

**資訊工程系碩士班**

**碩士學位論文**

**結合 SHAP 與反事實解釋之生成對抗模型**

**The Adversarial Generative Model Combining SHAP and Counterfactual Explanations**

**研究生：陳勝誠**

**指導教授：郭忠義 博士**

**中華民國一百一十五年七月**

摘 要

關鍵詞： SHAP、 Counterfactual Explanation、Generative Adversarial Network

隨著人工智慧在醫療、金融與法律等領域的應用日益廣泛，模型的可解釋性（Explainability）變得日益重要。SHAP（SHapley Additive exPlanations）是一種廣泛應用於解釋黑箱模型的特徵貢獻方法，而生成對抗網路（GAN）則在資料合成與增強方面展現出強大潛力。ACGAN（Auxiliary Classifier GAN）進一步結合了分類器與生成器，使生成樣本同時具備指定類別特徵。SHAP-GAN 將 SHAP 嵌入 ACGAN 的訓練流程中，讓生成樣本在外觀上像真的資料，且在 SHAP 貢獻上也與真實樣本一致，提升資料的可解釋性與分類效能。然而，SHAP-GAN 仍聚焦於生成靜態的、屬於特定類別的樣本，對於探討模型決策邊界與反事實推理尚未著墨。反事實解釋（Counterfactual Explanation）可提供決策反饋：「若想改變模型預測，應調整哪些特徵？」將反事實生成與 SHAP 結合，可望產出具有一致性與邏輯合理性的樣本，進一步強化分類器學習效果。

ABSTRACT

Keywords：SHAP、 Counterfactual Explanation、Generative Adversarial Network

With the growing adoption of artificial intelligence in fields such as healthcare, finance, and law, model explainability has become increasingly critical. SHAP (SHapley Additive exPlanations) is a widely used method for interpreting the feature contributions of black-box models, while Generative Adversarial Networks (GANs) have demonstrated strong capabilities in data synthesis and augmentation.Auxiliary Classifier GAN (ACGAN) integrates both a classifier and generator to produce synthetic samples conditioned on specific class labels. SHAP-GAN embeds SHAP values into the ACGAN training process, enabling generated samples to not only resemble real data in appearance but also align with the SHAP contributions of real samples. This improves both interpretability and classification performance.However, SHAP-GAN primarily focuses on generating static samples of specific classes and does not address model decision boundaries or counterfactual reasoning. Counterfactual explanations provide actionable feedback by answering the question: “What features need to change to alter the model’s prediction?” Combining counterfactual generation with SHAP may yield semantically meaningful and logically coherent samples, further enhancing the classifier's learning.

目 錄

[摘 要 i](#_Toc205212585)

[ABSTRACT ii](#_Toc205212586)

[目 錄 iv](#_Toc205212587)

[第一章 緒論 1](#_Toc205212588)

[1.1 研究動機與目的 1](#_Toc205212589)

[第二章 文獻探討 2](#_Toc205212590)

[2.1 SHAP 2](#_Toc205212591)

[2.2 ACGAN 與 SHAP-GAN 2](#_Toc205212593)

[2.3 反事實解釋 2](#_Toc205212594)

[第三章 研究方法 3](#_Toc205212596)

[3.1 問題案例描述 3](#_Toc205212597)

[3.2方法簡介 3](#_Toc205212598)

[3. 3 問題與解決方法定義 3](#_Toc205212599)

[3. 4 解決方法架構 4](#_Toc205212600)

[3. 5 實驗方法 4](#_Toc205212601)

[第四章 實作與成果分析 5](#_Toc205212602)

[4.1 系統架構 5](#_Toc205212603)

[第五章 結論與未來研究方向 6](#_Toc205212604)

[5.1 結論 6](#_Toc205212605)

[5.2 未來研究方向 6](#_Toc205212606)

[參考文獻 7](#_Toc205212607)

第一章 緒論

## 1.1 研究動機與目的

在許多真實世界的分類任務中，例如醫療風險評估，高風險樣本往往佔比極低，導致訓練資料呈現嚴重不平衡。這樣的資料不平衡會造成模型偏向多數類別，進而忽略對少數類別（如高風險病患）的正確識別，降低模型在關鍵場景下的實用性與可信度。

傳統的特徵重要性方法，如 SHAP，雖能揭示模型對各特徵的依賴程度，卻無法具體指出「要如何改變特徵」，才能讓少數類樣本被正確識別。反事實樣本生成正是解決此問題的有效策略之一，其目標是針對某筆資料，產生一筆與其相似、但會被模型判斷為目標類別（如高風險）的樣本，並確保特徵變動合理且可解釋。

以醫療應用為例，醫師可能想知道：「如果某位病患目前被模型判定為低風險，是否能透過提升 XX 基因表現、降低 YY 指標，使模型改判為高風險？」若能生成這類與目標類別對齊的反事實樣本，將有助於強化模型在少數類別上的理解與識別能力。

本研究提出結合生成對抗網路、反事實解釋與 SHAP，不僅可產生具備目標預測結果的反事實樣本，還能在 SHAP 特徵貢獻層面上模擬目標類別的解釋模式。這樣的樣本不僅有助於模型學習關鍵決策邊界，也能在訓練資料中補足少數類樣本的多樣性與代表性，進一步緩解資料不平衡所帶來的預測偏誤問題。

第二章 文獻探討

## 2.1 SHAP

## SHAP（SHapley Additive exPlanations）是一種以賽局理論為基礎的模型解釋方法，透過計算每個特徵對模型預測結果的邊際貢獻值，來量化特徵對預測結果的重要性。其理論基礎來自 Shapley Value，這是一種源自合作賽局理論的公平分配方式，用以分配多方共同造成的效益

Shapley value公式:

(1)

* N:特徵集合
* S:特徵子集
* :子及S的總貢獻
* :特徵i的Shepley value，代表特徵i的平均貢獻

## 2.2 ACGAN

生成對抗網路（Generative Adversarial Network, GAN）是一種由生成器（Generator）與判別器（Discriminator）組成的模型架構，能夠透過對抗式訓練的方式，從隨機向量生成樣本。但是傳統的 GAN 無法直接控制生成樣本的類別，因此限制了其在特定應用場域中的靈活性與可用性。

為解決上述問題，Odena 等人於 2017 年提出 ACGAN（Auxiliary Classifier GAN） [2]。該方法在 GAN 架構中引入輔助分類器（Auxiliary Classifier），使判別器除執行真偽判斷外，亦能同時預測輸入樣本的類別標籤。在訓練過程中，生成器同時接收隨機向量與類別條件（class label）作為輸入，並生成符合該類別特徵的樣本。判別器則負責輸出兩項結果：（1）樣本的真實性判斷；（2）樣本的類別預測。分類結果所產生的誤差將反向傳遞至生成器，以提升生成樣本的類別一致性與可控性。

ACGAN 的損失函數可分為兩部分：

1.對抗損失（Adversarial Loss）：用於判斷樣本之真實性，確保生成樣本在外觀與結構上接近真實資料。

2.分類損失（Classification Loss）：用於判斷樣本所屬之類別，確保生成樣本與指定類別一致。

透過此設計設計，ACGAN 能夠兼顧生成樣本的真實度與類別可控性，並應用於圖像生成、數據增強及其他需條件化生成的研究領域。

## 2.3 反事實解釋

反事實解釋（Counterfactual Explanation）是一種可解釋人工智慧（Explainable Artificial Intelligence, XAI）技術，其核心概念為找出能夠使模型輸出結果改變的最小輸入變化，藉此揭示模型決策的關鍵因素。此類解釋方式基於反事實推理（counterfactual reasoning）的邏輯，即「若輸入資料在某些特徵上略有不同，模型的預測結果將隨之改變」。

在實務應用中，反事實樣本（counterfactual instances）通常用於回答具有行動導向的問題，例如：「要使模型的預測結果由類別 A 變為類別 B，需要對輸入資料做出哪些具體調整？」此特性使反事實解釋特別適用於決策支援場景，如貸款審批、醫療診斷與信用評分等領域，能為使用者提供具體且可執行的建議。

反事實解釋方法的設計一般包含以下核心原則：（1）最小變化性（Minimality）：生成的反事實樣本應與原始樣本在特徵空間中保持最小差異；（2）有效性（Validity）：反事實樣本必須能夠使模型預測結果產生目標變化；（3）可行性（Feasibility）：生成的變更應符合現實可行性。

因此反事實解釋不僅有助於揭示模型的決策邏輯，也能在保持解釋性與操作性的同時，為使用者提供可落實的行動建議。

## 2.2相關研究比較探討

### 2.2.1SHAP-GAN

傳統的生成對抗網路（GAN）雖能合成擬真的資料樣本，但難以控制生成結果的類別。為了解決此問題，Odena 等人提出了 ACGAN（Auxiliary Classifier GAN）[2]，在訓練過程中引入輔助分類器，使生成器不僅能根據雜訊向量產出資料，還能根據輸入的類別條件生成特定類別樣本，同時藉由分類器反向訊息提高樣本的一致性與類別可控制性。

然而，ACGAN 僅專注於外觀與結構上的擬真，對於資料解釋層面（如特徵影響程度）則缺乏關注。為補足此不足，Lin 等人於 2023 年提出 SHAP-GAN [3]。該方法基於 ACGAN 架構，在損失函數中新增一項 SHAP 一致性損失（SHAP Consistency Loss），使生成樣本在特徵貢獻度（即 SHAP Value 向量）上與真實資料對齊。SHAP-GAN 不僅要求生成樣本在資料分佈上模擬真實樣本，更強化了它在解釋層面的合理性。

SHAP-GAN 的損失函數由兩部分組成：對抗損失以及 SHAP 一致性損失。此設計有助於產出不僅外觀相似、且SHAP的貢獻也一致的樣本，進而提升分類器的預測率。

本研究進一步在 SHAP-GAN 的基礎上引入反事實解釋與樣本距離損失，使生成樣本不僅解釋一致，並與原始資料保持最小差異，具備反事實推理價值。相比 SHAP-GAN 所生成的新樣本，本研究所生成的反事實樣本更貼近人類認知與決策邏輯，具更高的應用潛力與可信度。

第三章 研究方法

## 3.1 問題案例描述

本研究旨在，輸入原始樣本與期望的目標類別，生成一筆反事實樣本，其不僅能有效使判別器做出目標預測，且在 SHAP 層面呈現與該類別樣本相符的特徵貢獻，提升生成樣本的貢獻一致性與可解釋性。

## 3.2方法簡介

訓練一個條件式生成模型 G，使其輸入真實樣本 與目標類別時，輸出一個新的樣本，滿足：

1.分類結果

2.與差距最小

3.SHAP Value與類別真實樣本一致

## 3. 3 問題與解決方法定義

對於輸入樣本 與目標類別，尋找對應的反事實樣本 ，使得下式損失最小：

損失函數定義與說明：

為每一個損失函數的權重。

1.對抗損失:

* 意義:使生成器產生的樣本無法被判別器區分出來，從而生成更真實、更合理的資料分佈，有助於提升反事實樣本的一致性與資料品質。

2.分類損失

* 意義:使生成樣本能被判別器判斷為目標類別

3. SHAP 一致性損失:

* 意義:使生成的樣本在特徵貢獻模式上與目標類別樣本一致，強化可解釋性與一致性

4. 樣本距離損失

* 意義:控制生成樣本與原始樣本的差異度，越接近原始資料越好，符合反事實樣本應「最小修改」原則。

## 3. 4 研究方法架構

本架構包含:

* 生成器:以原始樣本與目標類別 為輸入，生成對應於目標類別的反事實樣本。
* 判別器 判斷生成樣本是否為真實資料，並確認其是否屬於目標類別。
* SHAP :計算每個樣本的特徵貢獻。

## 3. 5 實驗方法

評估項目：

* 將反事實樣本加入訓練集後，分類準確率提升
* 與目標類別平均 SHAP 值相似
* 類別平衡指標（如 F1、Recall for minority class）

參考文獻

[1] [S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions, "in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4768–4777, Dec. 2017.](https://dl.acm.org/doi/10.5555/3295222.3295230#bibliography)

[2] A. Odena, C. Olah, and J. Shlens, “Conditional Image Synthesis With Auxiliary Classifier GANs,” arXiv preprint arXiv:1610.09585, Jul. 2017

[3] [Cai, J., Lee, Z.-J., Lin, Z., & Yang, M.-R. (2025). A Novel SHAP-GAN Network for Interpretable Ovarian Cancer Diagnosis. Mathematics, 13(5), 882.](https://www.mdpi.com/2227-7390/13/5/882)

[4] [E. Albini, J. Long, D. Dervovic, and D. Magazzeni,“Counterfactual Shapley Additive Explanations,”in Proc. 2022 ACM Conf. on Fairness, Accountability, andTransparency(FAccT'22),pp.1054-1070,2022.doi:10.1145/3531146.3533168](https://dl.acm.org/doi/10.1145/3531146.3533168)

[5] [R. Richman and M. V. Wüthrich,"Conditional Expectation Network for SHAP,"](https://arxiv.org/abs/2307.10654v1)

[arXiv preprint arXiv:2307.10654, 2023.](https://arxiv.org/abs/2307.10654v1)

[6] X. Li, K. Mao, F. Lin, and Z. Feng,“Feature-aware conditional GAN for category text generation,”arXiv preprint arXiv:2308.00939, 2023.