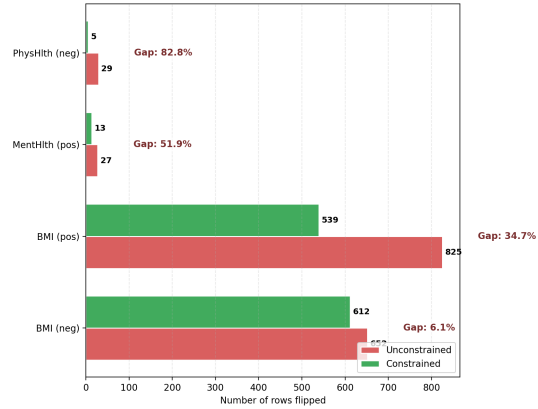
B. Model Training and Experimental Procedure

A single model M1 was trained per dataset which was a fully connected network with four hidden layers of 16 units each, $\tanh$ activations throughout, and a softmax output, satisfying the smooth, twice-differentiable requirement of Section III-D. The training used the Adam optimiser for 300 epochs with a learning rate of 0.01, minimising a class-weighted cross-entropy loss.

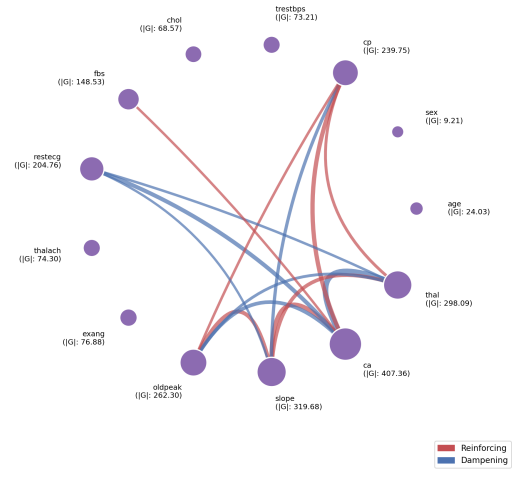
For each test instance, the margin's Jacobian and full Hessian were obtained in a single backward pass. Because



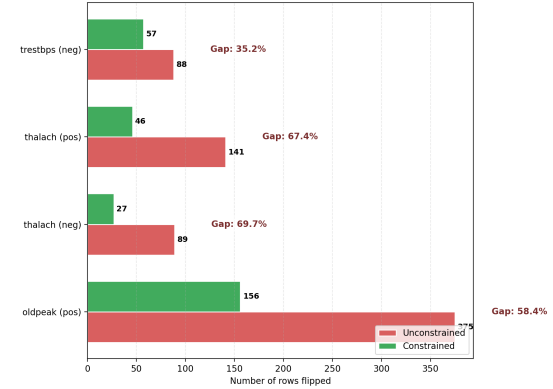
(a) Diabetes: interaction network



(c) Diabetes: plausibility audit



(b) Hypertension: interaction network



(d) Hypertension: plausibility audit

Fig. 2: Interaction networks (top) and plausibility audits (bottom) for the diabetes and hypertension models

TABLE I: Clinically plausible bounds $[L_a, U_a]$ for the eligible, mutable features in each domain. Units: *trestbps* in mmHg, *chol* in mg/dl, *thalach* in bpm; BMI in kg/m²; GenHlth on a 1–5 self-rated scale; MentHlth and PhysHlth in days per 30

Domain	Feature	Bound
Cardiovascular	<i>trestbps</i>	[70, 250]
	<i>chol</i>	[100, 600]
	<i>thalach</i>	[60, 220]
	<i>oldpeak</i>	[0, 8]
Diabetes	BMI	[12, 60]
	GenHlth	[1, 5]
	MentHlth	[0, 30]
	PhysHlth	[0, 30]

the Hessian computation is quadratic in the number of features per instance, the interaction network and plausibility experiments were computed on a fixed random subsample of 1,500 test instances per dataset. For each sampled instance, the eligible mutable drivers with the largest positive and largest negative first-order partial derivative were identified following Equation (2). Both the unconstrained sweep of Equation (3) and the plausibility-constrained sweep of Equation (4) were run for each, using a step size of $\delta = 0.05$

and a maximum multiplicative extent of $\gamma = 3.0$, consistent with the original binary formulation of [10]. The resulting per-instance Hessians were aggregated into the magnitude and signed matrices of Equation (6), and the resulting flip outcomes into the plausibility gaps of Equation (5). Table II reports the performance of each trained model on its complete held-out test dataset. The cardiovascular model is considerably stronger than the diabetes model on every measure, which is relevant to interpretation.

TABLE II: Performance of the trained networks on the complete held-out test dataset

Domain	Accuracy	Recall	F1
Cardiovascular	93.7%	93.3%	0.936
Diabetes	74.9%	74.9%	0.748

C. Interaction Networks

The interaction network for each model is drawn from the matrices, G and S , following Equations (6) and (7). The highest-magnitude feature pairs are retained as edges. The edge thickness is proportional to G_{ab} and the edge colour encodes the sign of S_{ab} (red for reinforcing, blue for dampening), and node size is proportional to a feature’s total coupling strength, $\sum_{b \neq a} G_{ab}$.

Reporting a single “strongest” pair per model requires care, since G and S can rank pairs differently. G ranks by average magnitude regardless of sign consistency across the population, while S ranks by the average signed coupling itself. For the diabetes model (Figure 2a), the pair of largest magnitude among reinforcing edges is BMI and MentHlth ($G = 0.956$), whereas the pair of largest signed value among all reinforcing edges is BMI and HvyAlcoholConsump ($S = +0.638$). The two criteria agree on the strongest dampening pair, BMI and GenHlth ($G = 5.127$, $S = -2.361$, as seen by the thickness of the connection). The same pattern holds for the hypertension model (Figure 2b). The largest-magnitude reinforcing pair is slope and ca ($G = 62.776$), while the largest signed reinforcing pair is cp and ca ($S = +37.582$). The two criteria again agree on the strongest dampening pair, that is, ca and thal ($G = 86.293$, $S = -49.399$). However, one finding is consistent. The strongest (thickest) edge is between the features with the largest nodes (in size).

D. Clinical Interpretation of the Dominant Interactions

This section discusses the signed-ranked strongest pairs of Section IV-C, that is, the pairs with the largest S_{ab} among reinforcing edges and the largest $|S_{ab}|$ among dampening edges, in clinical terms. As throughout this paper, these interactions describe structure, the trained model has learned from the data and are discussed as clinically *plausible* readings. They are not confirmed physiological or diagnostic mechanisms.

For the diabetes model, the strongest reinforcing pair is BMI and heavy alcohol consumption ($S = +0.638$). A reinforcing interaction here means the model treats these two behavioural and metabolic risk factors as compounding. At higher levels of reported heavy alcohol consumption, BMI contributes more strongly to the diabetes-risk margin. This is a plausible pattern, since chronic heavy alcohol use is independently associated with impaired glucose tolerance and altered insulin sensitivity [5]. The strongest dampening pair, BMI and general health rating ($S = -2.361$) points to the opposite pattern. As self-rated health worsens, BMI’s marginal contribution to the risk margin is reduced rather than amplified.

For the cardiovascular model, the strongest reinforcing pair is chest pain type and the number of major vessels affected, cp and ca ($S = +37.582$). This pairing is clinically intuitive. cp is a symptom-level classification reported by the patient, while ca is an objective anatomical finding from fluoroscopy. A reinforcing interaction between them mirrors standard diagnostic reasoning in cardiology, where a symptom pattern consistent with ischaemia and a confirmed burden of vessel involvement are treated as corroborating evidence for one another. The strongest dampening pair, ca and thal ($S = -49.399$), pairs two objective diagnostic findings. ca reflects vessel involvement on fluoroscopy, while thal reflects a perfusion defect on a thallium stress test. Both are, in effect, different imaging windows onto the same underlying construct, which is, the severity of coronary artery disease. Hence, the dampening relationship learned

by the model is consistent with these two findings carrying substantially overlapping information.

E. Plausibility Audit

The plausibility audit compares, for each driver feature and track (based on positive (pos) and negative (neg) first-order partial derivatives), the number of test instances for which the unconstrained sweep of Equation (3) finds a class-flipping counterfactual against the number for which the constrained sweep of Equation (4) also finds one [5]. The bottom row of Figure 2 presents this comparison as paired bars, with the annotated percentage giving the plausibility gap Δ_a , of Equation (5).

For the diabetes model (Figure 2c), the gap varies considerably by track even for the same feature. BMI on the positive track loses 34.7% of its unconstrained recourse (825 instances falling to 539), while BMI on the negative track loses only 6.1% (652 falling to 612). The two single-track features, PhysHlth and MentHlth, lose the largest proportions of either model, at 82.8% (29 to 5) and 51.9% (27 to 13) respectively. Both are self-reported wellbeing counts bounded at 30 days by definition, so the sweep exhausts their plausible range quickly. For the cardiovascular model (Figure 2d), the gaps are more uniform across drivers and tracks, ranging from 35.2% (trestbps, negative track, 88 to 57) to 69.7% (thalach, negative track, 89 to 27).

The Figures, 2c and 2d, carry a direct practical significance. A plausibility gap of a large size means that, for a substantial share of instances, a purely gradient-based counterfactual of exactly the kind reported as “valid” by the unconstrained sweep and by the Wachter and DiCE methods [8], [9] is not something a real patient could act upon. The change it prescribes, would require a feature value falling outside any physiologically plausible range. Left unaudited, such a counterfactual would appear indistinguishable from a genuinely actionable one since standard validity checks confirm only that the predicted class flips. They do not confirm that the flip is reachable by a real patient. The variation in Δ_a across tracks for the same feature, most visibly for BMI in the diabetes model, further shows that plausibility cannot be corrected with a single feature-level adjustment. Whether recourse is achievable depends on the direction of change as well as the feature itself, which is precisely what the per-driver, per-track audit in Figure 2 is designed to expose.

V. COMPARISON WITH BASELINE METHODS

Our proposed framework was compared with the Contrastive Explanation Method (CEM) [12], which searches for a pertinent negative through an elastic-net optimisation of the same decision margin $m(\mathbf{x})$, and with Integrated Hessians [13], which tests the interaction network directly. All the methods explain the same network architecture, data split and test subsample as Section IV-B. As the exact values depend on the trained model, all the results are reported as ranges over four independently initialised networks per dataset. Besides validity (the share of instances

whose predicted class was switched), sparsity and the total absolute change in the scaled space ($L1_{abs}$), we report *plausible validity* [5]. This is the share of instances that received a class-flipping counterfactual which leaves every immutable feature unchanged, maintains every feature within the bounds of Table I and keeps binary and categorical features at levels that exist in the data. CEM flipped the prediction more often (97.3% to 98.7% of instances for diabetes and 68.0% to 78.0% for hypertension, against 70.7% to 74.7% and 28.7% to 54.7% for our constrained sweep of Equation (4)). However, only 6.0% to 8.0% (diabetes) and 0.0% to 0.7% (hypertension) of the instances received a CEM counterfactual that passed all three plausibility checks. For the diabetes model, 89.1% to 90.5% of CEM’s valid counterfactuals altered an immutable attribute (like the age feature), and 79.6% to 87.2% placed a categorical feature at a level that does not exist (like a fractional age band). For the hypertension model, 99.1% to 100% altered an immutable attribute, most often age, followed by diagnostic findings such as *slope*, *ca* and *thal*. Our constrained recourse is plausible by construction as it changes one feature against roughly three (diabetes) and seven (hypertension) for CEM.

Integrated Hessians produces interaction scores but no recourse. So, it was compared with the mean absolute Hessian, G , of Equation (6) on the same instances and the same margin, using the per-feature median of the training split as its reference point. The Spearman rank correlation between the two methods’ pairwise interaction magnitudes was positive and significant for every network, with $\rho = 0.37$ to 0.42 over the 136 diabetes feature pairs ($p < 10^{-5}$) and $\rho = 0.81$ to 0.90 over the 78 hypertension feature pairs ($p < 10^{-18}$).

VI. CONCLUSION

This paper extended CHORD [10] with a population-level interaction network and a plausibility-constrained recourse procedure, evaluated on cardiovascular and diabetes risk classification. The interaction networks recovered clinically coherent structure with the strongest reinforcing and dampening pairs in each domain corresponding to conceptually distinct and conceptually overlapping sources of clinical evidence. The plausibility audit showed that between roughly a quarter and a half of the recourse, an unconstrained sweep would report as valid, which is not achievable within clinically defensible bounds.

When the method was compared with CEM, CEM flipped more predictions than our constrained sweep, but at most 8% of its instances received a counterfactual that respected immutability, clinical bounds and valid feature levels. Our constrained recourse changed one feature and was plausible by construction. Against Integrated Hessians [13], an axiomatic path-integrated second-order attribution method, the recovered interaction structure was corroborated, with a positive and significant rank correlation on both datasets. The proposed method also adds the population-level network construction and the plausibility audit that attribution alone does not provide. Extending the interaction-network component

to NHANES survey data [14], to examine cardiometabolic interaction structure in patients receiving GLP-1 receptor agonist therapy, is a natural next step. This extension remains at an early exploratory stage and is not reported here.

ACKNOWLEDGMENT

According to the rules of IEEE publications, it is acknowledged that Anthropic Claude was used to generate some text in this paper and has been cited where need be.

REFERENCES

- [1] N. Sattar, M. M. Y. Lee, S. L. Kristensen, K. R. H. Branch, S. Del Prato, N. S. Khurmi, C. S. P. Lam, R. D. Lopes, J. J. V. McMurray, R. E. Pratley, J. Rosenstock, and H. C. Gerstein, “Cardiovascular, mortality, and kidney outcomes with GLP-1 receptor agonists in patients with type 2 diabetes: A systematic review and meta-analysis of randomised trials,” *The Lancet Diabetes & Endocrinology*, vol. 9, no. 10, pp. 653–662, 2021.
- [2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bannetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [3] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.
- [4] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 3319–3328, 2017.
- [5] Anthropic, “Claude,” Large language model, <https://www.anthropic.com/claude>, 2026.
- [6] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [7] M. Tsang, D. Cheng, and Y. Liu, “Detecting statistical interactions from neural network weights,” *International Conference on Learning Representations (ICLR)*, 2018, available at: <https://arxiv.org/abs/1705.04977>.
- [8] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2017.
- [9] 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*)*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 607–617.
- [10] Z. Shahid, Y. Rahulamathavan, and S. Dogan, “Chord: A novel xai framework for guiding counterfactuals via higher-order sensitivity analysis,” in *2026 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2026, pp. 1817–1822.
- [11] Prosperchuks, “Diabetes, hypertension and stroke prediction,” Kaggle, 2022. [Online]. Available: <https://www.kaggle.com/datasets/prosperchuks/health-dataset>
- [12] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, “Explanations based on the missing: Towards contrastive explanations with pertinent negatives,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 590–601.
- [13] J. D. Janizek, P. Sturmels, and S.-I. Lee, “Explaining explanations: Axiomatic feature interactions for deep networks,” *Journal of Machine Learning Research*, vol. 22, no. 104, pp. 1–54, 2021.
- [14] Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS), “National health and nutrition examination survey data (2017–march 2020 pre-pandemic cycle),” [Online]. Available: <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020>, U.S. Department of Health and Human Services, Hyattsville, MD, 2021.