QUCE: The Minimisation and Quantification of Path-Based Uncertainty for Generative Counterfactual Explanations

DUELL, Jamie, SEISENBERGER, Monika, FU, Hsuan and FAN, Xiuyi

QUCE: The Minimisation and Quantification of Path-Based Uncertainty for Generative Counterfactual Explanations

Jamie Duell*, Monika Seisenberger†, Hsuan Fu‡, Xiuyi Fan*α
*Nanyang Technological University, Lee Kong Chian School of Medicine
† Swansea University, School of Mathematics and Computer Science
‡ Université Laval, Department of Finance, Insurance and Real Estate

***Abstract*—Deep Neural Networks (DNNs) stand out as one of the most prominent approaches within the Machine Learning (ML) domain. The efficacy of DNNs has surged alongside recent increases in computational capacity, allowing these approaches to scale to significant complexities for addressing predictive challenges in big data. However, as the complexity of the DNN models increases, interpretability diminishes. In response to this challenge, explainable models such as Adversarial Gradient Integration (AGI) leverage path-based gradients provided by DNNs to elucidate their decisions. Yet, the performance of path-based explainers can be compromised when gradients exhibit irregularities during out-of-distribution path traversal. In this context, we introduce Quantified Uncertainty Counterfactual Explanations (QUCE), a method designed to mitigate out-of-distribution traversal by minimizing path uncertainty. QUCE not only quantifies uncertainty when presenting explanations but also generates more certain counterfactual examples. We showcase the performance of the QUCE method by comparing it with competing methods for both path-based explanations and generative counterfactual examples. The code repository for the QUCE method is available at: https://github.com/jamie-duell/QUCE_ICDM.**

***Index Terms*—Explainable Artificial Intelligence, Deep Learning, Counterfactual, Uncertainty**

I. Introduction

The Path-Integrated Gradients (Path-IG) [1] formulation presents axiomatic properties that are upheld solely by path-based explanation methods. The Out-of-Distribution (OoD) problem is prevalent in the application of path-based explanation methods [2]; here the intuition is that traveling along a straight line path can incur irregular gradients and thus provide noisy attribution values [3]. Another known limitation of many Path-Integrated Gradient based approaches is the selection of a baseline reference; thus the Adversarial Gradient Integration (AGI) [4] method relaxes this constraint by generating baselines in adversarial classes. We note that AGI utilizes the path-based approach for generating counterfactual examples, and for this reason will be a primary baseline for our proposed method throughout this paper.

Counterfactual explanations [5] are often presented in the form of counterfactual examples [6]; here the goal is to provide a counterfactual example belonging to an alternative class with respect to a reference example. Counterfactual approaches aim to answer the question:

> "Given an instance, what changes can be made to change the outcome for that instance?"

Naturally, this allows for empirical observation as to which changes could provide an alternative outcome. The argument for using counterfactual methods is often developed from a causal lens [7]. It follows that to better evaluate this causal relationship, a promising avenue is to unify feature attribution with counterfactual examples, as demonstrated by the Diverse Counterfactual Explanations (DiCE) [6] method. Naturally, given quantitative approaches to feature attribution calculation such as these, ideally feature attribution methods should adhere to desirable axioms across XAI literature [8]. Thus, we aim to utilize state-of-the-art feature attribution assignment as to satisfy key axioms in our model development. Another concern with counterfactual examples, is the production of realistic paths to successfully create a counterfactual example; therefore we shall be exploring uncertainty.

Uncertainty quantification is not often considered when producing explanations, although some approaches have explored this. Autoencoder-based frameworks have been used to measure uncertainty for machine learning predictions and explanations [9]. The standard autoencoder approach evaluates the reconstruction error, which is often utilized in work surrounding anomaly detection [10]; instead, we explore the use of a variational autoencoder (VAE) for variational inference, and thus investigate counterfactuals generated with respect to the approximation of the true data distribution.

To address the above constraints, we propose the Quantified Uncertainty (Path-Based) Counterfactual Explanations (QUCE) method. The focus of the proposed method is threefold. We aim to

- minimize uncertainty and thus maximize the extent to which the generated paths and counterfactual examples are within distribution;
- relax the straight-line path constraints of Integrated Gradients;
- provide uncertainty quantification for counterfactual paths and counterfactual feature attribution.

This research is supported by LKCMedicine Start up grant funding from Ministry of Education Singapore.

αCorresponding Author. xyfan@ntu.edu.sg

In this work, we focus on the minimization of uncertain paths for counterfactual generation with quantifiable uncertainty measures on the generated counterfactual. QUCE's learning process relaxes IG's straight-line path restrictions as part of the generative process.

Intuitively, it is unclear in many scenarios if one single best path toward an alternative outcome exists; for example a patient's treatment path may be unclear [11], or there may be many viable paths to achieve the same outcome [12]. Therefore, QUCE utilises both a single and multiple-paths approach. We present the optimisation over the key metrics – *proximity* [5], *validity* [5] and *uncertainty* [13].

II. Explainable AI and Counterfactuals

In the work of [8] the authors present a desirable set of axiomatic foundations for XAI methods. As a brief informal overview, we consider the following axioms:

- **Success:** The explainer method should be able to produce explanations for any instance.
- **Explainability:** An explanation method should provide informative explanations. An empty explanation here is not recommended.
- **Irreducability:** An explanation should not contain irrelevant information.
- **Representativity:** An explanation should be possible on unseen instances.
- **Relevance:** Information should only be included if it impacts the prediction.

Counterfactual explanations can be presented both in the form of counterfactual examples and counterfactual feature attribution [6].

Definition 1 (Counterfactual Example)**.** *Given a probabilistic classifier $f : \mathbb{R}^J \to \{0, 1\}$, differentiable probabilistic function $F : \mathbb{R}^J \to [0, 1]$, and class $\tau = \{0, 1\}$ an instance $\mathbf{x} = \langle x^1, \ldots, x^J \rangle \in X$ where $X \in \mathbb{R}^{N \times J}$ and a classification threshold $\vartheta \in [0, 1]$ such that*

$$f(\mathbf{x}) = \begin{cases} \neg\tau, \textit{if } F(\mathbf{x}) \leq \vartheta \\ \tau, \textit{otherwise.} \end{cases} \tag{1}$$

Then a counterfactual example of $\mathbf{x}$ is some $\mathbf{x}^c$ where $f(\mathbf{x}^c) \neq f(\mathbf{x})$.

Counterfactual examples are often produced through the use of a learned generative function, examples of this can be seen in the work of [14]. Extending this, we define a *counterfactual generator*.

Definition 2 (Counterfactual Generator)**.** *Given an instance $\mathbf{x} \in X$ and a classifier f, a counterfactual generator is a function $\mathcal{G} : \mathbb{R}^J \to \mathbb{R}^J$ that takes an instance $\mathbf{x}$ and returns a counterfactual example $\mathbf{x}^c \notin X$.*

Feature attribution is a common form of XAI, and the feature attribution approach is seen in many methods [15], [16]. We continue by defining *feature attribution*:

Definition 3 (Feature Attribution)**.** *Given an instance $\mathbf{x}$, a feature attribution method is a function $\Phi : \mathbb{R}^J \to \mathbb{R}^J$, where Φ takes an instance $\mathbf{x}$ and returns a vector of feature attribution values $\langle \phi^1, \ldots, \phi^J \rangle$.*

Concatenating feature attribution methods and counterfactual examples leads to *counterfactual feature attribution* (CFA) which is defined as follows:

Definition 4 (Counterfactual Feature Attribution)**.** *A counterfactual feature attribution method is a function $\Phi_{CF} : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}^J$ that takes an instance $\mathbf{x}$ and a counterfactual example $\mathbf{x}^c$ and returns a vector of counterfactual feature attribution values $\langle \phi^1_{CF}, \ldots, \phi^J_{CF} \rangle$.*

III. Proposed Model: QUCE

A. Generating Counterfactuals with QUCE

To generate counterfactuals we propose a three-part objective function with a composite weighting vector $\Lambda = \langle \lambda_1, \lambda_2, \lambda_3 \rangle$, where each $\lambda \in \Lambda$ is an independent tolerance (weight) used to determine the influence of each part of the joint objective function presented in equation 2. By minimizing the objective function, we obtain the generated counterfactual $\mathbf{x}^c$. Informally, our objective function is composed of three parts:

- $\mathcal{L}_{pr}$, for the maximization the probability towards the desired class;
- $\mathcal{L}_{\delta}$, to minimize the distance between the instance and a generated counterfactual;
- $\mathcal{L}_{\epsilon}$, to minimize the uncertainty of both the generated paths and generated counterfactual examples.

Combining these terms with our weighting vector, we have

$$\mathcal{G}(\mathbf{x}) = \arg\min_{\mathbf{x}^c} \ \lambda_1 \mathcal{L}_{pr} + \lambda_2 \mathcal{L}_{\delta} + \lambda_3 \mathcal{L}_{\epsilon}. \tag{2}$$

Having constructed our generative objective function, we provide further notation to illustrate the learning process. First, we consider an iterative learning process such as gradient descent on $\mathbf{x}$ to produce a path from $\mathbf{x}$ to a generated $\mathbf{x}^c$. Thus, $\mathcal{G}(\mathbf{x})$ is minimized via gradient descent. We initially let $\mathbf{x}^c = \mathbf{x}$; $\mathbf{x}^c$ is updated via

$$\begin{aligned} &\mathbf{x}^c \Leftarrow \mathbf{x}_{\Delta_i}, \\ &\Delta_i = \varphi \nabla_{\mathbf{x}^c}(\mathcal{G}(\mathbf{x})), \\ &\mathbf{x}_{\Delta_i} = \mathbf{x}^c - \Delta_i. \end{aligned}$$

Let $\mathbf{x}$ be updated on a loop over $i (0 \leq i \leq n)$ iterations; when $i = n$ we let our $\mathbf{x}^c$ indicating our generated counterfactual. Here, φ represents the "learning rate," a small positive multiplier value $\varphi \in [0, 1] : \varphi << 1$. We store each update on $\mathbf{x}^c$ as a vector $\mathbf{x}^{\Delta} = \langle \mathbf{x}_{\Delta_0}, \ldots, \mathbf{x}_{\Delta_n} \rangle$.

B. Finding Counterfactuals

1) Valid Counterfactuals: A key concept in finding counterfactual examples is ensuring that the counterfactual is indeed *valid*, and thus we aim to produce counterfactual examples that belong to a counterfactual class.

Remark 1. *For the sake of simplicity we let $f(\mathbf{x}) = \neg\tau$ throughout the remainder of this work; thus τ is the counterfactual class for which we aim to generate some $\mathbf{x}^c$ such that $f(\mathbf{x}^c) = \tau$.*

In light of remark 1, given a target counterfactual class τ and probabilistic decision threshold ϑ we aim to find an instance $\mathbf{x}^c$ that satisfies

$$F(\tau|\mathbf{x}^c) \geq \vartheta. \tag{3}$$

Where ϑ is a probability threshold for the target class τ. Thus, we need a generator $\mathcal{G}$ that satisfies the condition in equation 3. To achieve this, we minimize the negative log-likelihood:

$$\mathcal{L}_{pr} = \Big[- log[F(\tau|\mathbf{x}^c)] \Big].$$

This constitutes the constrained optimisation problem with respect to some target class τ.

2) Proximity for Counterfactuals: Given an instance, in the production of counterfactual examples we often aim to find a counterfactual example that is "similar" in feature space to the instance. This is often termed *proximity*.

In this work, we use the l_2 norm as the proposed model focuses on producing counterfactuals from continuous features, defining proximity as follows:

Definition 5 (Proximity)**.** *Given an instance $\mathbf{x}$ and its counterfactual example $\mathbf{x}^c$, the proximity between the two instances is given by*

$$\mathcal{L}_{\delta} = \left[\frac{1}{2} ||\mathbf{x}^c - \mathbf{x}||^2 \right]. \tag{4}$$

3) Minimally Uncertain Counterfactuals: To maximize the certainty of counterfactual examples, we examine their complement—namely, the uncertainty associated with a counterfactual example. To explore this, we establish the concept of *counterfactual uncertainty*. Informally, we consider uncertainty to be the Evidence Lower Bound (ELBO) as measured by a Variational Autoencoder (VAE) framework. The objective function ELBO is comprised of two components, namely the Kullbeck–Leibler (KL) divergence and a reconstruction loss. This is defines the VAE loss (VAEL) as the following:

$$\text{VAEL}(\mathbf{x}) = \mathbb{E}_{q_\theta}[log\ q_\theta(\mathbf{z}|\mathbf{x}) - log\ p_\psi(\mathbf{z})] - \mathbb{E}_{q_\theta}\ log\ p_\psi(\mathbf{x}|\mathbf{z}),$$

where p and q are probability distributions and $\mathbf{z}$ is the latent representation of $\mathbf{x}$. The aim is to find a θ that successfully models the true training data distribution ψ, and thereby satisfying the following minimization problem:

$$\{\theta^*, \psi^*\} = \arg\min_{\theta,\psi} \text{VAEL}(\mathbf{x}). \tag{5}$$

Posterior to training the VAE we have a fixed representation shaping our distributions with the θ^* and ψ^* parameters, and thus we can provide counterfactual uncertainty as:

Definition 6 (Counterfactual Uncertainty)**.** *Given the fixed parameters θ^* and ψ^*, counterfactual uncertainty is given by:*

$$\mathcal{L}_{\epsilon} = \mathbb{E}_{q_{\theta^*}}[log\ q_{\theta^*}(\mathbf{z}|\mathbf{x}^c) - log\ p_{\psi^*}(\mathbf{z})] - \mathbb{E}_{q_{\theta^*}}\ log\ p_{\psi^*}(\mathbf{x}^c|\mathbf{z}).$$

C. Uncertainty in Counterfactual Explanations

Definition 6 determines how "good" the fit of the new instance is with respect to the training data distribution, and similarly how well a path fits into the data distribution. From definition 6, we have a quantifiable measure of uncertainty for the generated counterfactual $\mathbf{x}^c$. We define *Feature-wise Counterfactual Uncertainty* as follows:

Definition 7 (Feature-wise Counterfactual Uncertainty)**.** *The Feature-wise Counterfactual Uncertainty is given by:*

$$\epsilon_{\mathbf{d}} = \langle |d^1|, \ldots, |d^J| \rangle \ : \ \langle d^1, \ldots, d^J \rangle = \mathbf{d}; \tag{6}$$

$$\textit{where} \quad \mathbf{d} = \mathbf{x}^c - \hat{\mathbf{x}}^c, \tag{7}$$

where $\hat{\mathbf{x}}^c$ is the reconstruction of $\mathbf{x}^c$.

With this representation, we can then successfully update $\mathbf{x}^c$ by both adding and subtracting this vector of feature-wise counterfactual uncertainty as given by the reconstruction error, and thus we can calculate *Counterfactual Explanation Uncertainty*.

Definition 8 (Counterfactual Explanation Uncertainty)**.** *Given a CFA Φ_{CF} and the feature-wise counterfactual uncertainty $\epsilon_{\mathbf{d}}$, the counterfactual explanation uncertainty is given by*

$$\Phi_{CF}^{\epsilon_{\mathbf{d}}} = \Phi_{CF}(\mathbf{x}^c \pm \epsilon_{\mathbf{d}}|\mathbf{x}). \tag{8}$$

D. Path Explanations

The Path-Integrated Gradients [1] formulation is the only approach to our knowledge within the landscape of feature attribution methods that satisfies all the axioms in presented in [1]. Therefore, we adopt the path-integral formulation and relax the straight-line constraint present in IG. To achieve this, recall the set of learned updates on $\mathbf{x}$, namely $\mathbf{x}^{\Delta}$. It follows that we can produce explanations over $\mathbf{x}^{\Delta}$.

Formally, let the function F be a continuously differentiable function, the QUCE explanation takes the path integral formulation such that given a smooth function $\psi = \langle \psi^1, \ldots, \psi^J \rangle : [0, 1] \rightarrow \mathbb{R}^J$ defining a path in $\mathbb{R}^J$, where $\psi(\alpha)$ is a point along a path at $\alpha \in [0, 1]$ with $\psi(0) = \mathbf{x}_{\Delta_0}$ and $\psi(1) = \mathbf{x}_{\Delta_n}$, the single-path QUCE explainer is defined as:

$$\Phi_{\text{QUCE}}(\mathbf{x}^{\Delta}) := \int_{\mathbf{x}_{\Delta_0}}^{\mathbf{x}_{\Delta_n}} \nabla F(\psi(\alpha)) \cdot \psi'(\alpha) d\alpha. \tag{9}$$

It follows that the explanation uncertainty with respect to a single generated counterfactual $\mathbf{x}^c$ is given as

$$\Phi_{\text{QUCE}}^{\pm\epsilon_{\mathbf{d}}}(\mathbf{x}^c) := \int_{\mathbf{x}^c}^{\mathbf{x}^c \pm \epsilon_{\mathbf{d}}} \nabla F(\psi(\alpha)) \cdot \psi'(\alpha) d\alpha. \tag{10}$$

QUCE explanations can be easily computed by following the Riemann approximation of the integrals for each feature as defined in [1]. By taking the Riemann approximation of the explanation between points $\mathbf{x}_{\Delta i}$ and $\mathbf{x}_{\Delta i-1}$, we can bound the explanation to an error σ which we illustrate in proposition 1, following a construction similar to the proof of theorem 1 in [17].

Assumption 1. *∇F is monotonic along a path that is parameterised by ψ.*

Proposition 1. *The error σ for an explanation associated with two points $x^j_{\Delta i}$ and $x^j_{\Delta i-1}$, is bound by*

$$\sigma \leq |\mathcal{R}_P - \mathcal{L}_P|.$$

where $\mathcal{R}_P$ is the upper-bound and $\mathcal{L}_P$ is the lower-bound.

Proof. Under assumption 1, we can consider the upper-bound of the explanation to be given by the right Riemann sum ($\mathcal{R}_P$), for P steps in the Riemann approximation, such that:

$$\mathcal{R}_P = \frac{||x^j_{\Delta_i} - x^j_{\Delta_{i-1}}||}{P} \times \sum_{p=1}^{P} \frac{\partial F(\mathbf{x}_{\Delta_{i-1}} + \frac{p}{P}(\mathbf{x}_{\Delta_i} - \mathbf{x}_{\Delta_{i-1}}))}{\partial x^j}$$

$$\geq (x^j_{\Delta_i} - x^j_{\Delta_{i-1}}) \times \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}_{\Delta_{i-1}} + \alpha(\mathbf{x}_{\Delta_i} - \mathbf{x}_{\Delta_{i-1}}))}{\partial x^j} d\alpha$$

with a left Riemann sum ($\mathcal{L}_P$) giving a lower-bound, where:

$$\mathcal{L}_P = \frac{||x^j_{\Delta_i} - x^j_{\Delta_{i-1}}||}{P} \times \sum_{p=0}^{P-1} \frac{\partial F(\mathbf{x}_{\Delta_{i-1}} + \frac{p}{P}(\mathbf{x}_{\Delta_i} - \mathbf{x}_{\Delta_{i-1}}))}{\partial x^j}$$

$$\leq (x^j_{\Delta_i} - x^j_{\Delta_{i-1}}) \times \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}_{\Delta_{i-1}} + \alpha(\mathbf{x}_{\Delta_i} - \mathbf{x}_{\Delta_{i-1}}))}{\partial x^j} d\alpha$$

then the error σ for the explanation is bound by

$$\sigma \leq |\mathcal{R}_P - \mathcal{L}_P|.$$

□

Both attributed values from equation 10 illustrate uncertainty in feature attribution values given by the QUCE explainer.

Due to the potential stochastic nature of our model with potentially multiple minima (e.g. we may have two points equally "close" to the decision bound with different values), we consider a set of generated counterfactual examples to be given as $C = \langle \mathbf{x}^c_{1,1}, \ldots, \mathbf{x}^c_{1,k} \rangle$, where k is the number of generated counterfactual examples over some set $\mathbf{x}$. Given C, we can accumulate attribution over many counterfactuals by avoiding the specification of $\mathbf{x}^c$, so that we have:

$$\Phi_{\text{exQUCE}}(\mathbf{x}) := \int_{\mathbf{x}^c} \Big(\Phi_{\text{QUCE}}(\mathbf{x}^{\Delta}) \Big) p_C(\mathbf{x}^c) d\mathbf{x}^c \quad (11)$$

$$:= \mathop{\mathbb{E}}_{\mathbf{x}^c \sim C, \alpha \sim \mathcal{U}(0,1)} \Big[\Phi_{\text{QUCE}}(\mathbf{x}^{\Delta}) \Big]. \quad (12)$$

Here we let $\alpha \sim \mathcal{U}(0,1)$ indicate interpolation over α for n counterfactual steps in the generator function. Informally, we get the expectation of the gradients over the piecewise-linear path between counterfactual steps of the generator. We take a similar approach to the Expected Gradients [18] formulation, except we instead sample from a set of generative counterfactual examples. We make two arguments as to why we use this approach:

- In explaining a counterfactual outcome, we do not know the specific path taken and thus we can average over many paths.
- We can invert the path to explain $\mathbf{x}$ and therefore we can have many generative baselines. This relaxes the specified baseline of many existing path-based explanation methods.

To proceed, we show via proposition 2 that completeness[1] holds when working with the many-paths approach for expected values, ensuring axiomatic guarantees.

Proposition 2. *The QUCE method satisfies the following equality:*

$$\mathop{\mathbb{E}}_{\mathbf{x}^c \sim C} \Big[\Phi_{QUCE}(\mathbf{x}^{\Delta}) \Big] \quad (13)$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}^c \sim C} \Big[F(\mathbf{x}^c) - F(\mathbf{x}) \Big] \quad (14)$$

Proof. Due to the *completeness* axiom the following holds true:

$$F(\mathbf{x}_{\Delta_n}) - F(\mathbf{x}_{\Delta_0}) = \quad (15)$$

$$\int_{\mathbf{x}_{\Delta_0}}^{\mathbf{x}_{\Delta_n}} \nabla F(\psi(\alpha)) \cdot \psi'(\alpha) d\alpha \quad (16)$$

given $\mathbf{x}_{\Delta_n} = \mathbf{x}^c$ and $\mathbf{x}_{\Delta_0} = \mathbf{x}$ respectively, we have

$$F(\mathbf{x}^c) - F(\mathbf{x}).$$

By relaxing a strict definition of $\mathbf{x}^c$, where we instead use a set of generated counterfactuals C, yielding

$$\int_{\mathbf{x}^c} \left(\int_{\mathbf{x}}^{\mathbf{x}^c} \nabla F(\psi(\alpha)) \cdot \psi'(\alpha) d\alpha \right) p_C(\mathbf{x}^c) d\mathbf{x}^c \quad (17)$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}^c \sim C, \alpha \sim \mathcal{U}(0,1)} \Big[\nabla F(\psi(\alpha)) \cdot \psi'(\alpha) \Big] \quad (18)$$

and since $F(\mathbf{x})$ is a constant, we have:

$$\mathop{\mathbb{E}}_{\mathbf{x}^c \sim C} \Big[F(\mathbf{x}^c) \Big] - F(\mathbf{x}) = \mathop{\mathbb{E}}_{\mathbf{x}^c \sim C} \Big[F(\mathbf{x}^c) - F(\mathbf{x}) \Big], \quad (19)$$

equations 13 and 14 are equivalent. □

Given we can compute many-paths explanations, it follows that we can also take the expected gradients for the explanation uncertainty computed by QUCE along each path, such that

$$\Phi^{\pm\epsilon_{\mathbf{d}}}_{\text{exQUCE}}(\mathbf{x}^c) := \mathop{\mathbb{E}}_{\mathbf{x}^c \sim C} \Big[\Phi^{\pm\epsilon_{\mathbf{d}}}_{\text{QUCE}}(\mathbf{x}^c) \Big]. \quad (20)$$

We proceed to evaluate QUCE against the axioms in section II. We show that it is inherently straightforward to prove that our proposed QUCE method satisfies these desirable axioms.

Proposition 3. *The QUCE method satisfies success, explainability, irreducability, representativity and relevance.*

Proof. As a direct implication of the generative learning process, QUCE will achieve an explanation satisfying success. Next, explainability holds assuming that different instances

[1] *Completeness*: The difference in prediction between the baseline and input should be equal to the sum of feature attribution values.

generated by QUCE do not have the same prediction probability with respect to the target class. Furthermore, we characterize irrelevance under our own interpretation: since a feature that does not change does not affect the predicted outcome, it should be assigned zero attribution. Then directly from the definition of QUCE it is clear that irreducability holds, as the gradients are multiplied by a zero-value scalar for the same valued features. It is easy to see that any instance with the same dimensionality of the instances from a training dataset can utilize the QUCE approach, satisyfing representativity. Relevance holds as a direct implication of irreducability and the fact that explanations utilise the model gradients, thereby ensuring model-specific relevance.

□

IV. Quantitative Evaluation

A. Datasets

1) The Simulacrum: The Simulacrum[2] is a synthetic dataset used in this study, the Simulacrum is a large dataset developed by Health Data Insight CiC and derived from anonymous cancer data provided by the National Disease Registration Service, NHS England. We consider five subsets of patient records based on ICD-10 codes corresponding to lung cancer, breast cancer, skin cancer, lymphoma and rectal cancer. These datasets are organised as survival time classification problems, where patients are predicted a survival time of either at least 6 months or less than 6 months.

2) COVID Rate of Infection: The COVID rate of infection dataset contains details on control measures, temperature, humidity and the daily rate of infection for different regions of the UK. Details on data collection are provided in [19]. This dataset is a binary classification task identifying an increased rate of infection against a non-increased rate of infection.

3) Wisconsin Breast Cancer: The Wisconsin Breast Cancer (W-BC) [20] dataset, provided in the scikit-learn library[3], is a binary classification dataset that classifies malginant and benign tumours given a set of independent features from breast mass measurements.

B. Baselines

1) Diverse Counterfactual Explanations: DiCE [6], a counterfactual generator, provides feature attribution values for an instance with respect to its counterfactual examples. We use the DiCE method as a comparison for generating counterfactual examples, as DiCE is not a path-based explainer, we can only compare the generated counterfactuals.

2) Integrated Gradients: IG [1] produces explanations for instances in a given dataset. We modify IG in our experiments so that the baseline becomes the instance to be explained, while the target instance is the counterfactual generated by QUCE. This way, we can evaluate the straight-line path solution against the QUCE-generated path.

[2]https://simulacrum.healthdatainsight.org.uk/

[3]https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

3) Adversarial Gradient Integration: AGI [4] provides feature attribution values and generative counterfactual examples. The AGI method is the only path-based generative counterfactual method currently available to our knowledge and thus forms the primary focus of our comparison.

C. Results

To evaluate the QUCE method, we provide a comparison of uncertainty along a path. To do this, we use a pre-trained VAE, feeding all generated instances along any path into the VAE to determine the reconstruction error for all given instances along a path. The intuition behind this is that a smaller reconstruction error is associated with a path that better follows the data distribution and is therefore more "realistic". In Table I we observe that the QUCE method provides paths that better follow the data distribution when compared against both IG and AGI on average.

We also analyze the average VAE loss and reconstruction error per instance across all 100 instances on both the training and test datasets. This highlights the closeness of the reconstructed sample against the ground truth counterfactual generated by different methods. In Table II we observe that QUCE provides a lower value with respect to uncertainty measurement. In Table III we observe that our proposed QUCE method provides better reconstructed counterfactuals.

To compare counterfactual feature attribution methods we evaluate the deletion score, a common metric used for evaluating feature attribution methods for identifying important features. The deletion score is used in various studies [21], [22]; here, a lower value indicates better performance. In Table IV we observe that the QUCE method performs better on average than both DiCE and AGI for counterfactual feature attribution performance.

V. Conclusion

In this paper, we provide a novel approach that combines generative counterfactual methods and path-based explainers, minimizing uncertainty along generated paths and for generated counterfactual examples. We provide an analysis of the proposed QUCE method on path uncertainty, generative counterfactual example uncertainty, counterfactual reconstruction error and deletion score metrics. Our approach provides paths that are less uncertain in their interpolations, so that more reliable gradients and explanations can be extracted, to facilitate this, we also provide a clear explanation of uncertainty associated with assigned feature attribution values.

In whole we observe from the results, that not only do we provide more reliable gradients on the majority of datasets, more importantly we produce counterfactual examples and paths that conform better to the underlying data distribution.

References

[1] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *ICML*, 2017, p. 3319–3328.

[2] J. Duell, M. Seisenberger, T. Zhong, H. Fu, and X. Fan, "A formal introduction to batch-integrated gradients for temporal explanations," in *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, nov 2023, pp. 452–459.

Table I: Comparison of the average path uncertainty on the generated counterfactual instances. This is experimented over 100 instances from the training and testing sets of each dataset. Here we have 1000 steps (path interpolation instances) for the Riemann approximation of every path-based approach, thus effectively 100×1000 instances. Here, the lower value the better.

Path $\mathcal{L}_\epsilon$	Lung	Breast	Skin	Lymph	Rectal	COVID	W-BC
Train							
QUCE	**0.92±0.32**	**0.82±0.26**	**0.94±0.41**	**0.74±0.06**	**0.86±0.28**	**1.36±0.15**	**0.82±0.16**
IG-QUCE	0.95±0.32	0.84±0.27	0.96±0.41	0.76±0.06	0.86±0.29	1.39±0.11	0.84±0.20
AGI	1.94±1.86	1.49±0.95	1.80±1.47	0.92±0.28	2.19±2.23	2.01±0.15	0.93±0.36
Test							
QUCE	**0.82±0.34**	**0.91±0.28**	**0.82±0.31**	**0.69±0.19**	0.80±0.29	**0.61±0.07**	**0.83±0.29**
IG-QUCE	0.83±0.33	**0.91±0.28**	**0.82±0.31**	0.70±0.19	**0.79±0.29**	0.67±0.05	0.85±0.33
AGI	1.22±1.16	1.83±0.97	1.34±1.20	0.89±0.28	1.57±1.35	0.82±0.11	0.96±0.55

Table II: Comparison of the average VAE loss for generated counterfactual (CF) examples. This is experimented over 100 instances on each dataset. Here we observe that the proposed QUCE method performs best across all datasets.

CF $\mathcal{L}_\epsilon$	Lung	Breast	Skin	Lymph	Rectal	COVID	W-BC
Train							
QUCE	**1.01**	**0.78**	**0.97**	**0.71**	**0.85**	**1.25**	**0.93**
DiCE	1.97	1.02	1.48	1.10	1.27	1.57	3.63
AGI	2.93	2.07	2.61	1.01	3.48	2.40	1.22
Test							
QUCE	**0.93**	**0.91**	**0.86**	**0.67**	**0.83**	**0.76**	**1.08**
DiCE	1.95	1.09	1.33	1.18	1.22	0.92	3.00
AGI	1.68	2.75	1.87	1.02	2.38	0.84	1.41

Table III: Comparison of the average sum of feature-wise reconstruction error (CR) between original instances and their generated counterfactual examples. This is experimented on 100 instances for each dataset.

CR	Lung	Breast	Skin	Lymph	Rectal	COVID	W-BC
Train							
QUCE	**0.95**	**0.73**	**0.90**	**0.66**	**0.78**	**1.16**	**0.73**
DiCE	1.80	0.95	1.33	1.01	1.18	1.38	1.04
AGI	2.55	1.89	2.25	0.91	2.91	2.19	0.80
Test							
QUCE	**0.88**	**0.85**	**0.80**	**0.63**	**0.78**	**0.57**	**0.76**
DiCE	1.81	0.95	1.32	0.99	1.18	1.38	1.04
AGI	1.53	2.41	1.66	0.92	2.08	0.79	0.81

Table IV: Comparison of the deletion scores for counterfactual generative methods that provide feature attribution values (see definition 4). This is experimented over 100 instances on each test dataset. Here the lower the value the better.

Deletion	Lung	Breast	Skin	Lymph	Rectal	COVID	W-BC
QUCE	**0.556**	0.689	0.656	0.611	**0.669**	**0.699**	**0.632**
DiCE	0.561	0.688	**0.649**	0.619	**0.669**	0.710	0.637
AGI	0.559	**0.683**	0.659	**0.607**	0.670	0.728	0.648

[3] A. Kapishnikov, S. Venugopalan, B. Avci, B. Wedin, M. Terry, and T. Bolukbasi, "Guided integrated gradients: an adaptive path method for removing noise," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2021, pp. 5048–5056.

[4] D. Pan, X. Li, and D. Zhu, "Explaining deep neural network models with adversarial gradient integration," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 8 2021, pp. 2876–2883.

[5] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, Apr. 2022.

[6] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAccT*, 2020, p. 607–617.

[7] M. Höfler, "Causal inference based on counterfactuals," *BMC Medical Research Methodology*, vol. 5, no. 1, Sep. 2005.

[8] L. Amgoud and J. Ben-Naim, "Axiomatic foundations of explainability," in *IJCAI*, 2022, pp. 636–642.

[9] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a CLUE: A method for explaining uncertainty estimates," in *International Conference on Learning Representations*, 2021.

[10] F. Angiulli, F. Fassetti, and L. Ferragina, "Reconstruction error-based anomaly detection with few outlying examples," 2023.

[11] L. E. Beutler, K. Someah, S. Kimpara, and K. Miller, "Selecting the most appropriate treatment for each patient," *International Journal of Clinical and Health Psychology*, vol. 16, no. 1, p. 99–108, Jan. 2016.

[12] J. W. Bull, N. Strange, R. J. Smith, and A. Gordon, "Reconciling multiple counterfactuals when evaluating biodiversity conservation impact in social-ecological systems," *Conservation Biology*, vol. 35, no. 2, p. 510–521, Sep. 2020.

[13] A. Sagar, "Uncertainty quantification using variational inference for biomedical image segmentation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, jan 2022, pp. 44–51.

[14] J. Ma, R. Guo, S. Mishra, A. Zhang, and J. Li, "CLEAR: Generative counterfactual explanations on graphs," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 2022.

[15] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017, pp. 4765–4774.

[16] A. Dejl, H. Ayoobi, M. Williams, and F. Toni, "Cafe: Conflict-aware feature-wise explanations," 2023.

[17] J. D. Janizek, P. Sturmfels, and S.-I. Lee, "Explaining explanations: axiomatic feature interactions for deep networks," *J. Mach. Learn. Res.*, vol. 22, no. 1, jan 2021.

[18] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature Machine Intelligence*, p. 620–631, 2021.

[19] M. Kapcia, H. Eshkiki, J. Duell, X. Fan, S. Zhou, and B. Mora, "Exmed: An ai tool for experimenting explainable ai techniques on medical data analytics," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 841–845.

[20] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995.

[21] P. Yang, N. Akhtar, Z. Wen, and A. Mian, "Local path integration for attribution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 3173–3180, Jun. 2023.

[22] N. Akhtar and M. A. A. K. Jalwana, "Towards credible visual model interpretation with path attribution," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, 23–29 Jul 2023, pp. 439–457.