應用資料擴增技術於異常浪機率預警系統之研究

陳盈智[1]、陳威成[2]、林芳如[3]、潘琦[3]、蔡政翰[4]、董東璟[1,2]

[1]國立成功大學近海水文中心 [2]國立成功大學水利與海洋工程學系

[3]中央氣象署海象氣候組 [4]國立臺灣海洋大學海洋環境資訊系

114年天氣分析與預報研討會
2025.09.02~04

2025/09/03

前言與研究目的

海岸瘋狗浪(Coastal Freak Wave)

- **海岸瘋狗浪尚無明確定義，只要是在海岸邊激起的巨浪並足以將人沖倒，就可稱為海岸瘋狗浪**
- 發生時機可分為兩種類型：

1. 天氣不佳不斷有大浪襲擊岸邊或礁岩
2. 在風平浪靜或僅有微風的天氣，突然激起大浪將人捲入海中

- **中央氣象署(2019)統計台灣2000至2019年共發生360件合計589人落海**

  海岸瘋狗浪的形成機制複雜，尚未有理論可完整解釋

  ➔難以預測何時何地會發生

圖片來源：三立新聞網

圖片來源：TVBS NEWS

人工智慧方法(artificial intelligent, AI)

近年來因電腦硬體與技術增強，**利用AI分析複雜的自然現象之研究層出不窮**

研究目的

海岸瘋狗浪發生機制複雜且難預警其發生時間，而近幾年已證實人工智慧技術能被成功運用

- **海岸瘋狗浪機率預警系統可透過AI方法建立，然而如果實際案例數量稀少，則難以進行。為解決此問題，本研究嘗試應用資料擴增方法於真實資料案例不充足時，仍可建立瘋狗浪預測模型。本研究採用三種資料擴增技術進行比較與應用，包含噪聲注入、SMOTE和Mixup等方法。**

研究區域與資料

研究區域

- 本研究的主要研究區域為**新北市貢寮區龍洞海岸之龍洞灣岬**

  龍洞灣岬左側臨山，右側臨海，因長期海蝕作用，形成懸崖陡峭的地形

  位於台灣東北海岸，易受**東北季風的影響，經常會出現海岸瘋狗浪**

- 中央氣象署(2019)**建立光學監視站長期拍攝龍洞灣岬**

  以人工檢查是否有超過1.5倍人高的浪花，判斷是否有瘋狗浪發生

  每小時發生一件或一件以上的海岸瘋狗浪，即視為一筆海岸瘋狗浪案例

  彙整發生瘋狗浪之時間，結合鄰近龍洞浮標海氣象數據，建立海岸瘋狗浪資料

研究資料

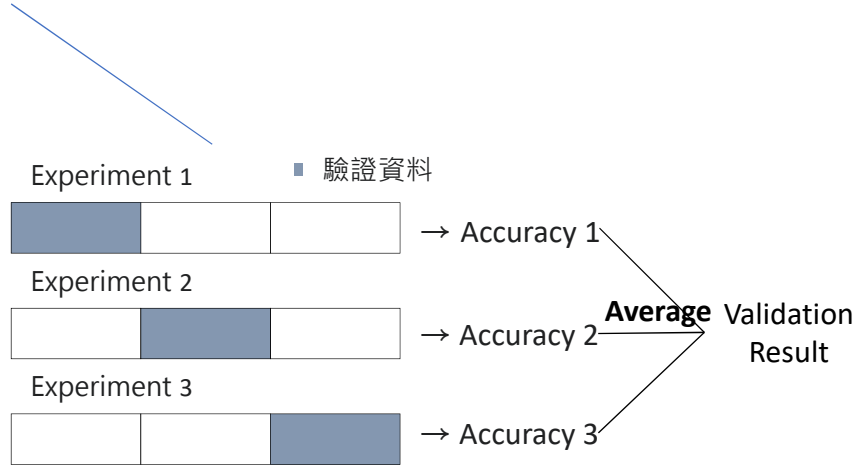
*資料正規化(Normalization)
將原始數據縮放至特定範圍，使得各個海氣象因子數據的尺度一致

使用2016-2019年期間的龍洞瘋狗浪案例

一共使用**1400筆資料**(有瘋狗浪與無瘋狗浪資料各占一半)，並將資料進行正規化(normalization)處理後，作為模型學習的資料庫並劃分為訓練、驗證與測試資料。

- 80% (1120筆)➜ 訓練與驗證資料。其中，驗證資料由K-fold交叉驗證產生
- 20% (280筆) ➜ 測試資料

訓練資料用於模型的學習，驗證資料用於模型調整訓練參數時比較依據，測試資料用於評估最終模型的性能。



瘋狗浪潛在因子

- 參考Doong et al. (2020)的研究，選取可能影響瘋狗浪發生的潛在因子

海象類別因子	示性波高
	平均週期
	尖峰週期
氣象類別因子	平均風速
方向類別因子	波向
	風向
湧浪類別因子	波浪尖銳度
	湧浪波高
	湧浪週期
波浪群性類別因子	窄度係數
	譜尖參數
非線性類別因子	BFI值

➢ 波浪群性因子

波譜的**窄度係數(narrowness)和Goda(2000)提出的譜尖參數(Qp)可用來描述波浪的群性**，在群性明顯的波群中，其能量通常較為密集

窄度係數: $Narrowness = \sqrt{\frac{m_0 m_2}{m_1^2 - 1}}$

譜尖參數: $Qp = \frac{2}{m_2^2}\int fS^2(f)df$

➢ 非線性類別因子:

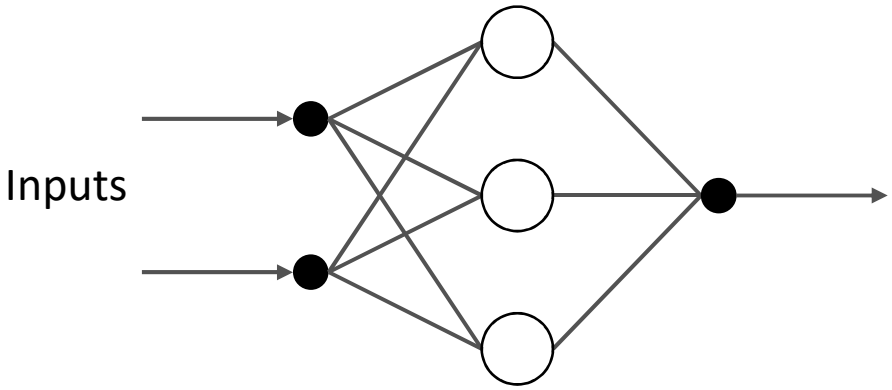
班傑明非線性指數(Benjamin Feir Index, BFI)可由波浪尖銳度以及譜尖參數計算得到，能**量化波浪不穩定現象的程度**

BFI值: $BFI = \sqrt{2\pi m_0} k_p Q_p$

研究方法

類神經網路(Artificial Neural Network, ANN)

- 類神經網路是一種模仿人類大腦神經元結構的機器學習模型，由**多個神經元(Neurons)組成龐大複雜網狀架構，神經元間透過權重(Weight)相連，形成數學函式**
- 模型在訓練過程會不斷地計算與更新權重，目標是減少模型輸出與實際值之間的誤差，最終得到擬合欲解決問題的函式



輸入層(Input layer)：接收外界給予的資訊

隱藏層(Hidden layer)：對輸入資訊進行計算和特徵提取

輸出層(Output layer)：輸出最終的預測結果

→ 類神經網路可利用將神經網路輸出層的啟動函數設定為Sigmoid函數，讓模型的輸出限制在0~1的範圍來得到機率

$$P = \hat{y} = f\left(\sum_{i=1} y_i^{(l-1)} * w_i^l + b_i^l\right)$$

資料擴增方法-噪聲注入

本研究提出**在案例數尚不充足的情況下，利用資料擴增方法(data augmentation)增加系統建置結果的可信度**

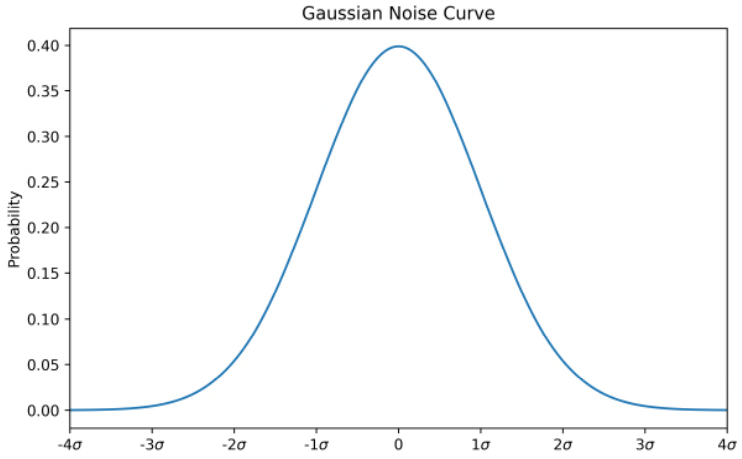
➔選擇了三種常見且廣泛使用的方法進行實驗與比較

1. 噪聲注入(Noise Injection)

透過在**輸入資料中添加隨機噪聲，模擬真實環境中的變化**，幫助模型學習到更多潛在的資料特徵或規律

$$x_i' = x_i + n_i$$

其中，x_i為原始樣本，x_i'為加入噪聲後的新樣本，n_i為隨機產出的噪聲



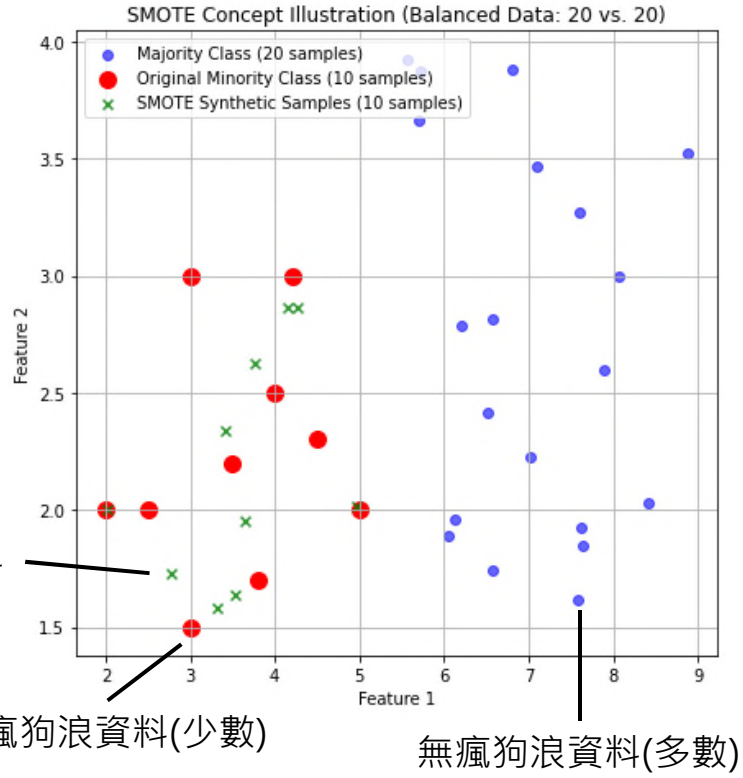
Ex: **高斯噪聲(Gaussian Noise)**為一種被廣泛運用的隨機噪聲，符合常態分布的特性

$$\mathrm{N}(\mu, \sigma^2) = p_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

資料擴增方法-合成少數過採樣方法

2. 合成少數過採樣方法(Synthesized Minority Oversampling Technique, SMOTE)

分析資料集中數量較少的類別(即瘋狗浪資料)，合成新的資料並加入，平衡有瘋狗浪與無瘋狗浪之間的資料比例



① 隨機選擇少數類別中的一個資料

② 利用K-近鄰演算法(K Nearest Neighbors Algorithm)計算歐式距離，**找出該資料在多維特徵空間中的K個同類別鄰近資料**

③ 在這些鄰近資料中選取其中一個資料

④ 在**原始選取資料與所選鄰近資料之間，產生新的資料**

$$x_i' = x_i + \beta \cdot (x_{zi} - x_i)$$

其中，x_i為隨機選擇的原始樣本，x_i'為合成後產生的新樣本，x_{zi}為x_i鄰近的同類別樣本，β為0到1之間的隨機數

資料擴增方法-混合擴增方法

3.混合擴增方法(Mixup)

透過**將兩個不同資料按照隨機比例進行混合，且同時對數據及其標籤進行線性插值產生新的資料**(Zhang et al., 2017)

$$x_i' = \lambda x_i + (1-\lambda)x_j$$

$$y_i' = \lambda y_i + (1-\lambda)y_j$$

其中，(x_i, y_i)與(x_j, y_j)為兩個隨機選取資料的數據與類別，λ則為0到1之間的隨機數，可代表從原始資料產生新樣資料的權重

綜以上所述，本研究選取三種資料擴增方法進行瘋狗浪資料擴增，並將擴增後的資料輸入模型進行訓練。**比較不同資料擴增方法的成效與特性，探索適用於瘋狗浪資料擴增與建模的技術**。

研究結果

尚未資料擴增之模型

為比較資料擴增對瘋狗浪預測模型訓練效果的影響，本研究先**將原始瘋狗浪資料自行刪減至原數量的一半(280筆)，並以此作為訓練資料，建立一個尚未經資料擴增的模型**，作為**對照組**以進行比較

- ✓ 瘋狗浪資料筆數：280筆 (尚未資料擴增)
- ✓ 無瘋狗浪資料筆數：560筆

訓練結果	正確率	76.1
測試結果	正確率	63.9
	反查率	39.3
	回應率	77.5

明顯出現過擬合問題 (overfitting)

正確率：所有瘋狗浪或無瘋狗浪的資料中，模型預測正確的比率

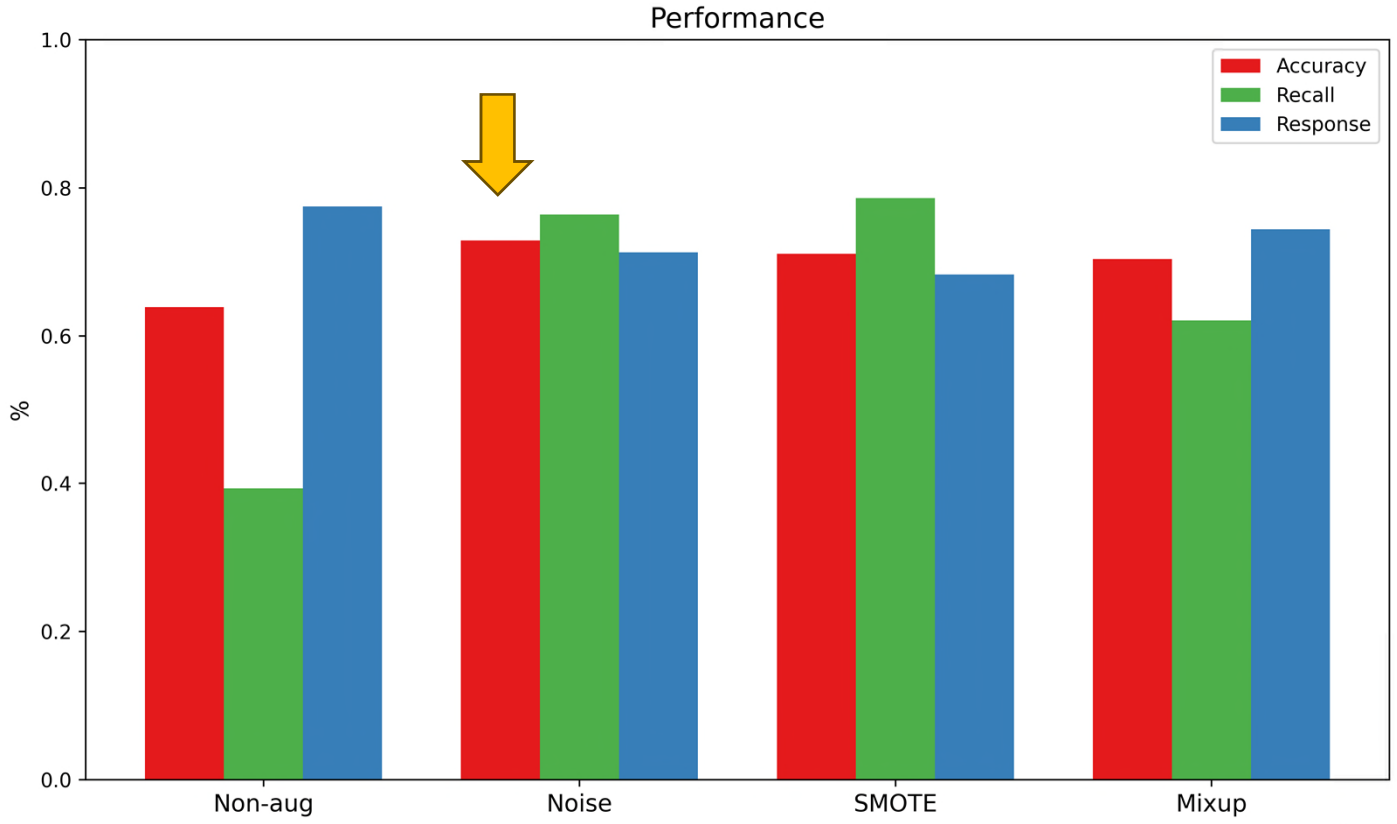
回應率：模型預測為瘋狗浪會發生的結果中，有多少比例實際真的有發生

反查率：實際有瘋狗浪發生的資料中，模型有正確預測到瘋狗浪發生的比率

資料擴增後之模型

使用三種資料擴增方法：**噪聲注入、合成少數過採樣(SMOTE)、混合擴增**

- ✓ 瘋狗浪資料筆數： 280筆 → 560筆 (資料擴增後)
- ✓ 無瘋狗浪資料筆數：560筆

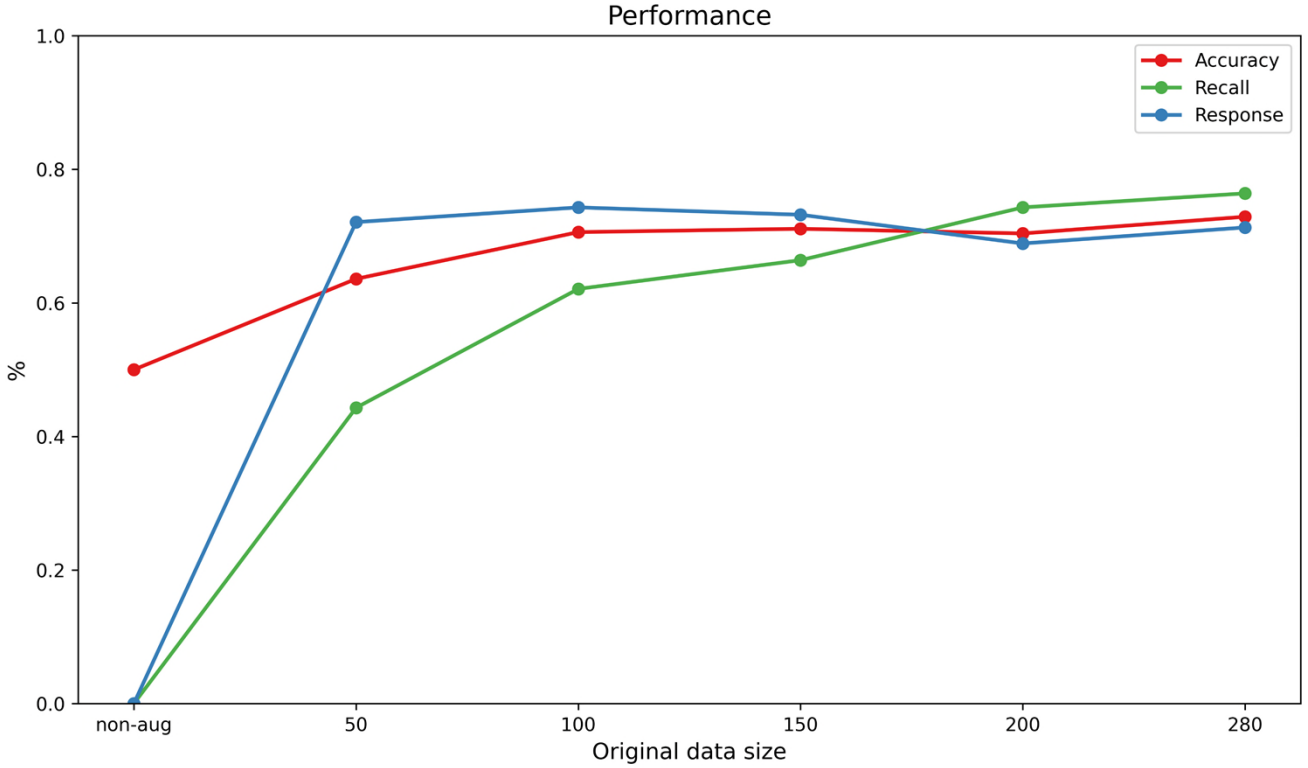


- ➢ **「噪聲注入」方法提升的訓練效果為最佳**
- ➢ 其**正確率提升最為顯著 (63.9%→72.9%)**
- ➢ 其他兩種方法的反查率獲回應率雖較高，但其他指標低於七成，非平衡的預測表現

不同原始資料數量擴增後之模型

使用不同原始資料數量進行擴增後並訓練，比較不同模型之訓練效果

- ✓ 瘋狗浪資料筆數： 50,100,150,200,250,280筆 → 560筆 (噪聲注入方法擴增後)
- ✓ 無瘋狗浪資料筆數：560筆



- ➢ 隨著**原始資料數量的增加，模型訓練結果愈加穩定**，各項評估指標皆能維持在七成以上
- ➢ 在**少量的原始資料量**下進行資料擴增，已**可獲得一定的訓練效果**

使用真實案例擴增之模型

不使用資料擴增方法，而是使用真實瘋狗浪資料擴增，**建立僅有真實案例的瘋狗浪預測模型**進行比較

瘋狗浪資料筆數： 280筆 → 560筆 (經資料擴增方法/真實案例擴增)

無瘋狗浪資料筆數：560筆

		尚未資料擴增	經資料擴增	真實案例擴增
訓練結果	正確率	76.1	74.1	71.8
測試結果	正確率	63.9	72.9	73.6
	反查率	39.3	76.4	81.4
	回應率	77.5	71.3	70.4

尚未資料擴增前的模型: 訓練效果明顯較差
真實案例擴增後的模型: 正確率明顯上升，反查率可突破八成
經過資料擴增後的模型: **雖不及真實案例擴增，但已可接近真實案例擴增後的訓練結果**

➔顯示資料擴增方法在真實案例數量較少時，可用於彌補訓練資料不足的問題，幫助產出良好的訓練效果

結論

結論

本研究選取三種資料擴增方法進行瘋狗浪資料擴增，並將擴增後的資料輸入模型。比較不同資料擴增方法的成效與特性，探索適用於瘋狗浪資料擴增與建模的技術

✓ 三種資料擴增方法中，以**「噪聲注入」方法提升的訓練效果為最佳**，其正確率提升最為顯著

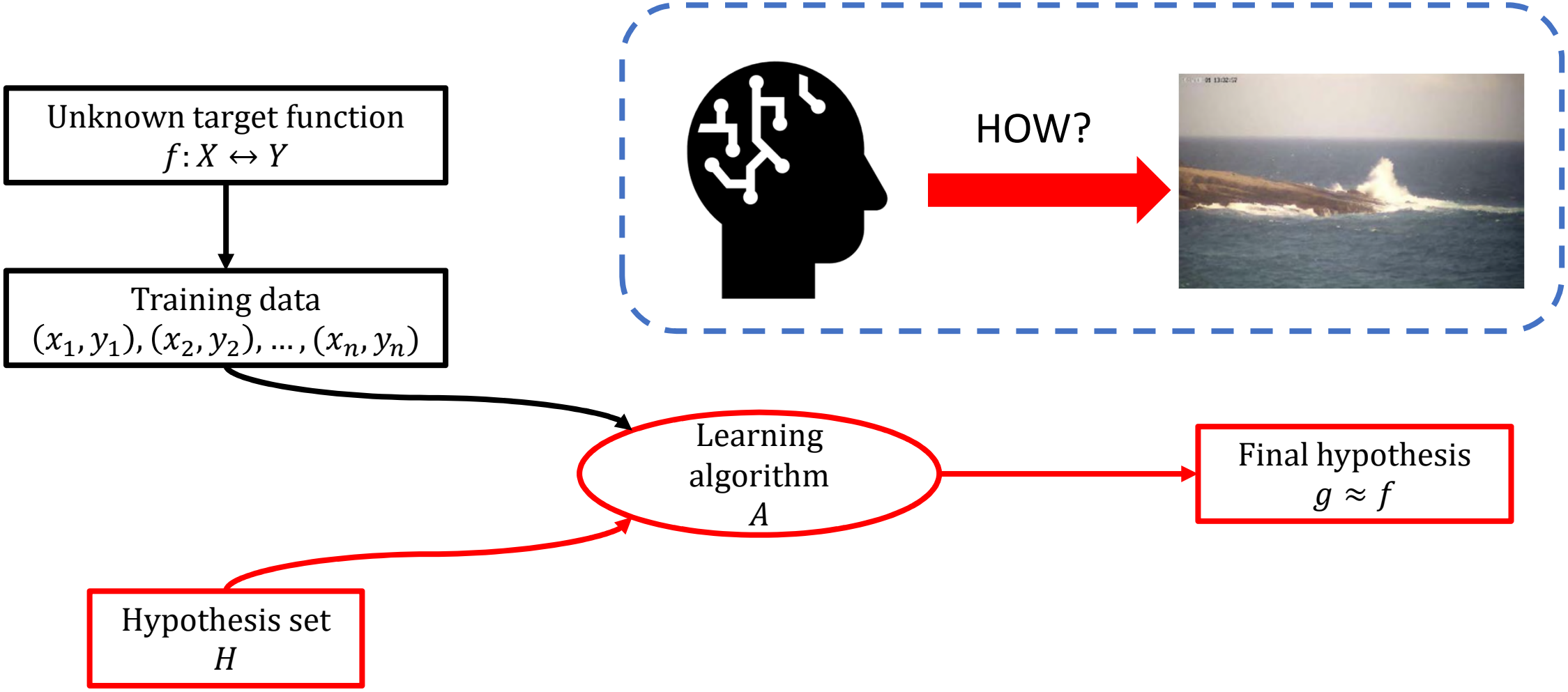
✓ 在少量的原始資料量下進行資料擴增，已可獲得一定的訓練效果

✓ 經過資料擴增後的模型，雖不及真實案例擴增後的訓練成果，但已可接近真實案例擴增的結果

✓ 顯示**資料擴增方法在瘋狗浪實際案例數量稀少時，可有效幫助模型訓練，完成瘋狗浪機率預警系統建置**

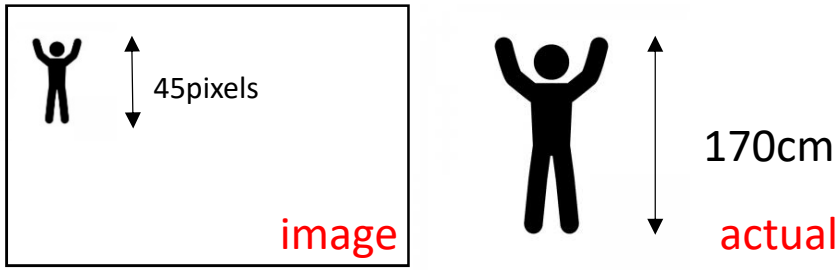
Thanks for your attention

AI應用瘋狗浪預測?

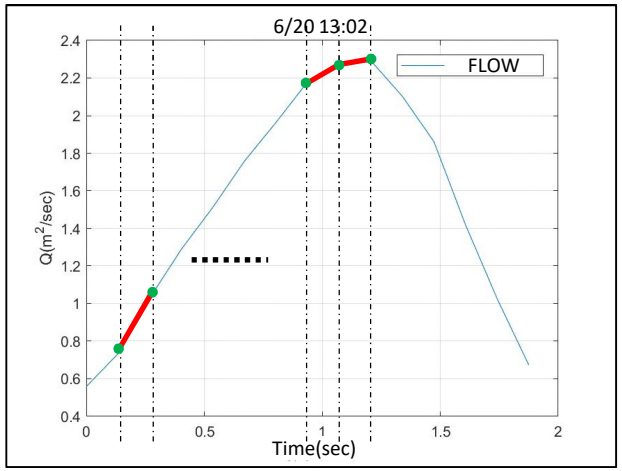

Unknown target function
$f: X \leftrightarrow Y$
Training data
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
HOW?
Learning algorithm
A
Final hypothesis
$g \approx f$
Hypothesis set
H

瘋狗浪影像分析

$$V_{CFW} = \rho_{splash} \int_{t_s}^{t_e} \eta(s,t) ds$$



➡ Scale is 3.8cm per pixel.

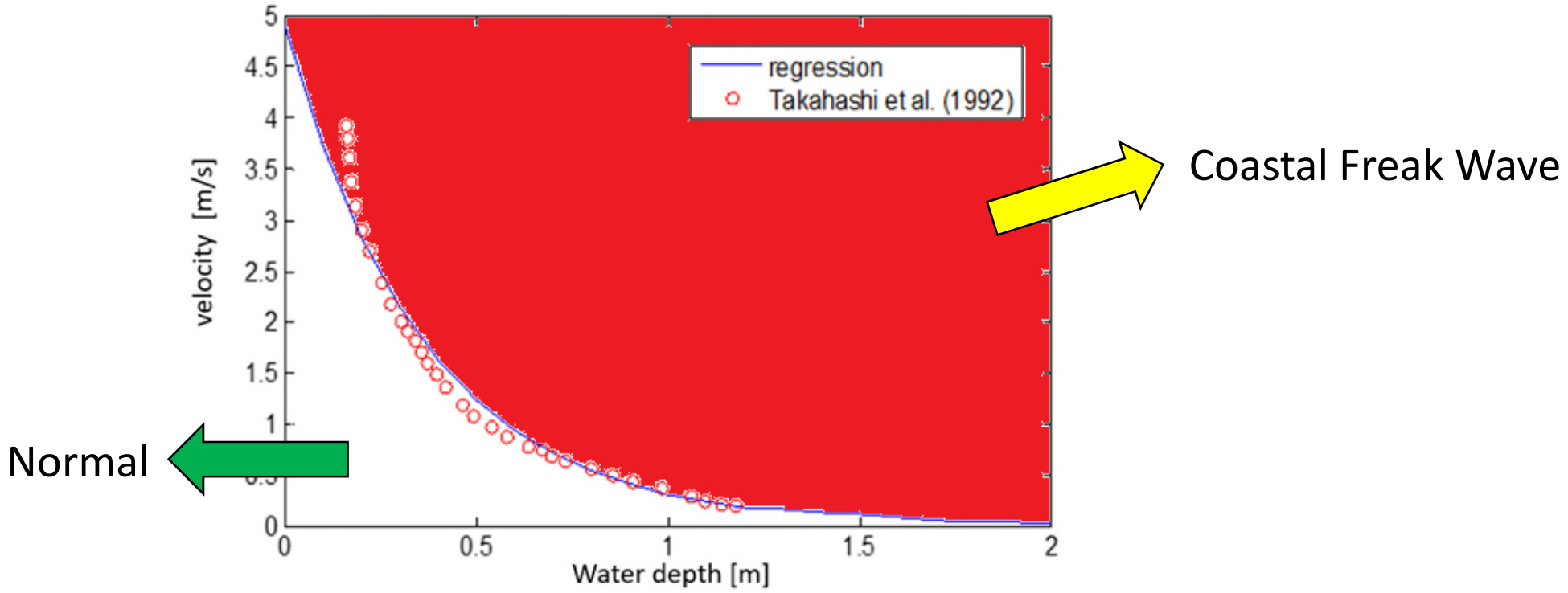


$$Q_{CFW} = \frac{1}{t_p} \int_{t_s}^{t_p} d\,V_{CFW}/dt \;\; dt$$

t_{pick} : *duration of freak wave arrive* max *discharge*

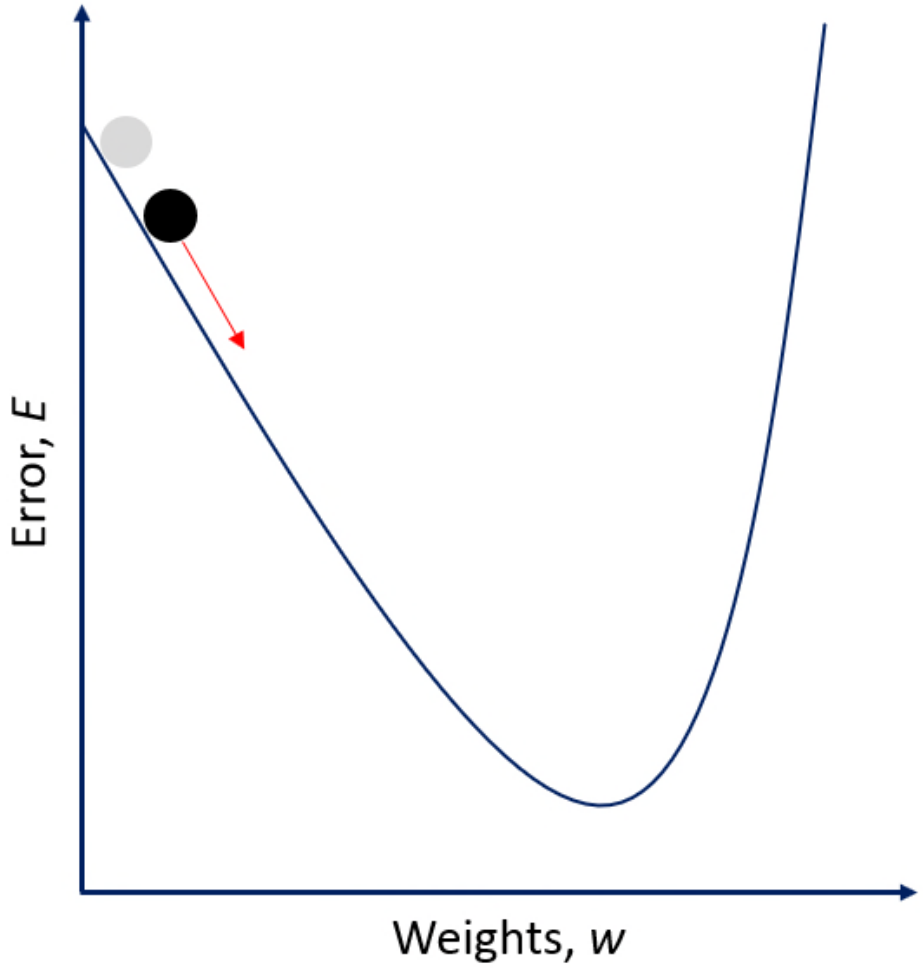
$$u_{CFW} = dQ_{CFW}/dH_{CFW}$$

瘋狗浪影像分析



海嘯實驗與實驗迴歸結果 (Takahashi et al., 1992)

梯度下降法(ANN)



$$W_{new} = W_{old} - \eta \nabla L(w)$$

四種評估指標

正確率(Accuracy) 正確預測瘋狗浪會與不會發生的比率

$$Accuracy = \frac{TP + TN}{T}$$

反查率(Recall Rate) 實際有瘋狗浪時，能正確預測的比率

$$Recall = \frac{TP}{TP + FN}$$

回應率(Response Rate) 模型預測有瘋狗浪時，實際亦正確的比率

$$Response = \frac{TP}{TP + FP}$$

預兆得分(Threat Score) 考慮偽陽性和偽陰性兩種誤報並正確瘋狗浪的比率

$$Threat\ Score = \frac{TP}{FN + FP + TP}$$

		True Condition	
	Total Population (T)	Positive	Negative
Predicted Outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

超參數率定

訓練超參數(hyperparameters)影響模型在訓練時的架構與策略，**不同的超參數組合會有不同的訓練效果**

→ 隨機森林: 分類標準(criteria)、決策樹數量(number of trees)與最大選擇特徵數量(maximum of features)

→ 支撐向量機: 核函數(kernel function)、映射係數(mapping coefficient)與懲罰係數(penalty coefficient)

→ 類神經網路: 啟動函數(activation function)、隱藏層神經元數量(number of neurons)與學習率(learning rate)

本研究使用**k-fold交叉驗證(k-fold cross-validation)**與**網格搜尋法(grid search)**率定模型最佳的超參數組合

- **K-fold交叉驗證**: 將訓練資料切成k等份，k-1等份訓練模型，剩下資料驗證模型得到正確率，直到每一等份的資料都用作驗證過，將每次的結果平均起來，即為k-fold交叉驗證的結果
- **網格搜尋法**: 在限定範圍內輸入所有有可能的超參數組合，比較出訓練效果最佳的組合(以k-fold交叉驗證為依據)

超參數率定結果

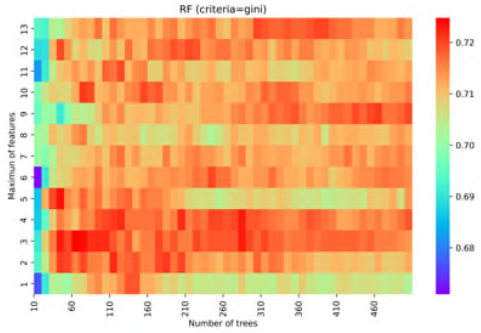
→ 隨機森林不需要太多棵決策樹就可以有不錯的訓練效果

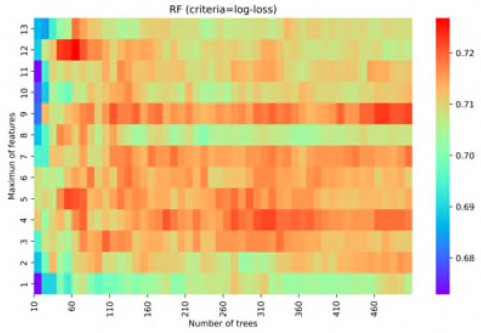
→ 支撐向量機比起使用線性核函數，使用多項式核函數會有更好的訓練效果

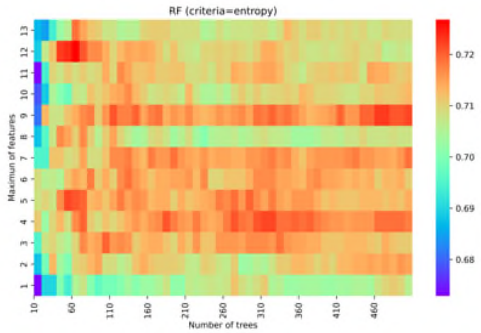
→ 類神經網路隱藏層使用較多的神經元可以有比較好的訓練效果

	RF		SVM		ANN	
Hyper-parameters	N_{trees}	60	γ	4	**Hidden layer**	(9,10,10)
	f_{max}	12	C	0.156	η	0.001
	Criteria	Entropy	**Kernel function**	Poly	**Activation function**	ReLU
Validation result	**71.0%**		**73.3%**		**73.2%**	

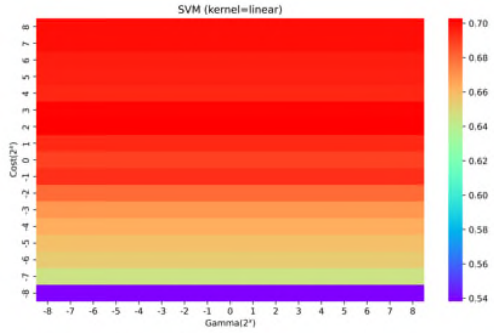
隨機森林

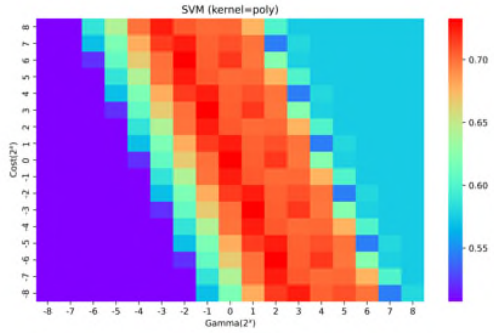


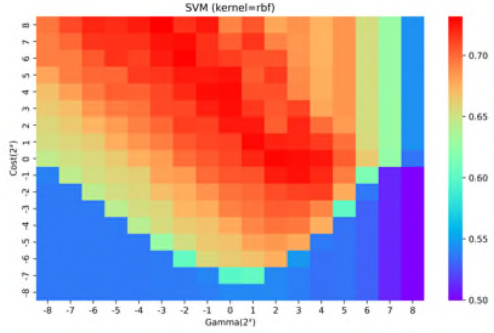




支撐向量機







類神經網路