

**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模型**

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

**研究生：陳勝誠**

**指導教授：郭忠義 博士，王秉豐 博士**

**中華民國一百一十五年七月**

摘 要

ABSTRACT

目 錄

[摘 要 i](#_Toc207137611)

[ABSTRACT ii](#_Toc207137612)

[目 錄 iii](#_Toc207137613)

[第一章 緒論 1](#_Toc207137614)

[1.1 研究動機與目的 1](#_Toc207137615)

[第二章 文獻探討 2](#_Toc207137616)

[2.1 相關研究背景介紹 2](#_Toc207137617)

[2.1.1 SHAP 2](#_Toc207137618)

[2.1.2 ACGAN 2](#_Toc207137620)

[2.1.3 反事實解釋 3](#_Toc207137621)

[2.1.4 模型蒸餾 3](#_Toc207137622)

[2.2 相關研究比較探討 3](#_Toc207137623)

[2.2.1 SHAP-GAN 3](#_Toc207137624)

[第三章 研究方法 5](#_Toc207137625)

[3.1 研究議題 5](#_Toc207137626)

[3.2系統架構 5](#_Toc207137627)

[3.3方法簡介 6](#_Toc207137628)

[3.4 問題與解決方法定義 6](#_Toc207137629)

[3.5 實驗方法 7](#_Toc207137630)

[第四章 實驗與討論 8](#_Toc207137631)

[參考文獻 9](#_Toc207137632)

第一章 緒論

## 1.1 研究動機與目的

在許多現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導致模型偏向多數類別，進而忽略對少數類別（如高風險病患）的正確分類，降低模型在關鍵場合下的實用性與可信度，而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分類。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。

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，是否能透過提升某個基因的表現或是降某個指標，使模型改判為高風險？」若能生成這類與目標類別一致的反事實樣本，將有助於強化模型在少數類別上的理解與分類能力。

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 特徵貢獻層面上與目標類別擁有相似的特徵貢獻。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 相關研究背景介紹

### 2.1.1 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
* :子集S的總貢獻
* :特徵i的Shepley value，代表特徵i的平均貢獻

### 2.1.2 CAVE

Conditioned Variational Autoencoder (CAVE) 是變分自編碼器（VAE）的擴展版本，其核心在於將條件變數（例如類別標籤）納入編碼器與解碼器，使模型能夠生成帶有特定屬性的樣本。在此架構下，編碼器會根據輸入樣本與條件變數推得潛在機率分布，而解碼器則在同樣條件下重建輸入或生成新樣本。這種條件化設計不僅能保證生成樣本的類別一致性，也提升了生成過程的可控性與多樣性。由於其能針對少數類別進行樣本合成，CAVE 尤其適合用於類別不平衡或小樣本學習的場景，因此在醫療與生物資訊領域的應用研究中被廣泛採用。

### 2.1.3 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。

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。

### 2.1.4 知識蒸餾

知識蒸餾是一種以教師（teacher）和學生（student）為核心的知識轉移與模型壓縮方法。其核心在於：先以性能優良的教師模型學得豐富而穩健的決策行為，再以教師的輸出訊息作為監督訊號，訓練參數量較少、計算成本更低的學生模型，使其在顯著降低推論延遲與資源占用的同時，仍能接近甚至匹配教師模型的泛化能力。相較於只使用「硬標籤」的傳統監督學習，蒸餾引入了經由溫度參數 𝑇平滑化 softmax 機率（「軟標籤」），其所揭示的類別間相似度與不確定性常被視為「暗知識（dark knowledge）」，能在學生模型上重建更合理的決策邊界並提供正則化效果。

在常見的情況中，學生模型同時最小化對真實標籤的交叉熵損失與對教師軟標籤的散度損失。常見的總損失函數可表示為:

其中， ，α∈[0,1] 為權重係數。引入溫度（temperature）參數 >1 可在訓練階段軟化softmax 輸出，將原本尖銳的機率分布平滑化，以此顯示類別間的相似度與教師模型的不確定性訊號，用於學生模型的學習。實務上，蒸餾訓練時常取較大的 ；而在推理階段不再進行溫度縮放（等價於設 = 1），直接以學生模型的 logits 計算機率。

## 2.2 相關研究比較探討

### 2.2.1 SHAP-GAN

傳統的生成對抗網路雖能合成擬真的資料樣本，但難以控制生成結果的類別。為解決此問題，Odena 等人提出 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，使生成樣本不僅在SHAP特徵貢獻上一致，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。

### 2.2.2 SMOTE

Prasher 等人（2023）針對卵巢癌檢測提出結合 SMOTE 與 XGBoost 的方法，以解決小樣本與類別不平衡的問題。SMOTE（Synthetic Minority Over-sampling Technique）透過在少數類別樣本與其鄰近樣本之間進行插值，生成新的合成資料點，相較於單純複製樣本，更能有效增加少數類別的代表性並改善分類器對少數類別的識別能力。在該研究中，作者整合來自 Kaggle 的六個卵巢癌微陣列資料集（GSE6008、GSE9891、GSE18520、GSE38666、GSE66957、GSE69428），共計 529 筆樣本與 20 個基因特徵，先透過 SMOTE 進行資料平衡，再利用 XGBoost 的特徵重要性分數選取關鍵基因（如 EIF1、SPON1、ATG4B、SFN、AP3S1），最後以 XGBoost 分類器進行建模與評估。實驗結果顯示，該方法在測試集上達到 94.14% 的準確率，表現優於 CNN（約 93%）、SVM（約 90%）及其他集成式機器學習方法（約 85%），顯示資料增強與特徵選取結合傳統機器學習模型，能在醫學基因表達資料的分類任務中展現出高效能與良好泛化能力

第三章 研究方法

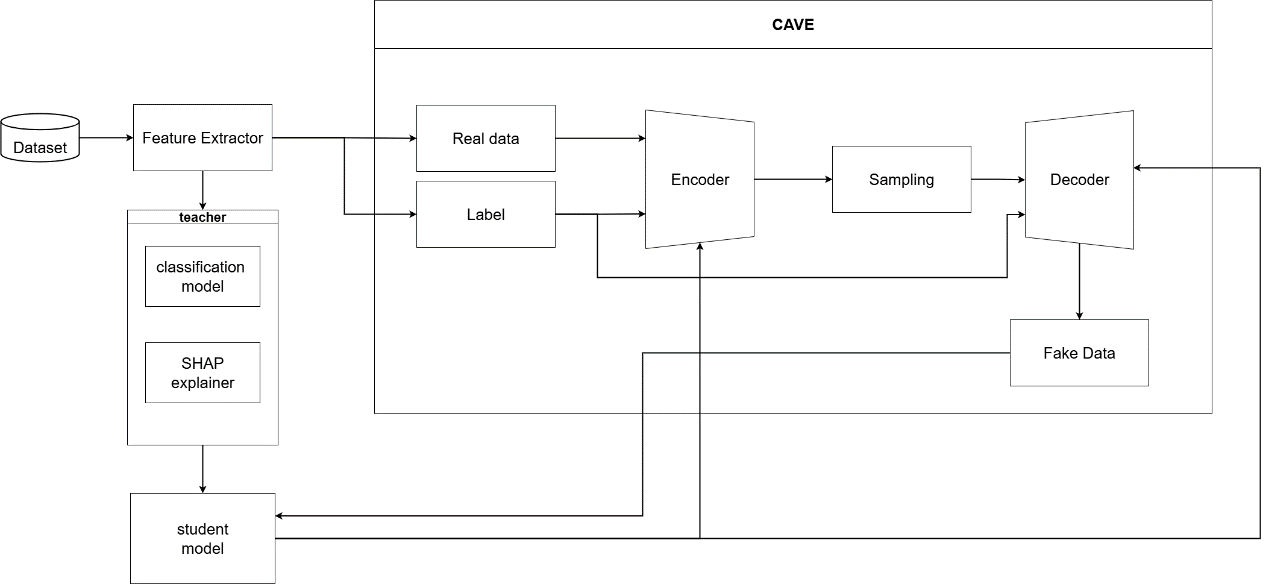
## 3.1 研究議題

在醫療領域中，許多資料因收集成本高、受試者條件限制或疾病罕見等因素，普遍存在樣本數量稀少與類別分佈不平衡的問題。例如，卵巢癌資料集 GSE6008 僅包含有限的樣本，且不同癌症亞型間的樣本數存在顯著的差異。在此情況下，模型在訓練過程中容易Overfitting多數類別，導致少數類別的特徵學習不足，最終造成分類性能下降與泛化能力不足。

本研究針對上述問題，結合 SHAP 特徵貢獻指引與反事實解釋指引，並透過生成對抗網路生成合理的少數類別反事實樣本；其中，SHAP 指引確保生成樣本在 SHAP的特徵貢獻模式與目標類別一致，反事實指引則限制改動僅發生在必要特徵且幅度最小。此方法目的在補足資料不足並改善類別不平衡的問題，期望能夠提升模型對各類別特徵的學習能力與整體分類表現。

## 3.2系統架構

本研究的系統架構主要由 CAVE（Conditional Variational Autoencoder）、教師模型（teacher model）與學生模型（student model）三大模組構成。首先，資料集經過 SHAP explainer 計算每個特徵的貢獻值，再根據平均貢獻度選取前 K 個重要特徵，進行特徵選取以保留最具代表性的輸入資訊。接著，將真實樣本輸入到教師模型，透過 SHAP explainer 為每筆樣本產生對應的特徵貢獻向量，並以此作為新的標籤訓練學生模型，使其能近似 Shapley value，以此建構一個可微分的 SHAP 模型。在 CAVE 訓練階段，編碼器(Encoder)會接收真實樣本及其對應的類別標籤，經由解碼器(Decoder)生成假樣本。這些合成樣本會送入學生模型以預測其Shapley value，並與目標類別在真實資料中的 Shapley value 進行比較。所得的 Shapley value會作為一項附加損失，納入 CAVE 的損失函數中，使模型生成的資料在特徵層級上與真實資料擁有相似的解釋性。



#### 圖1 系統架構

## 3.3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.在特徵空間中與原始樣本 𝑥 保持高度相似，僅進行達成目標分類 所需的最小改動

3.SHAP Value與類別真實樣本相似

## 3.4 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：

(2)

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:

* (3)
* 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

* (4)
* 意義:使生成樣本能被分類器判斷為目標類別

3. SHAP 一致性損失:

* (5)
* 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，強化與真實樣本的一致性

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。

## 3.5 實驗方法

評估項目：

* 將生成的樣本加入訓練集後，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）

第四章 實驗與討論

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.