

Margin-Guided Counterfactuals: Extending Higher-Order Sensitivity Analysis to Multi-Class Clinical Risk Stratification

Amber Fatima¹ and Zohaib Shahid²

¹ Department of Public Health and Social Care, QA Higher Education, London, United Kingdom

² Institute for Digital Technologies, Loughborough University London, United Kingdom

`amber.fatima@qa.com, z.shahid@lboro.ac.uk`

Abstract. Building trustworthy AI systems requires robust and mechanically transparent explanations. Counterfactual explanations offer useful ‘what-if’ scenarios, but they rarely tell about the decision mechanisms behind them. A previous study introduced a framework called CHORD (Counterfactuals via Higher-Order Relational Derivatives) that bridged sensitivity analysis and counterfactual generation using higher-order partial derivatives to quantify feature interactions. This framework was binary only, where switching a prediction meant crossing a single threshold. This study extends it to clinical risk stratification with three or more ordered classes. Our main contribution is a scalar decision margin, defined as the gap between the predicted class and its nearest competitor. This margin reduces the multi-class output to one quantity, so the existing first-order (Jacobian) and second-order (Hessian) machinery operates unchanged. We also defined a clinically targeted margin which measures recourse toward a specific class of concern, such as high risk. We evaluated this approach on three independent three-class datasets, namely, Maternal Health Risk, Fetal Health (cardiotocography), and Thyroid Disease. Across all three, pre-perturbing the most strongly coupled feature, based on partial derivatives, reduced the change required in the primary driver to switch the predicted class. We further compared our method with two gradient-based baselines. Compared to the Contrastive Explanation Method, our approach yields a far more parsimonious recourse, albeit with lower validity. Against the Integrated Hessians, the recovered interaction structure shows a positive rank correlation, corroborating our findings. The resulting explanations are actionable, mechanistically interpretable, and grounded in the network curvature.

Keywords: Explainable AI, Feature interactions, Counterfactual examples, Sensitivity analysis, Clinical risk stratification

1 Introduction

The deployment of machine learning models in high-stakes decision-making, particularly in clinical settings, has intensified the need for interpretability and

transparency. In medicine, models are increasingly used to stratify risk but a prediction alone offers a clinician little actionable insight. A prominent and psychologically intuitive class of explanation methods addresses the question, that is, “*What would have to change in the input for the prediction to change to the desired outcome?*”. These are known as counterfactual explanations [25, 17], and they are attractive because they provide actionable recourse rather than a static attribution score.

Although counterfactuals are compelling, most established methods treat input features as independent levers. They follow the local gradient to identify the shortest path across the decision boundary [4]. However, neural networks are non-linear, and their features interact. A prior study introduced a framework called CHORD (Counterfactuals via Higher-Order Relational Derivatives) which uses first-order (Jacobian) and second-order (Hessian) sensitivity information to quantify these interactions. It demonstrated that pre-perturbing a coupled feature (the most impactful feature based on the second-order sensitivity analysis) reduces the change required in the primary driver (the most impactful feature from the first-order sensitivity analysis) to switch a binary prediction [22]. However, that work was formulated exclusively for binary classification, where switching the prediction had an unambiguous meaning, that is, crossing a single sigmoid threshold.

Regarding our extension of that framework (Fig. 1), first, a neural network (M1) is trained (given any tabular dataset) to produce a vector of class logits $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \dots, z_K(\mathbf{x}))$ for a K -class problem. Then, 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. Then, concerning the handling of multiple classes (our novelty), a scalar decision margin, $m(\mathbf{x})$, is calculated from the neural network which is defined as the difference between the predicted class logit and its closest competitor. By construction, $m(\mathbf{x}) \geq 0$ and $m(\mathbf{x}) = 0$ are exactly on the decision boundary between the predicted class and its nearest competitor. So, a prediction switches precisely when a perturbation drives $m(\mathbf{x})$ to zero.

A second clinically targeted margin, $m_t(\mathbf{x})$, is also calculated and is defined instead as the difference between the predicted class logit and the logit of a fixed class of interest t (such as high risk). Both margins are scalar despite the network producing K vector-valued outputs. This allows the remainder of the framework, originally formulated for a single binary output, to operate on $m(\mathbf{x})$ or $m_t(\mathbf{x})$ without a further change in structure. The framework will now go from one row to another and pick the features with the highest positive and negative first-order partial derivatives of the margin or the output, indirectly (most impactful features). Note that the highest negative partial derivative has the highest absolute value (with a 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.

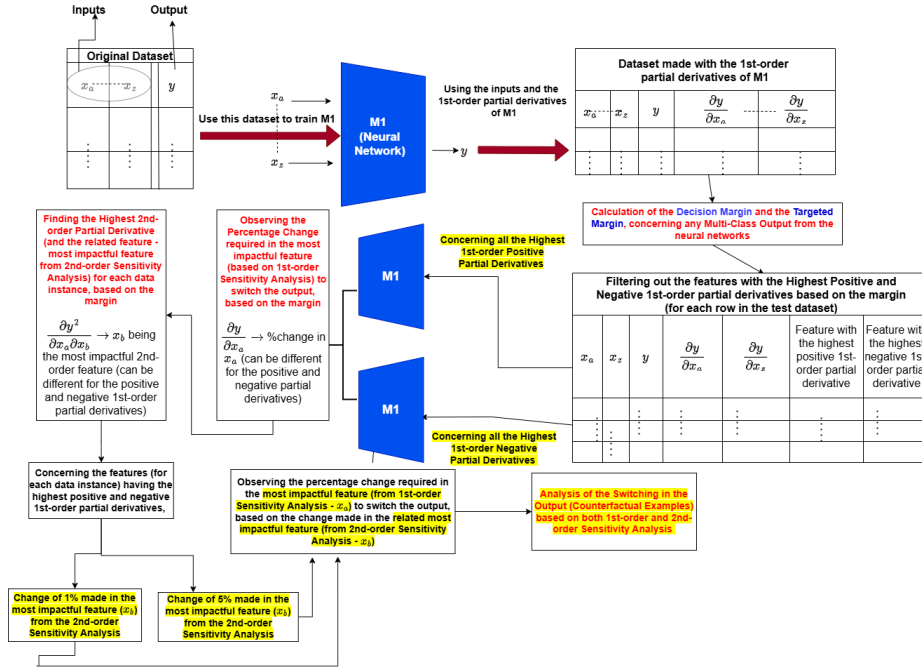


Fig. 1. Overview of the extended framework: the decision margin reduces class logits to a single scalar for first-order and second-order sensitivity analysis.

The logic behind the first-order sensitivity analysis is that for a feature with the highest positive partial derivative, the margin (hence, 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 margin 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 predicted class) is noted for each data point after passing the relevant test datasets through M1 again. The feature interactions are studied by finding the most impactful features (through the highest second-order partial derivative or the Hessian of $m(\mathbf{x})$ or $m_t(\mathbf{x})$) and changing them (per row of the dataset) by 1% and 5% (as per the experiments in Section 4). Then, by keeping that feature at the percentage change of 1% or 5%, it is observed whether the output can be switched to the nearest competitor class or target class at a lower average percentage change of the most impactful feature (based on the highest positive/negative first-order partial derivative) as compared to first-order sensitivity analysis.

Clinical risk stratification is rarely binary. Maternal health risk is routinely graded as low, mid, or high [15, 24], and fetal status derived from Cardiotocography is graded as Normal, Suspect, or pathologic [1, 7]. Thyroid function is likewise graded as hyperthyroid, hypothyroid, or normal [18, 26]. In a multi-class

setting, there is no single threshold. The output space is partitioned into several regions by pairwise decision boundaries. The question “*What is the smallest change that flips the prediction?*” is ill-defined until one specifies which boundary is crossed. Resolving this ambiguity in a principled way is the central contribution of this late-breaking work.

This study makes four contributions. First, we introduce a scalar *decision margin* that generalizes the original framework to an arbitrary number of classes, while leaving the rest of the framework structurally unchanged. Second, we define a clinically *targeted* margin that measures recourse toward a specific class of interest (high risk or pathologic or hypothyroid, as indicated in the concerned datasets), which is often more useful to clinicians than recourse to the nearest class. Third, we apply the extended framework to three independent three-class clinical datasets, the Maternal Health Risk dataset [24], the Cardiotocography (Fetal Health) dataset [7], and the Thyroid Disease dataset [18, 26] to analyze the resulting counterfactuals based on first- and second-order sensitivity behaviour. Fourth, we benchmark the extended framework against two gradient-based baselines. We compare recourse quality against the Contrastive Explanation Method (CEM) [9] and the recovered interaction structure against Integrated Hessians [12].

The remainder of this study is organized as follows. Section 2 reviews counterfactual explanation methods and the role of sensitivity analysis. Section 3 presents the multi-class margin formulation and the rest of the proposed framework in detail. The results of the experiments on the framework and their discussion are presented in Section 4. Section 5 compares the framework against the two baselines, mentioned previously. The limitations of this framework and the conclusion follow in Section 6.

2 Literature Review

2.1 Counterfactual Explanations

Explainable AI (XAI) develops methods to interpret opaque black-box models [5]. Post-hoc methods, which explain a prediction after it is made, dominate the field. Feature-attribution methods such as LIME [19] and SHAP [16] assign an importance value to each input feature, while gradient-based saliency methods such as Integrated Gradients [23, 4] and Grad-CAM [20] highlight influential inputs. These methods answer the question of *which* features matter, but not what a subject would need to change to obtain a different outcome.

Counterfactual explanations address precisely that question. The foundational formulation of Wachter et al. [25] framed the search for a counterfactual as an optimisation problem by minimizing the distance between the original instance and a perturbed instance subject to the model’s output flipping to a desired class. This distance-minimising view was subsequently criticized for neglecting other desirable properties. The DiCE framework of Mothilal et al. [17] explicitly optimizes for diversity in addition to proximity, and formalized the

now-standard properties of proximity, sparsity, and actionability, noting for instance that a counterfactual advising a change to an immutable attribute is not actionable, whereas advising a change to a modifiable one may be.

2.2 Sensitivity Analysis and Feature Interactions

While foundational counterfactual methods rely on first-order gradients to locate the shortest path across a decision boundary, the gradient is by construction blind to feature interactions. The gradient of the output with respect to an input feature measures that feature’s local importance which is a principle also central to the construction of adversarial examples [11]. Second-order information, encoded in the Hessian, captures how the effect of one feature on the output depends on the value of another [21]. The prior framework [22] exploited this by locating, for each instance, the feature most strongly coupled with the primary driver (feature with the highest first-order positive or negative partial derivative), and showed that perturbing this coupled feature (feature with the highest second-order partial derivative) reduces the change required in the driver to flip a binary prediction. The present work extends that idea to a multi-class clinical setting.

2.3 Machine Learning in Maternal Health, Fetal Health and Thyroid Disease

Machine learning has been widely applied to maternal and perinatal risk prediction. The Maternal Health Risk dataset [24] was collected through an IoT-based monitoring system and records physiological markers with an associated low, mid, or high risk label. In fetal medicine, cardiotocography (CTG) is the standard tool for assessing fetal wellbeing, and the SisPorto system of Ayres-de-Campos et al. [8] produced the widely used CTG feature set that underlies the Fetal Health dataset, in which expert obstetricians labelled each trace as Normal, Suspect, or Pathologic. The Thyroid Disease dataset [18, 26] was compiled from clinical thyroid function tests originally collected by the Garavan Institute, with each patient’s diagnosis grouped into hyperthyroid, hypothyroid, or normal categories. Machine learning has also been applied to pre-eclampsia prediction using antenatal markers [13] and to the classification of thyroid disease using serum hormone measurements [2]. These works focus on predictive accuracy. Comparatively, little attention has been given to mechanistically interpretable counterfactual recourse in the multi-class clinical setting, which is the gap this work addresses.

3 Proposed Framework

To find the connection between higher-order sensitivity analysis and counterfactual examples for a neural network with multi-class outputs, 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 proposed framework has been described briefly in Section 1. Further details will be given in this section. The single conceptual change required to generalize the original framework [22] to multiple classes is the introduction of a scalar decision margin, described in Section 3.2. Every subsequent stage operates on this margin exactly as the original framework operated on the binary sigmoid output.

3.1 Trained Neural Networks and Datasets

We evaluate our proposed framework on three independent and publicly available three-class clinical datasets. Both exhibit the ordered risk structure that motivates the decision-margin formulation.

Maternal Health Risk. This dataset comprises 1014 records collected through an IoT-based risk monitoring system in hospitals and community clinics [24]. Each record contains six physiological features (**Age**, **SystolicBP**, **DiastolicBP**, **Blood Sugar (BS)**, **BodyTemp**, and **HeartRate**) and an ordinal target label of low, mid, or high risk [4]. The class distribution is 406 low risk, 336 mid risk, and 272 high risk. All six features are naturally non-zero physiological measurements, so no feature selection is required, and the clinically targeted margin is defined against the high-risk class.

Fetal Health (Cardiotocography). This dataset comprises 2126 records derived from cardiotocogram (CTG) examinations, each labelled by expert obstetricians as Normal, Suspect, or Pathologic [7]. The class distribution is 1655 Normal, 295 Suspect, and 176 Pathologic which shows a substantially heavier imbalance than the maternal dataset [4]. Of the 21 available attributes, we retain the seven clinical CTG event features (**baseline value**, **accelerations**, **fetal_movement**, **uterine_contractions**, **light_decelerations**, **severe_decelerations**, and **prolongued_decelerations**). The remaining histogram summary statistics are descriptive metrics of the heart-rate distribution and are not used in this experiment. The clinically targeted margin is defined against the pathologic class.

Thyroid Disease . The three classes of the Thyroid Disease dataset [18, 26] were derived from the diagnosis codes supplied with the dataset, that is, negative (no condition requiring comment), hyperthyroid (diagnosis groups A to D) and hypothyroid (diagnosis groups E to H). Records whose diagnosis belonged to the remaining metadata groups, namely binding protein, concurrent non-thyroidal illness, replacement therapy, antithyroid treatment and miscellaneous, were excluded. It was because they do not describe thyroid function status and would therefore corrupt a three-class formulation [4]. The TBG assay was discarded as it was missing in over 96% of records. Six continuous features were retained

(TSH, T3, TT4, T4U, FTI and **age**). After complete-case filtering and the removal of physiologically impossible ages, 4,917 records remained, of which 984 formed the concerned test set. All features are scaled to the range $[0.05, 1.0]$ and the three datasets were put through an 80/20 stratified split for the experiments. The strictly positive lower bound is required because the counterfactual sweep changes a feature multiplicatively, and a feature value of exactly zero cannot be moved in this way.

We train a fully connected network M1 with four hidden layers of ten units (each with `tanh` activation to produce Hessians carrying genuine feature-interaction values as ReLU-based neural networks produce zero second-order partial derivatives) for the three datasets. The output layer produces K logits, with $K = 3$ for all the datasets, trained with a cross-entropy objective. Class probabilities are obtained via softmax, though all sensitivity analysis is performed on the raw logits. The moderate performance stats are reported in Table 1 (concerning the chosen architectures designed to gain good performance, overall). The M1 for Thyroid Disease dataset achieved the strongest predictive performance of the three datasets studied, with an accuracy of 93.4% and a macro-averaged recall of 94.3%, despite a prominent class imbalance (4,293 negative, 454 hypothyroid and 170 hyperthyroid records). Class weighting was applied during training to prevent the minority classes from being ignored.

Table 1. Performance metrics of the chosen datasets

Dataset	Accuracy of M1	Recall of M1	F-1 Score of M1
Maternal Health Risk	68.5%	67.1%	0.674
Fetal Health (CTG)	84.3%	66.7%	0.659
Thyroid Disease	93.4%	94.3%	0.813

3.2 The Decision Margin

Let M1 be the trained network, $f_\theta : \mathbb{R}^p \rightarrow \mathbb{R}^K$, mapping a p -dimensional input, $\mathbf{x} = (x_1, \dots, x_p)$, to a vector of class logits, $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \dots, z_K(\mathbf{x}))$. The predicted class is, as calculated by Eq. (1).

$$c(\mathbf{x}) = \arg \max_{k \in \{1, \dots, K\}} z_k(\mathbf{x}) \quad (1)$$

To generalize the notion of switching the output, we define a scalar *general decision margin*, $m(x)$, as the gap between the predicted class logit and its nearest competitor, as given by Eq. (2).

$$m(\mathbf{x}) = z_{c(\mathbf{x})}(\mathbf{x}) - \max_{k \neq c(\mathbf{x})} z_k(\mathbf{x}) \quad (2)$$

where, we denote the nearest competing class by $c'(\mathbf{x}) = \arg \max_{k \neq c(\mathbf{x})} z_k(\mathbf{x})$. By construction, both $m(\mathbf{x}) \geq 0$ and $m(\mathbf{x}) = 0$ are exactly on the decision boundary between $c(\mathbf{x})$ and $c'(\mathbf{x})$. A prediction switches precisely when a perturbation drives $m(\mathbf{x})$ to zero. The binary formulation of the original framework [22] is recovered as the special case, $K = 2$, where the boundary $z_1(\mathbf{x}) = z_2(\mathbf{x})$ coincides with the sigmoid threshold. Crucially, $m(\mathbf{x})$ is scalar despite the network having K vector-valued outputs, so its gradient (Jacobian) and Hessian are computed using exactly the same machinery as in the binary case.

The nearest competitor is not always the clinically relevant one. For a patient predicted as low risk, the nearest competitor may be mid risk. The urgent question is what would reclassify the patient as high risk (in the case of Maternal Health Risk dataset) or Pathologic (Fetal Health Dataset) or Hypothyroid (Thyroid Disease Dataset) (that is, a class of particular interest). We, therefore, also define a *targeted margin*, $m_t(x)$ against a fixed clinical class of interest t , as shown in Eq. (3).

$$m_t(\mathbf{x}) = z_{c(\mathbf{x})}(\mathbf{x}) - z_t(\mathbf{x}) \quad \text{for } \mathbf{x} \text{ with } c(\mathbf{x}) \neq t. \quad (3)$$

We report results under both Eq. (2) and (3). The general margin characterizes the local decision geometry of the neural network. The targeted margin characterizes recourse specifically toward the class of greatest clinical consequence.

3.3 First-Order Sensitivity Analysis

For each test instance, we compute the gradient (Jacobian) of the margin, as shown in Eq. (4), by automatic differentiation.

$$\nabla_{\mathbf{x}} m(\mathbf{x}) = \left(\frac{\partial m}{\partial x_1}, \dots, \frac{\partial m}{\partial x_p} \right), \quad (4)$$

Since m is a difference of two logits, its gradient is the difference of the two corresponding logit gradients (or indirectly, the gradients of the outputs), obtained in a single backward pass. Per row, we select the feature with the largest positive partial derivative, $x_a^+ = \arg \max_i \partial m / \partial x_i$, and the feature with the most negative partial derivative, $x_a^- = \arg \min_i \partial m / \partial x_i$. Because of the prediction (output) switching as $m \rightarrow 0$, a feature with $\partial m / \partial x_i > 0$ widens the margin when increased and is therefore decreased to force a flip, while a feature with $\partial m / \partial x_i < 0$ is increased.

Each selected **primary driver** (most impactful feature), x_a , is then increased/decreased multiplicatively, up/down to three times its original magnitude, and the first percentage change at which $c(\mathbf{x})$ changes is recorded. Rows for which no flip is achievable within this range, or for which the required change comes out as zero, are excluded. This is consistent with the setup of experiments done with the original framework [22].

3.4 Second-Order Sensitivity Analysis

For each instance we compute the Hessian of the margin, as depicted by Eq. (5).

$$H_{ij}(\mathbf{x}) = \frac{\partial^2 m(\mathbf{x})}{\partial x_i \partial x_j} \quad (5)$$

$H_{ij}(\mathbf{x})$ remains well defined and non-trivial owing to the smooth activation. For the first-order driver, x_a , we examine row a of the Hessian and exclude the diagonal self-interaction term, H_{aa} . Then, we identify the **coupled feature** (as per the highest positive or negative second-order partial derivative), based on Eq. (6).

$$x_b = \arg \max_{j \neq a} |H_{aj}(\mathbf{x})|. \quad (6)$$

The coupled feature, x_b , is then pre-perturbed by 1% and by 5%, in the direction consistent with the sign of H_{ab} . Then, the first-order sweep (increase/decrease) for x_a is repeated under this perturbation. Comparing the percentage change in x_a required to flip the prediction (where a lower average percentage change, as compared to the first-order sensitivity analysis, is of interest), before and after perturbing x_b , yields the central comparison of the framework.

4 Discussion of the Results

4.1 Counterfactual Sensitivity: Percentage Change Required to Switch the Prediction

The main results are summarized for all the datasets (concerning test data only: 203 rows for Maternal Health Risk [24], 426 rows for the Fetal Health [7] and 984 rows for the Thyroid Disease [18, 26]) in Tables 2, 3, 4 and 5, 6 and 7. These results clearly show, in the majority of the cases, that a smaller average percentage change in x_a (**primary driver** or first-order feature or the most impactful feature from the first-order sensitivity analysis) is required to switch the output once x_b (**coupled feature** or second-order feature or the most impactful feature from the second-order sensitivity analysis) is pre-perturbed. Concerning the results for the Maternal Health Risk dataset (Tables 2 and 3), it can be clearly seen that the change of 5% in the second-order feature leads to a bigger impact on the first-order feature as it is able to switch the output at a much lower average percentage change value.

Concerning the Fetal Health dataset (Tables 4 and 5), the average percentage change in the first-order feature is indeed lowered by pre-perturbing the second-order feature in some cases but it is a very nominal change. It can also be seen that the average percentage change required in the first-order feature (before and after pre-perturbing the second-order feature) is well beyond 100% and this is worth researching further. This could lead to new insights into the way a neural network works on data which is quite highly imbalanced. The results of the

Thyroid Disease dataset (Tables 6 and 7) denote almost the same trend. Considering the general margin (positive first-order partial derivatives), the average percentage change required in the primary driver x_a to switch the predicted class falls from 52.33% to 45.30% after a 1% perturbation of the coupled feature x_b . It falls further to 24.40% after a 5% perturbation. This represents a reduction of more than half, and is the largest effect observed across all three datasets examined.

The targeted margin, evaluated toward the hypothyroid class, exhibits the same behaviour even more strongly [4]. The required change decreases from 124.43% to 106% and then to 49.09%. This is again more than halving. However, this behaviour is confined to the positive first-order partial derivatives. For the negative first-order partial derivatives, the required percentage change increases slightly rather than decreasing (from 26.90% to 30.20% under the general margin and from 17.42% to 18.19% under the targeted margin). This follows the asymmetry already observed in the Maternal Health Risk and Fetal Health datasets where the effect was similarly more reliable for features with a positive first-order partial derivative. The consistency of this asymmetry across three independent datasets suggests it is a property of the method rather than of any single dataset. Understanding its origin is a clear direction for further research.

Overall, even with the introduction of the scalar decision margin (general and targeted), it can be observed that the results are consistent with the ones, presented in [22]. It can be said that counterfactuals can indeed be understood better by grounding them in feature interactions, which change a little based on the type of the decision margin. This has been described further in Section 4.2.

Table 2. First-order sensitivity results under the general margin for the maternal health risk dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	92%	92%	89.4%
Negative	91.07%	91.45%	86.26%

Table 3. First-order sensitivity results under the targeted margin (high risk class) for the maternal health risk dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	86.71%	85.73%	82.68%
Negative	76.1%	76.1%	71.23%

Table 4. First-order sensitivity results under the general margin for the fetal health dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	112.22%	112.50%	113.29%
Negative	145.18%	144.34%	138%

Table 5. First-order sensitivity results under the targeted margin (pathologic class) for the fetal health dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	119%	118.71%	118.3%
Negative	155.69%	156.49%	154.08%

Table 6. First-order sensitivity results under the general margin for the Thyroid Disease dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	52.33%	45.30%	24.40%
Negative	26.90%	27.73%	30.20%

Table 7. First-order sensitivity results under the targeted margin (Hypothyroid class) for the Thyroid Disease dataset, before and after second-order feature perturbation.

Type of 1st Order Partial Derivative	Before	After 1% change	After 5% change
Positive	124.43%	106%	49.09%
Negative	17.42%	17.60%	18.19%

4.2 Feature Coupling: Reinforcing and Dampening Interactions

To understand the feature interactions, the **feature couples** (x_a , the primary driver from the first-order sensitivity analysis and x_b , the coupled feature from

the second-order sensitivity analysis) were studied as shown in Table 8 for both datasets, concerning the general and targeted margin. For the Maternal Health Risk dataset, across both types of margins, every feature couple produces a dampening interaction, as depicted by the average negative values of the second-order partial derivatives or H_{ab} . In every case, the negative sign indicates that as x_b increases, the marginal influence of x_a on the decision margin (hence, the output) is weakened rather than being strengthened. The magnitude of this dampening is not uniform [4]. The **BS** and **SystolicBP** pairing, which appears independently under both margin definitions, produces a dampening effect quite significantly larger than the **Age** and **BS** or **DiastolicBP** and **BodyTemp** pairings. This suggests that glucose and blood pressure carry a substantially stronger redundant relationship in the internal representation of the neural network than the other maternal feature combinations.

This pattern is consistent with a model which has learned to treat several of its physiological markers as overlapping proxies for a shared underlying risk state. It implies that once one marker signals elevated risk, a correlated second marker contributes comparatively less to the margin on top of it, rather than compounding the risk assessment further. The treatment of glucose and blood pressure by the neural network is consistent with the known clinical overlap between gestational diabetes and hypertensive disorders of pregnancy. This is where the two conditions occur together frequently and are not treated as additive risks in clinical scoring either [10, 4].

The neural network, trained on the fetal health dataset, shows the exact opposite pattern. Every feature couple identified, again across both margins, produces a reinforcing interaction, denoted by the positive sign of the average H_{ab} . Here, the positive values indicate that the coupled feature amplifies, rather than weakens the influence of the primary driver (x_a) on the margin. Notably, the two significantly large reinforcing effects, between x_a and x_b concerning the targeted margin, indicate something interesting. When the analysis is narrowed specifically to escalation toward the Pathologic class (targeted class), the cardiotocographic features learned by the model interact more strongly than when the nearest competing class is analyzed.

The pairing of **accelerations** and **prolongued decelerations** is particularly notable for being mutually coupled. They appear as each other’s top partner concerning both positive and negative first-order sensitivity analyses of the targeted margin. This suggests that these two features function as a jointly diagnostic unit for the neural network’s most severe classification, rather than as independently informative signals. Overall, this model appears to treat its cardiotocographic event features as compounding rather than redundant. This is consistent with how CTG traces are read clinically where the co-occurrence of multiple concerning features is interpreted as worse than any single feature in isolation [6].

In contrast to the Maternal Health Risk dataset, where every coupled pair was found to be dampening, all four couplings identified for the Thyroid Disease dataset are reinforcing. The average Hessian values range from +2080 to +3492.

Table 8. Primary driver and coupled feature pairs by dataset and margin type, with interaction type and average Hessian value.

Dataset	1st Order Partial Derivative Type	Primary Driver (x_a)	Coupled Feature (x_b)	Interaction	Avg. H_{ab}
Maternal Health Risk (general margin)	+ve	Age	BS	Dampening	-3.64
	-ve	BS	SystolicBP	Dampening	-127.5
Maternal Health Risk (targeted margin)	+ve	DiastolicBP	BodyTemp	Dampening	-0.62
	-ve	BS	SystolicBP	Dampening	-191.3
Fetal Health (general margin)	+ve	accelerations	severe_decelerations	Reinforcing	+81.9
	-ve	severe_decelerations	baseline value	Reinforcing	+42.81
Fetal Health (targeted margin)	+ve	accelerations	prolongued_decelerations	Reinforcing	+113
	-ve	prolongued_decelerations	accelerations	Reinforcing	+93.5
Thyroid Disease (general margin)	+ve	TSH	T3	Reinforcing	+3492.5
	-ve	TSH	TT4	Reinforcing	+2080.1
Thyroid Disease (targeted margin)	+ve	T3	TSH	Reinforcing	+3178.7
	-ve	TSH	T3	Reinforcing	+2353.8

These magnitudes are substantially larger than those recorded for either of the previous datasets. This indicates a markedly stronger curvature in the learned decision surface.

The identity of the coupled pairs is notable. Under the general margin, TSH is the primary driver on both tracks, coupled with T3 on the positive track and with TT4 on the negative track. Under the targeted margin, the pairing becomes mutual, with T3 driving and TSH coupled on the positive track and the reverse on the negative track. TSH, T3 and TT4 are precisely the assays through which thyroid function is clinically assessed, and they are linked through the hypothalamic-pituitary-thyroid feedback axis [3]. So, the model has evidently learned to weigh them jointly rather than independently. This alignment between the recovered interaction structure and established clinical practice is a useful validation signal for the method. It must be emphasised, however, that these couplings describe the behaviour learned by the neural network from this dataset. They are not offered as confirmation of any physiological mechanism.

5 Comparison with State-of-the-Art Gradient-Based Methods

To validate our algorithm according to the existing literature, it was compared against two gradient-based methods, each chosen to isolate a distinct component of the claim of this framework. The Contrastive Explanation Method (CEM) [9] finds a pertinent negative through a projected fast iterative shrinkage-thresholding optimisation. It is notable here because its objective is a hinge loss on the gap between the predicted class logit and the strongest competing logit. This is structurally the same scalar that this study terms as the decision margin. It therefore tests whether the presented algorithm’s advantage over a comparably principled first-order optimiser comes specifically from its use of second-order interaction information, rather than from the margin formulation alone. Integrated Hessians [12] is an axiomatic and path-integrated second-order attribution method. It, therefore, tests the interaction-detection component directly rather than the recourse component.

5.1 Experimental Protocol

Both methods explain the *same* trained network on each dataset, under the same 80/20 stratified split and the same random seed. This has been done so that any difference in outcome is attributable to the explanation method, rather than to the model. Both CEM and Integrated Hessians require a reference or a baseline point. The per-feature median of the *training* split was used throughout. The justification is that clinical measurements have no natural origin denoting absence. For example, a serum TSH of zero does not represent an absent assay but a non-survivable state, so the zero baseline, conventionally used, for image data is meaningless in this setting [4]. The training median is the closest tabular analogue of an unremarkable patient. It is class-agnostic and lies inside the observed data manifold. Any deviation from it is precisely what carries diagnostic signal. Restricting the median to the training split prevents test information from leaking into the explanation.

A common currency is also required. Our algorithm reports a percentage change in a single driver feature whereas CEM moves several features simultaneously. So, comparing the algorithm’s single-feature figure directly against a multi-feature method would flatter our algorithm for reasons unrelated to second-order information. Both the total relative change ($L1_{rel}$) and the absolute movement ($L1_{abs}$) in the scaled space are therefore reported alongside validity and sparsity, as shown in Table 9.

5.2 Comparison Metrics

Four quantities (Table 9) are reported for each method and dataset, since a method strong on one measure can conceal a weakness on another.

1. Validity is the proportion of test instances for which the method found a perturbation which flips the class. It indicates only whether a counterfactual was found, not its size, so it must be read alongside the other three metrics.
2. $L1_{rel}$ (Total Relative Change) is $\sum_i |x'_i - x_i|/|x_i|$, summed over altered features. Expressing change as a proportion of each feature’s starting value allows features on different scales to be compared. This generalises the presented algorithm’s own percentage-change figure.
3. $L1_{abs}$ (Total Absolute Change) is the summed absolute change in the normalised input space. It is reported alongside $L1_{rel}$ because a feature with a small starting value can register a large relative change from a tiny absolute movement. $L1_{abs}$ protects against this distortion.
4. Sparsity is the mean number of features altered. Lower sparsity means simpler and more clinically actionable advice.

Overall, validity shows whether a method solves an instance. Sparsity shows how many features it acts upon, and $L1_{rel}/L1_{abs}$ show how far those features moved. High validity achieved by moving many features a long way is not better recourse than resolving fewer instances with one small and targeted change.

5.3 Counterfactual Recourse

Table 9 presents the recourse comparison. Our algorithm produces markedly more frugal counterfactuals than CEM on every dataset. In terms of total relative change, it is between 9.8 and 23.5 times smaller. In absolute terms, it is between 7.1 and 13.9 times smaller [4]. It also alters 2 features by construction, against roughly 5 for CEM. It is similar to a recourse that a clinician can state in a single sentence, namely adjust one measurement while accounting for one coupled measurement. This is materially more useful at the bedside than one requiring simultaneous movement in five variables.

This advantage does come with a cost, though. Our algorithm achieves lower validity than CEM on every dataset. The cause is structural rather than incidental. The presented algorithm deliberately confines itself to sweeping a single driver feature, with one coupled feature held at a fixed offset, and bounds that sweep at three times the original magnitude. When no single feature can carry the instance across the decision boundary within those limits, our algorithm correctly reports that no counterfactual was found, whereas a method, which is free to move every feature at once, will almost always succeed.

Therefore, it can be said that our algorithm explains fewer instances, but explains them far more economically. The preferred behaviour is a deployment decision rather than a purely technical one. It is also worth noting that CEM’s advantage in validity is achieved by producing, by far, the largest perturbations of the two methods compared, on all the datasets. Since CEM’s elastic-net penalty controls sparsity through a single coefficient, this reflects the particular operating point chosen rather than a fixed property of the method. A more aggressive sparsity setting would trade validity back for parsimony.

Table 9. Comparison of our algorithm against the Contrastive Explanation Method (CEM) on all three datasets.

Dataset	Method	Validity	$L1_{rel}$	$L1_{abs}$	Sparsity
Maternal Health Risk	Our Algorithm	85.3%	60.08%	0.232	2.00
	CEM	92.0%	696.27%	1.650	5.25
Fetal Health (CTG)	Our Algorithm	60.0%	88.83%	0.138	2.00
	CEM	100.0%	2088.27%	1.919	4.91
Thyroid Disease	Our Algorithm	70.0%	78.33%	0.112	2.00
	CEM	96.0%	766.20%	1.485	5.17

5.4 Interaction Structure

Table 10 compares the interaction structure recovered by the Hessians of our algorithm against Integrated Hessians. Integrated Hessians produces interaction scores but no recourse, so this comparison is confined to interaction detection only. ρ is the Spearman rank correlation between the two methods’ pairwise interaction magnitudes across all feature pairs [4]. Rank correlation between the two methods is positive on all three datasets with ρ ranging from 0.434 to 0.825. It reaches significance on Thyroid Disease ($p = 1.5 \times 10^{-4}$) and Fetal Health ($p = 0.049$) and remains marginal on Maternal Health Risk ($p = 0.056$). Two independent second-order methods, thus, agree broadly on which feature pairs are strongly coupled. This corroborates the interactions reported earlier in this study.

The agreement is imperfect in an instructive way. The single highest-ranked pair differs on every dataset, even where ρ is strongest. On the Thyroid Disease ($\rho = 0.825$), only one of the top three pairs is shared, with our algorithm ranking T3-TSH highest and Integrated Hessians ranking TSH-TT4 highest. This follows from what each method measures. The Hessian (from the presented algorithm) is local curvature at the instance itself while Integrated Hessians accumulates curvature along the path from the median reference to the instance. So, broad agreement with divergence at the top is expected while comparing a point estimate to a path integral. Notably, the disagreeing pairs on Thyroid Disease all fall within the same TSH/T3/TT4 panel. So, the divergence concerns ordering within a clinically related set of assays rather than a different mechanism.

Overall, our algorithm does not dominate CEM on every axis and should not be presented as doing so. It offers substantially more economical recourse at the cost of resolving fewer instances. The interaction structure, it recovers, is broadly corroborated by an independent axiomatic method. It differs mainly in fine-grained ordering for reasons that follow from the local nature of the pointwise Hessian.

Table 10. Agreement between the interaction structure recovered by our algorithm’s Hessian and by integrated Hessians, both computed on the same scalar decision margin with the training-set feature medians as the reference point.

Dataset	ρ	p	Our Algorithm’s top pair	IH top pair	Top-3 overlap
Maternal Health Risk	0.504	0.056	BS, SystolicBP	Age, BS	3/3
Fetal Health (CTG)	0.434	0.049	accelerations, prolonged_decel.	accelerations, baseline value	2/3
Thyroid Disease	0.825	1.5×10^{-4}	T3, TSH	TSH, TT4	1/3

6 Conclusion

This study extended a prior algorithm [22] from binary to multi-class classification through a single conceptual device, that is, a scalar decision margin. This margin reduces an arbitrary number of class logits to the one quantity the original first- and second-order sensitivity analysis was built to operate on. Applied to maternal health risk stratification, fetal health classification from cardiotocography, and thyroid disease classification, the extended framework consistently recovered the central signature of [22]. By pre-perturbing the feature most strongly coupled to the primary driver based on partial derivatives, our algorithm was able to reduce the change required in that driver to switch the predicted class, under both the general and clinically targeted margin. The effect was strongest on the thyroid dataset, where the required change more than halved under both margins.

Comparison against two gradient-based baselines explores the effectiveness of this algorithm further. Against the Contrastive Explanation Method [9, 4], the presented algorithm produced substantially more parsimonious recourse, altering roughly two features against five and requiring an order of magnitude less total change, but at the cost of lower validity. This happened because its bounded single-feature sweep cannot always resolve an instance where an unconstrained multi-feature search can. Against Integrated Hessians [12], the recovered interaction structure showed positive rank correlation on all three datasets, corroborating the interaction findings in aggregate, even though the single top-ranked pair differed for every dataset. This is a divergence attributable to comparing a local Hessian against a path-integrated one, rather than to either method being unreliable.

These findings should be read with caveats. All three datasets are modest in size and class-imbalanced, so minority-class interactions carry greater uncertainty. The generalisation beyond these populations is not established. The recovered interactions are properties of the trained model’s decision surface, not confirmed physiological mechanisms [14]. The margin formulation treats risk categories as unordered despite their natural ordinal structure. The underlying recourse can also propose changes which are not proximate, sparse, or feasible for the individual concerned [17], particularly for immutable or single-timepoint features. Perturbing one feature in isolation risks physiologically implausible profiles [14, 25]. Thus, accordingly, the presented framework is intended to support, not replace, clinical judgement. It is indeed open to follow-up research which extends this work to further clinical domains.

Declaration on Generative AI. According to the rules of the FTC 2026, it is acknowledged that Anthropic Claude was used to generate some of the text in this study. The citation for Claude has been added to the sections where any included content was generated by it.

References

1. Akmal, H., Hardalaç, F., Ayturan, K.: A fetal well-being diagnostic method based on cardiotocographic morphological pattern utilizing autoencoder and recursive feature elimination. *Diagnostics* 13(11), 1931 (june 2023)
2. Alshayeji, M.H.: Early thyroid risk prediction by data mining and ensemble classifiers. *Machine Learning and Knowledge Extraction* 5(3), 1195–1213 (2023)
3. Ambasta, A., Grenet, G., Otte, J., Thompson, W., Bassett, K., Perry, T.: Thyroid testing in primary hypothyroidism. *African journal of primary health care & family medicine* 17(1), e1–e4 (june 2025)
4. Anthropic: Claude. Large language model, <https://www.anthropic.com/claude> (2026), accessed: 2026. Version: Claude Opus 4.8
5. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barabado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, 82–115 (2020)
6. Brown, J., Kanagaretnam, D., Zen, M.: Clinical practice guidelines for intra-partum cardiotocography interpretation: A systematic review. *The Australian & New Zealand journal of obstetrics & gynaecology* 63(3), 278–289 (Mar 2023)
7. Campos, D.A.d., Bernardes, J.: Cardiotocography data set. <https://archive.ics.uci.edu/dataset/193/cardiotocography> (2010), accessed 2026
8. Ayres-de Campos, D., Bernardes, J.a., Garrido, A., Marques-de Sá, J., Pereira-Leite, L.: SisPorto 2.0: A program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal Medicine* 9(5), 311–318 (2000)
9. Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K., Das, P.: Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 31, pp. 590–601 (2018)

10. Erkamp, J.S., Geurtsen, M.L., Duijts, L., Reiss, I.K.M., Mulders, A.G.M.G.J., Steegers, E.A.P., Gaillard, R., Jaddoe, V.W.V.: Associations of maternal early-pregnancy glucose concentrations with placental hemodynamics, blood pressure, and gestational hypertensive disorders. *American Journal of Hypertension* 33(7), 660–669 (Apr 2020)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: *International Conference on Learning Representations (ICLR)* (2015), <https://arxiv.org/abs/1412.6572>
12. Janizek, J.D., Sturmfels, P., Lee, S.I.: Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research* 22(104), 1–54 (2021), <http://jmlr.org/papers/v22/20-1223.html>
13. Jhee, J.H., Lee, S., Park, Y., Lee, S.E., Kim, Y.A., Kang, S.W., Kwon, J.Y., Park, J.T.: Prediction model development of late-onset preeclampsia using machine learning-based methods. *PLOS ONE* 14(8), e0221202 (2019)
14. Kenny, E.M., Keane, M.T.: On generating plausible counterfactual and semi-factual explanations for deep learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 11575–11585 (2021)
15. Khadidos, A.O., Saleem, F., Selvarajan, S., Ullah, Z., Khadidos, A.O.: Ensemble machine learning framework for predicting maternal health risk during pregnancy. *Scientific Reports* 14(1), 21483 (sept 2024)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 30, pp. 4765–4774 (2017)
17. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*. pp. 607–617. Association for Computing Machinery, New York, NY, USA (2020)
18. Quinlan, J.R.: Thyroid disease. UCI Machine Learning Repository (1987), <https://doi.org/10.24432/C5D010>, accessed 2026
19. Ribeiro, M.T., Singh, S., Guestrin, C.: “why should I trust you?”: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1135–1144 (2016)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128(2), 336–359 (2020)
21. Shahid, Z., Rahulamathavan, Y., Dogan, S.: 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)*. pp. 193–200 (2024)
22. Shahid, Z., Rahulamathavan, Y., Dogan, S.: Chord: A novel xai framework for guiding counterfactuals via higher-order sensitivity analysis. In: *2026 IEEE Conference on Artificial Intelligence (CAI)*. pp. 1817–1822. IEEE (2026)
23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)* 70, 3319–3328 (2017)
24. UCI Machine Learning Repository: Maternal health risk data set. <https://archive.ics.uci.edu/dataset/863/maternal+health+risk> (2020), donated by Marzia Ahmed. Accessed 2026
25. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31(2), 841–887 (2017)

26. Werr, E.: Thyroid disease data. Kaggle (2022), <https://www.kaggle.com/datasets/emmanuelwerr/thyroid-disease-data>, accessed 2026