111 學年度專題報告競賽

題目: 探討黑色素瘤基於系統化圖像特徵擷取的智能辨識

系所班別：統計學系三年級
姓名學號：郭依璇(410978004)
譚靖蓉(410978011)
謝瑋芸(410978044)
黃名揚(410978056)
李博業(410978058)
報告日期：2023/06/09

目錄

摘要

黑色素瘤患者死亡率極高，若未及早治療，待黑色素瘤擴散後，放射治療與化學治療的改善程度都有限。如果可以在黑色素瘤擴散之前及早發現、治療，便能大幅提高治癒的機率，對於此症狀的辨識目前大多是利用深度卷積神經網路 (CNN) 進行判斷，然 CNN 訓練會需要許多照片，與非常好的設備，且無法得知特徵，本研究嘗試直接擷取影像特徵，未來可以輔助判讀者快速找尋黑色素瘤。

本研究從公開資料蒐集黑色素瘤與非黑色素瘤照片，使用自動化裁剪圖片，並使用余佳蓉 (2021) [4] 建議的判定準則提取特徵，利用支援向量機 (SVM) 、羅吉斯迴歸 (Logistic Regression) 、隨機森林 (Random Forest) 、K-近鄰分類器 (KNN) 、單純貝氏 (Naive Bayes)、神經網路 (Neural Network)、六個分類器來辨識黑色素瘤。

對於黑色素瘤的辨識與預測，以 SVM 的模型表現最好，辨識黑色素瘤的準確度可達 76.5%。另外，在擷取的特徵中，顏色和紋理為辨識過程中最關鍵的特徵。

1 前言

黑色素瘤為一種與痣高度相似的皮膚癌，美國癌症協會統計 [1]，2023 年有 97,610 的新患者，預期會有 7,990 人會死亡，致死率約 0.08 % 。雖然致死率看似不高，但根據美國皮膚科學會 [2] 的資料，因皮膚癌而死的患者中，大部分都是源自於黑色素瘤，可知黑色素瘤對於人體影響甚大，但因其辨別不易，導致人民時常不自覺。

目前研究致力於用 CNN 來診斷黑色素瘤，相較於傳統皮膚科醫師憑影像及經驗判斷，CNN 有更好的判別效果，其確診率遠勝於醫生肉眼判斷的結果。現在 CNN 所訓練的模型可以協助醫師作為輔助判斷的工具，但模型仍存在著一些問題，例如 CNN 對於設備的要求較高，且需要使用大量的資料來達到較高的準確度，再者，最後 CNN 模型無法解釋，無法提供判讀者有效的資訊。

因此本研究期望在設備較差且資料量較少的情況下，找出能精確判別黑色素瘤的統計、機器模型，除了可以減少醫療、時間成本外，還能找出圖片分類的關鍵特徵，期望未來自動

診斷有望輔助現今的臨床皮膚檢測，有效降低臨床評估的負荷。

2 文獻回顧

文獻有許多黑色素瘤特徵擷取的演算法，而 ABCD 法是最常被提到的方法，A (Asymmetry) 代表非對稱，B (Border) 為邊界，C (Color) 為顏色，D (Diameter) 表示直徑，Majumder 等人 (2019) [3] 根據 ABCD 法則，加上將 A、B 與 D 分段處理特徵，使用 ANN (Artificial Neural Network) 模型辨識黑色素瘤的模型準確度 98.2 %、敏感度 98 %、特異性 98.2 %，但此文獻訓練及僅使用 200 張影像，而測試集則只有 22 張影像，且沒有交代下載來源。

(余佳蓉, 2021) [4] 提出一個系統，除 ABCD 準則，再加入 E (Enlargement) 細胞擴展的特徵，該系統可以在任何地方分析病患的皮膚鏡影像，提高早期黑色素瘤的發現率，從而提高治癒率。該系統的特徵提取使用 HSV 和 RGB 色彩空間，將影像轉換為灰度並進行高斯濾波，以濾除影像中的雜訊和毛髮，該系統僅有提供準確度達到 99.20 %，再者，此文獻提出的方法能夠解決雜訊和頭髮等問題，對早期發現黑色素瘤有很大幫助，但沒有探討是具體是如何將特徵值數據化的，也沒有詳細的分析過程，而且所使用的資料集的痣與瘤的圖片皆處理得非常乾淨、清晰、高清，且不需任何裁切，這是不符合實際的。
此文獻提出的方法能夠解決雜訊和頭髮等問題，對早期發現黑色素瘤有很大幫助，但沒有探討是具體是如何將特徵值數據化的，也沒有詳細的分析過程，而且所使用的資料集的非瘤與瘤的圖片皆處理得非常乾淨、清晰、高清，且不需任何裁切，這是不符合實際情況的。

Gessert 等人 (2020) [5] 使用 ISIC 2019 皮膚病變比賽所提供的黑色素瘤與非黑色素瘤的資料，運用 CNN 模型辨識黑色素瘤，但由於測試集得到最佳敏感度 74.2 %。在此文獻中用了大量的資料，但最終的模型無法解釋。

文獻提到 ABCDE 準則可以有效的截取特徵，但都僅以少量圖片說明其判讀的成效，本研究希望探討這個 ABCDE 準則是否適用 ISIC 2019 皮膚病變比賽所提供非典型的資料。由於該比賽的資料非常雜亂，本研究會先使用 Rehman 等人 (2022) [6] 介紹如何解決拍攝皮膚鏡圖片時所遇到的角落邊框的問題，與如何去除圖片中的毛髮，本研究先確定角落邊框的最外和最內輪廓，並依此裁剪圖片，以自動化方式先前處理比賽的大量圖片，再用數據較平衡的資料、尋找更多測試集以避免實驗誤差、用較不完美的圖片，通過前處理後達到相同的效果，並實踐特徵值數據化且將其套入分類器探討辨識效果。最終我們期望用較少的資料、較普通的電腦設備進行機器學習，達到相同、甚至更好的分類效果並可以解釋黑色素瘤的模型結果與特徵值。

3 研究方法及步驟

3.1 資料來源

本研究使用了 PH^2 資料集 [7] 中的 40 張黑色素瘤圖片、 MED-NODE 資料集 [8] 中的 70 張黑色素瘤圖片、ISIC Challenge 資料集 [9] 中的 162 張黑色素瘤圖片、Kaggle 資料集 [10] 中的 584 張黑色素瘤圖片以及 32042 張非黑色素瘤圖片，總共 856 張黑色素瘤的圖片，與 32042 張非黑色素瘤的圖片。

3.2 資料取樣

本研究收集到的可用黑色素瘤圖片總共 856 張，所用訓練集與測試集的比例為 7:3，按照比例將黑色素瘤圖片分為 600 張訓練集與 256 張測試集。

考慮到會有數據不平衡的情況，先將三萬多張非黑色素瘤圖片分為 7:3 的訓練集和測試集，再從訓練集中隨機取 600 張非黑色素瘤圖片當作樣本，從測試集中隨機取 256 張非黑色素瘤圖片當作樣本。

但隨機取樣的樣本可能不夠具有代表性，因此本研究從訓練集中隨機取十組 600 張的非黑色素瘤圖片當作訓練集，從測試集中隨機取十組 256 張的非黑色素瘤圖片當作測試集，如圖 1。故一組訓練集有 600 張黑色素瘤圖片與 600 張非黑色素瘤圖片，加起來共 1200 張圖片；一組測試集有 256 張黑色素瘤圖片與 256 張非黑色素瘤圖片，加起來共 512 張圖片，並在後續的訓練結果中比較十組訓練集和十組測試集的差異，以證明非黑色素瘤圖片樣本的代表性。

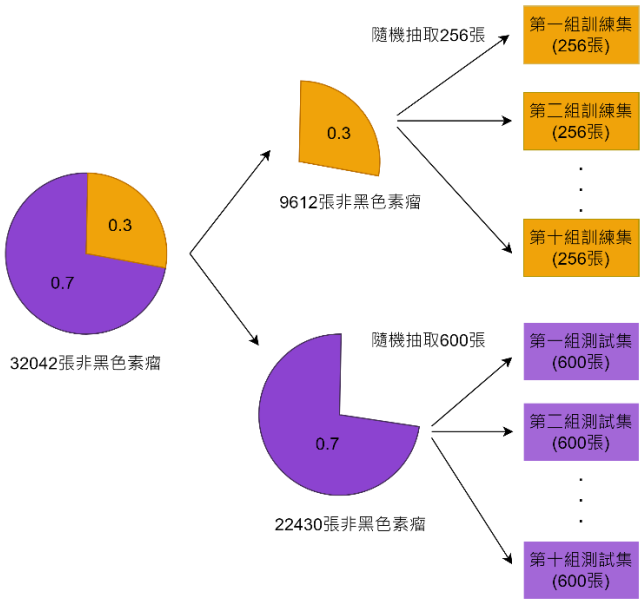

圖 1: 非黑色素瘤訓練集測試集取樣示意圖

3.3 資料預處理

由於收集到的原始圖片有許多問題，如圖片外圍的黑框白框、圖片中的毛髮、圖片中的其他雜訊，這些問題會導致後續特徵擷取處理上的困難，造成結果不理想。而通常在影像處理的領域中會使用開操作以及閉操作 (OpenCV, n.d.) [11] 來去除影像中的毛髮，並將圖片轉成二值化圖 (OpenCV, n.d.) [12] 以利後續擷取病徵位置。因此本次研究將圖片根據圖 2 流程進行處理，以下將詳細介紹各處理步驟。

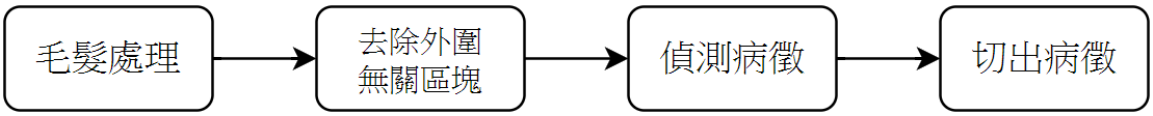

圖 2: 圖片預處理流程圖

3.3.1 毛髮處理

一、 閉操作

閉操作是對圖片先膨脹後腐蝕，填補圖片中的小黑洞以去除黑色的噪點。因此我們先對彩色圖閉操作以去除毛髮，成果如圖 4，可以看出毛髮有些微減少。

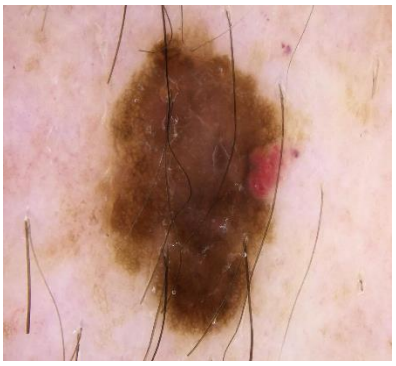

圖 3 : 閉操作前的原圖

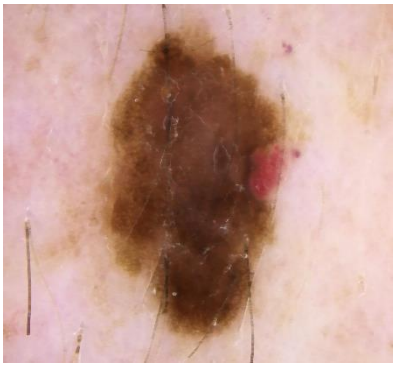

圖 4 : 圖 3 閉操作後的圖

二、 開操作

開操作是對完成閉操作的圖片先腐蝕後膨脹，去掉小白噪點並填充，用於移除小亮斑點，進而排除病徵外部的雜訊。

由於開操作適用於去除白色噪點，因此本研究先將圖片二值化，把殘留的毛髮轉成白色，再使用開操作將其去除，成果如圖 6。

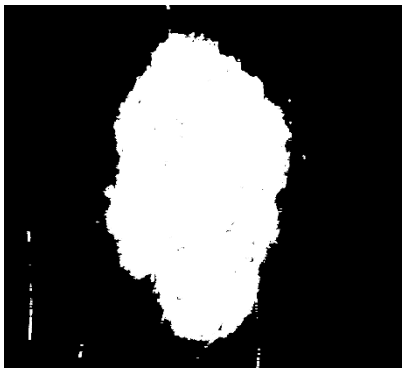

圖 5： 開操作前的二值化圖

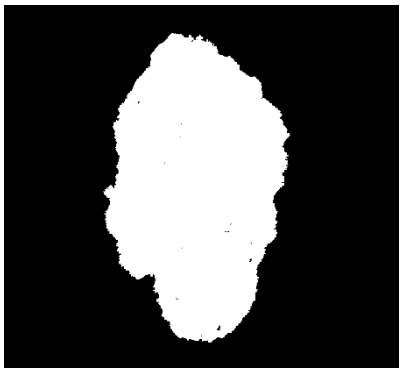

圖 6： 開操作後的二值化圖

3.3.2 去除外圍無關區塊

參考 Rehman 等人 (2022) [6]，從圖片中心點出發，運用程式去除外圍與本次研究無關的區塊。為了後續特徵擷取，必須透過二值化把圖片簡單地做分割。首先把彩色圖片，如圖 6，轉成灰度圖，並透過高斯濾波器抑制噪聲、平滑圖片，再對此灰度圖由 otsu 演算法 (Murzova, A. and Seth, S. , 2020) [13] 尋找閾值 (Threshold)，若像素灰度大於閾值則令為黑色，小於閾值為白色，成果如圖 8。

圖 7： 彩色原圖

圖 8： 圖 7 的二值化圖

由於收集的部分圖片外圍有黑圓框，如圖 7，而這是不需要且會影響後續特徵擷取的雜訊；為了切除干擾區塊，本次研究將二值化後的原圖，從圖片中心點開始，沿著圖片對角線的四個方向逐步確認其像素值，如圖 9，抓出紅色虛線與橘色虛線的交接點作為圖片的邊框端點，再由這些點去除外圍無關的區塊，成果如圖 10。

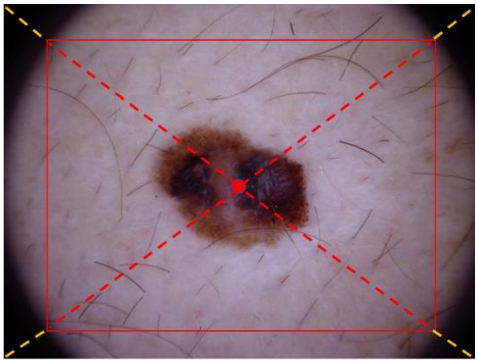

圖 9： 去除外圍無關區塊示意圖

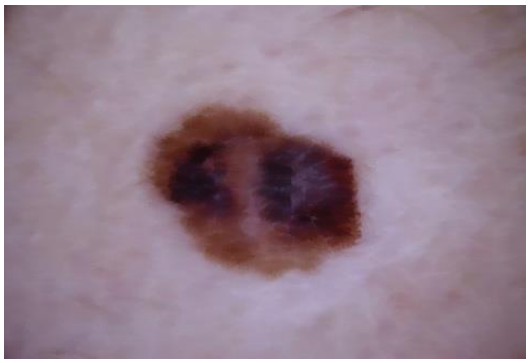

圖 10： 圖 7 去除外圍無關區塊完成圖

3.3.3 偵測病徵

一、 邊緣偵測

即使去除外圍無關區塊，圖片仍有會影響特徵擷取的雜訊，如圖 11，所以利用邊緣偵測畫出圖片中可能為病徵的所有輪廓。

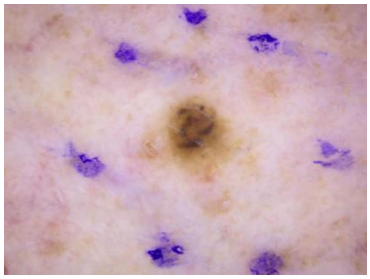

圖 11 :彩色原圖
(病徵外圍有明顯的塗料)

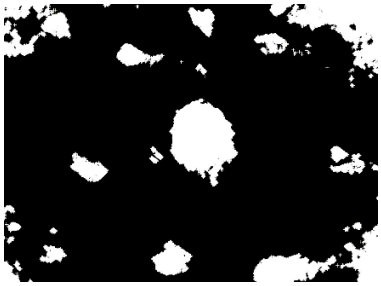

圖 12 : 圖 11 進行二值化後

圖 13 : 圖 12 邊緣偵測的結果

二、 最大輪廓

在切除無關區域的圖片中，病徵的輪廓通常比其他雜點輪廓要大，因此本次研究把各個輪廓，如圖 13，依長度排序，並提取最長的輪廓，將其視為病徵輪廓，如圖 14。

圖 14： 圖 13 中的最大輪廓

3.3.4 切出病徵

一、最小外接矩形

根據病徵輪廓，畫出其最小外接矩形，如圖 15，再把最小外接矩形作為遮罩蓋在原圖之上，如圖 16，並依照矩形角度旋轉圖片 ，如圖 17。

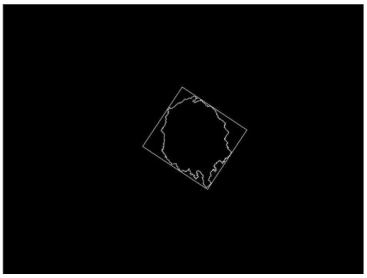

圖 15： 在病徵輪廓的最小外接矩形

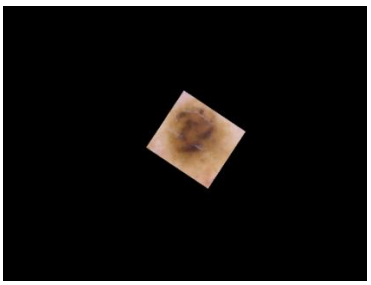

圖 16： 原圖蓋上最小矩形遮罩後的圖

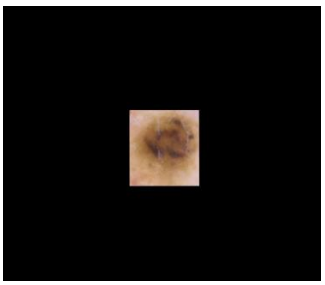

圖 17： 依矩形角度旋轉後的圖 16

二、裁剪圖片

利用程式裁剪旋轉後的矩形圖片 ，提取病徵圖片作為後續進行特徵擷取的最終特徵圖，如圖 18。

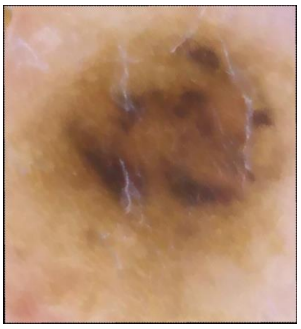

圖 18: 最終提取的特徵圖

3.4 特徵擷取

根據 ABCDE 法則以及黑色素瘤的特性，本研究歸納出幾個特徵擷取的方向，分別為不對稱性、不規則、顏色、紋理。

3.4.1 不對稱性

先將圖片二值化並切成四等分，如圖 19，並求出每一等分內病徵的面積比例，即每一等分白色面積比例，最後計算四個面積比例之標準差，作為不對稱的特徵值 (sym)。

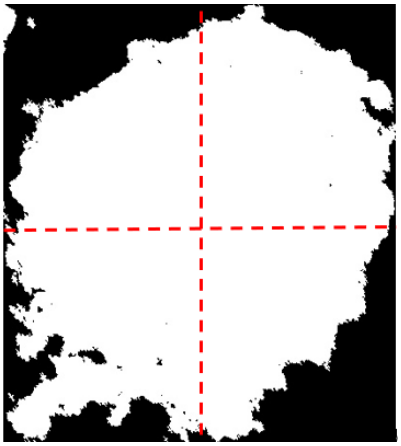

圖 19: 不對稱性特徵擷取示意圖

3.4.2 不規則

基於不規則特徵，本次研究在病徵輪廓上畫出擬合橢圓，如圖 20，並計算病徵與橢圓面積差，再將其除以病徵總面積。而藉由以上方法計算出的比例，作為本次研究不規則特徵的特徵變數。命名為不規則比例－橢圓（ae_diff）。

另一不規則特徵則是根據病徵輪廓畫出其最小外接矩形，並計算兩者相差面積占病徵面積比例，如圖 21。命名為不規則比例－矩形（ar_diff）。

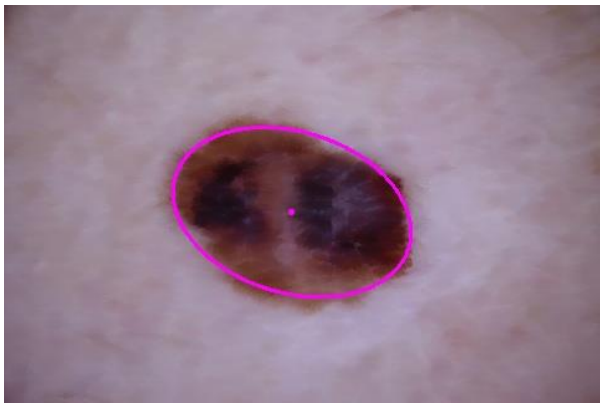

圖 20：在圖 10 的病徵輪廓上畫出擬合橢圓

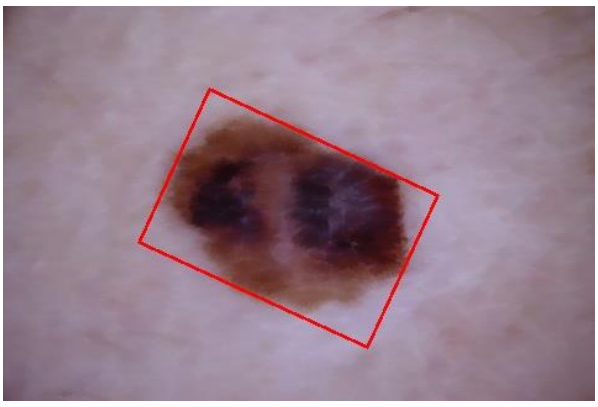

圖 21：在圖 10 的病徵輪廓上畫出最小外接矩形

3.4.3 顏色

由 ABCDE 法則的 C 可以知道黑色素瘤大多會有顏色分佈不均的現象，因此本次研究根據病徵圖片的顏色通道 (HSV)，分別計算各通道的平均值、標準差及偏度，總共九個特徵，色相平均值 (h_mean)、色相標準差 (h_std)、色相偏度 (h_skw)、飽和度平均值 (s_mean)、飽和度標準差 (s_std)、飽和度偏度 (s_skw)、明度平均值 (v_mean)、明度標準差 (v_std)、明度偏度 (v_skw)。hsv 色彩三屬性請見附錄 B。

3.4.4 紋理

黑色素瘤的表面崎嶇，容易產生比非黑色素瘤更複雜的紋理，因此本次研究計算病徵圖片的灰度共生矩陣，得到四個角度 (0 度、45 度、90 度、135 度) 的五種統計量 : 對比 (Contrast)、同質性 (Homogeneity)、差異性 (Dissimilarity)、相關性 (Correlation)、能量 (Energy)，總共 20 個特徵。詳細方法請見附錄 C。

3.4.5 病徵佔原圖比例

圖片占比 (Proportion) 特徵是考量到每張圖內病徵面積大小占比不一，故本研究增加病徵占原圖面積比例為 ABCDE 法則外的特徵。

3.5 特徵值標準化

圖 22 是黑色素瘤各變數未經標準化的盒形圖，可以發現變數之間的尺度差距大，若不標準化，在做後續分析時可能會造成影響和誤差。圖 23 為黑色素瘤各變數標準化後的盒形圖，觀察資料的分布，並沒有變數分布有明顯差異，因此後續會再使用不同的分類器對特徵進行統計分析。

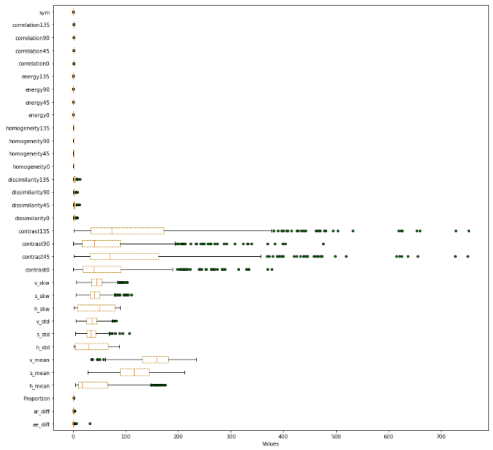

圖 22: 黑色素瘤各變數未經標準化盒形圖

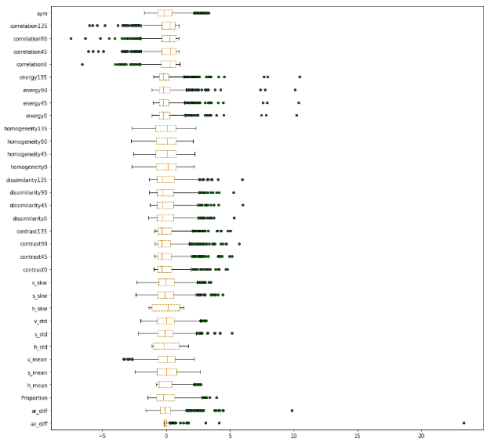

圖 23: 黑色素瘤各變數經標準化後盒形圖

3.6 維度縮減

擷取出來的部分特徵值存在高度相關，因此需要對其進行降維的動作。由圖 24 可以看出，灰度共生矩陣所產出的 5 個變數對比度 (contrast)、差異性 (dissimilarity)、同質性 (homogeneity)、能量 (energy) 及相關性 (correlation)，每個變數的 4 個角度間存在高度相關，故分別對其進行 PCA 維度縮減。

進行維度縮減後，產生 5 個新變數，分別為對比度 (conf1)、同質性 (homof1)、差異性 (disf1)、相關性 (corrf1)、能量 (energyf1)。

由圖 25 可以看出，hsv 三個通道各自的標準差及偏度存在高度相關，也分別對其進行 PCA 維度縮減。

進行維度縮減後，產生 3 個新變數，分別為色相標準差偏度 (h_ss)、飽和度標準差偏度 (s_ss)、明度標準差偏度 (v_ss)。維度縮減前後所有變數的相關係數圖參考附錄 D。

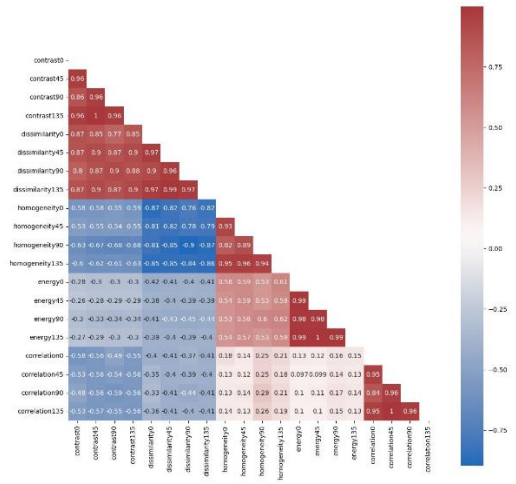

圖 24： 對比度、差異性、同質性、能量及相關性的相關係數圖

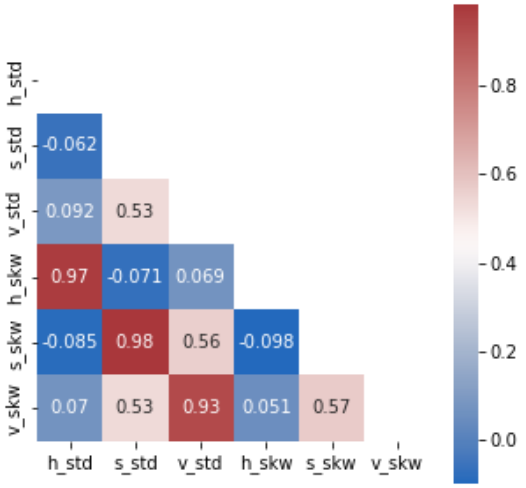

圖 25：s 和 v 的平均、標準差及偏度相關係數圖

3.7 模型開發

3.7.1 分類器模型介紹

本研究使用下述七種分類器。

1. 支援向量機（Support Vector Machine，簡稱：SVM）
2. 隨機森林（Random Forest）
3. 單純貝氏分類器（Naive Bayes classifier）
4. K-近鄰演算法（k nearest neighbor，簡稱 ：KNN）
5. 羅吉斯迴歸（Logistic Regression）
6. 神經網路（Neural Network，簡稱：NN）
7. 深度卷積神經網路（Convolutional Neural Network，簡稱 ：CNN）

詳細方法說明，請見附錄 E。

3.7.2 評估指標

本研究的評估指標分為兩個方面，統計量化指標和程式執行時間。

一、統計量化指標

各模型分析結果將呈現如表 1 形式，分為四個部分

(1) 真陽性 (TP)，代表該病徵實際為黑色素瘤且預測該病徵為黑色素瘤。

(2) 真陰性 (TN)，代表該病徵實際為非黑色素瘤且預測該病徵為非黑色素瘤。

(3) 假陽性 (FP)，代表該病徵實際為非黑色素瘤但預測該病徵為黑色素瘤。

(4) 假陰性 (FN)，代表該病徵實際為黑色素瘤但預測該病徵為非黑色素瘤。

依此建構各項指標評估統計方法分類表現之優劣，各項統計評估指標陳列於表 2

表 1: 模型分析結果

		實際狀況	
		+ (黑色素瘤)	- (非黑色素瘤)
表預測結果	+ (黑色素瘤)	TP	FP
	- (非黑色素瘤)	FN	TN

表 2: 各項統計評估指標

評估指標	定義	公式	意義
靈敏度 (Sensitivity)	真實狀態為陽性且其預測結果亦為陽性的比率	$\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}}$	測試正確識別疾病患者的能力
特異度 (Specificity)	真實狀態為陰性且其預測結果亦為陰性的比率	$\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$	測試正確識別未患病者的能力
準確率 (Accuracy)	預測正確占整體的比率	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$	測試正確識別所有患者的能力

二、程式執行時間

由於每個分類器模型的內部運行方式不同，因此運行時間也會所不同，故本研究亦將時間成本納入評估標準之一。

3.8 研究流程

完整研究流程，參考圖 26

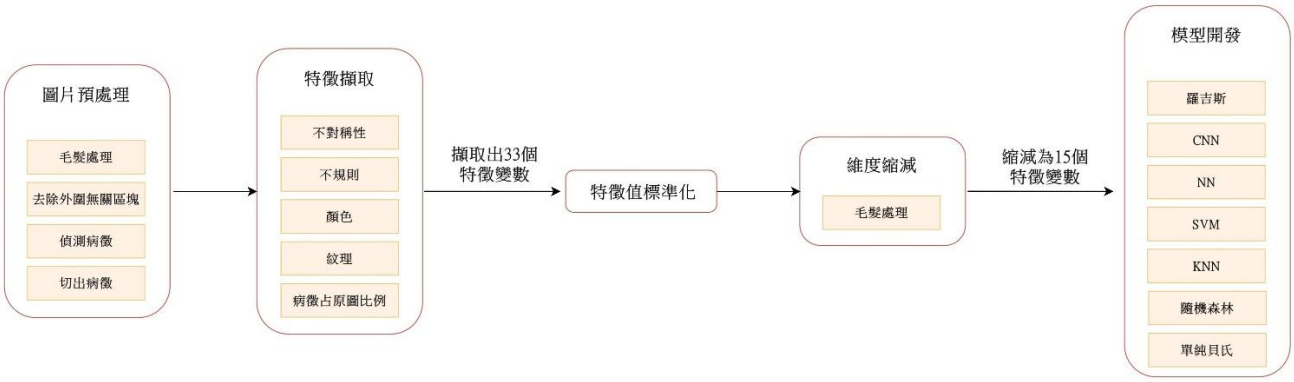

圖 26: 研究流程圖

4. 研究結果與討論

4.1 樣本抽樣對結果的影響

在進行樣本抽樣時，可能會出現樣本與母體之間存在差異的情況，這種差異會對研究結果的準確性和代表性產生影響，因此本研究將比較十組訓練集的差異，以證明非黑色素瘤圖片樣本的代表性。

本研究將十組訓練集放入七種分類器中訓練，並比較其結果，由圖 27 可以看出 10 組訓練集在各分類器的準確度四分位距差異不大，即 10 組訓練集經過各個分類器訓練後的準確度分布情況沒有顯著不同，表示非黑色素瘤樣本的取樣是具有代表性的。

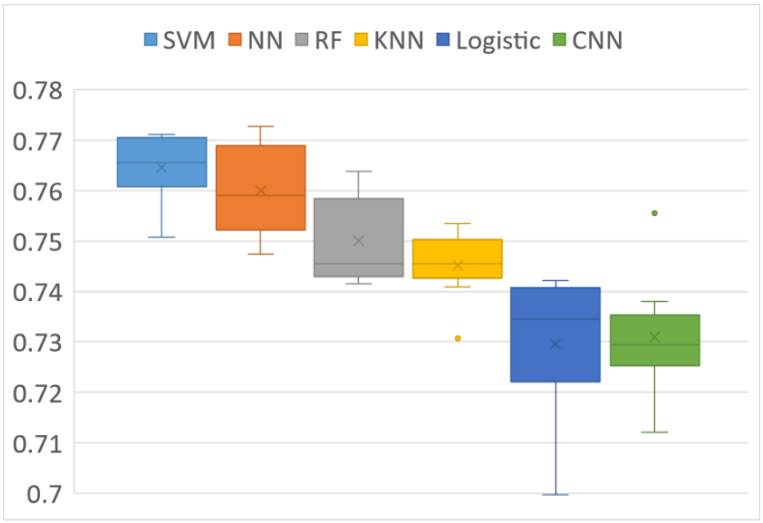

圖 27 : 10 組訓練集經過各個分類器訓練後的準確度盒形圖

4.2 各分類器模型訓練結果

4.2.1 評估指標分析

本研究對七種分類器做比較，表 3 是 10 個訓練集經過各種分類器訓練後的平均結果，表 4 是 10 個訓練集各自經過 10 個測試集測試後的平均結果，可以看出 :

一、準確度

七種分類器的準確度表現都相當良好，其中 SVM、隨機森林，以及 Neural Network 的表現特別優異。

二、程式執行時間

除了 NN 和 CNN 之外，其他 5 個分類器的程式執行時間都在 0.5 秒之內，而其中程式執行時間最短的分類器是單純貝氏分類器。

表 3: 各分類器模型訓練結果 (10 個訓練集的平均值)

分類器種類	特異度 (%)	靈敏度 (%)	準確度 (%)	執行時間 (秒)
SVM	84.1	84.4	84.2	0.21904
隨機森林	100	100	100	0.60724
單純貝氏	45.7	86.5	66.1	0.00090
KNN	100	100	100	0.02534
羅吉斯迴歸	76.2	73.7	75.0	0.00788
NN	100	100	100	12.40170
CNN	100	100	100	32.35060

表 4: 各分類器模型測試結果 (10 個訓練集各 10 個測試集的平均值)

分類器種類	特異度 (%)	靈敏度 (%)	準確度 (%)	執行時間 (秒)
SVM	75.8	77.	76.5	0.07623
隨機森林	73.2	76.8	75.0	0.20994
單純貝氏	57.5	70.3	63.9	0.00054
KNN	74.7	74.4	74.5	0.00684
羅吉斯迴歸	71.4	74.5	73.0	0.00573
NN	74.6	77.3	76.0	9.14600
CNN	76.6	68.4	72.6	32.35060

4.2.2 特徵重要性分析

一、SVM

SVM 使用線性核函數可以計算特徵權重，圖 28 是將特徵權重取絕對值再排序後的結果，即每個特徵對於分辨黑色素瘤的重要性如圖 28，其中紅色為正面影響，表示該變數越大，圖片被辨識為黑色素瘤的機率越大；藍色則為負面影響，表示該變數越大，圖片被辨識為黑色素瘤的機率越小。

由圖 28 得知，明度平均值 (v_mean)、 飽和度平均值 (s_mean)、差異性 (disf1)、圖片占比 (Proportion) 為使用 SVM 判斷黑色素瘤時較為重要的特徵。

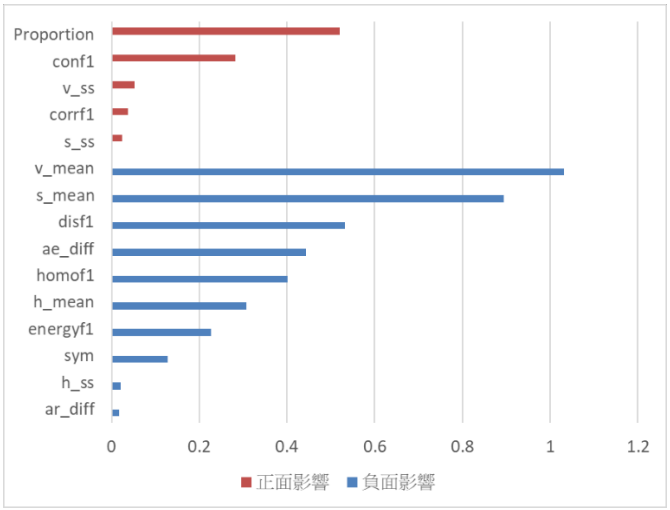

圖 28：SVM 特徵重要性長條圖

(長條圖的長度是十組訓練集的 SVM 特徵係數平均值)

二、隨機森林

隨機森林中的特徵重要性計算的結果是一個相對的分數，該分數衡量了每個特徵對於模型預測能力的貢獻。較高的 Gini 重要性值表示該特徵對於模型的預測有較大的影響，圖 29 為 10 組訓練集在隨機森林的 Gini 重要性平均值。

由圖 29 得知，在隨機森林模型中，差異性 (disf1)、同質性 (homo1)為使用隨機森林判斷黑色素瘤時較為重要的特徵。

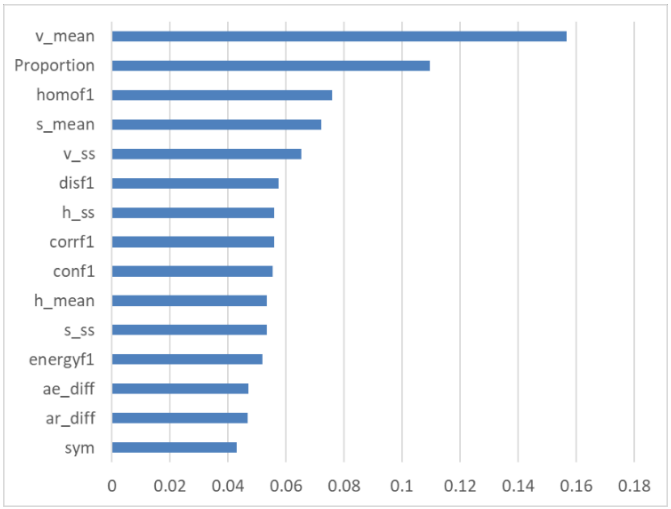

圖 29： 隨機森林特徵重要性長條圖

(長條圖的長度是十組訓練集的隨機森林特徵權重平均值)

三、羅吉斯迴歸

圖 30 為 10 組訓練集各變數的迴歸係數經標準化後的平均估計值，依據大小排序後可以得出各變數在羅吉斯模型中的重要性如圖 30，其中紅色為正面影響，表示該變數越大，圖片被辨識為黑色素瘤的機率越大；藍色則為負面影響，表示該變數越大，圖片被辨識為黑色素瘤的機率越小。

由圖 30 得知，不規則比例-橢圓 (ae_diff)、明度平均值 (v_mean)、差異性 (disf1)、同質性 (homof1) 為使用羅吉斯模型判斷黑色素瘤時較為重要的特徵。

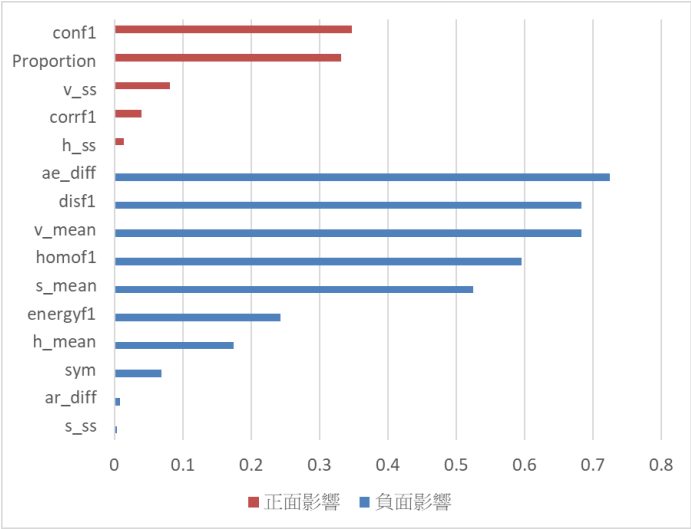

圖 30： 羅吉斯標準化係數估計平均值長條圖

(長條圖的長度是十組訓練集的標準化係數平均值)

4.3 討論

4.3.1 結果比較與討論

就分類器而言，綜合所有結果，SVM 不論是在靈敏度、特異度、準確度與程式執行時間皆表現優異。雖然單純貝氏分類器的執行時間最短，但準確度卻是所有分類器中最低的，推測可能是因為單純貝氏分類器的假設和本次研究的資料不適配。

在本次研究的分類器中，SVM、隨機森林和羅吉斯迴歸模型皆能看出各個特徵變數的重要性，整理各模型重要特徵變數如下表 5。

表 5：各分類器辨別黑色素瘤的關鍵特徵變數

分類器種類	關鍵特徵變數					
	明度平均值	飽和度平均值	不規則比例 - 橢圓	同質性	病徵佔原圖比例	不規則比例 - 橢圓
SVM	✓	✓	✓	✓	✓	
隨機森林	✓	✓		✓	✓	
羅吉斯迴歸	✓	✓		✓		✓

以下針對前述所提及之特徵擷取方向進行討論。

一、不對稱性

不對稱性在模型開發中皆為重要性偏低的變數，推測可能是不對稱性特徵擷取方式不夠精確，抑或是因為非黑色素瘤的種類太繁雜，存在許多不對稱的形狀，因此不對稱性對於辨別黑色素瘤可能不是一個適合的特徵變數。

二、不規則程度

不規則比例-橢圓為負向影響，在模型開發中為重要性偏高的變數，表示形狀越規則，辨別為黑色素瘤的機率越高，這與本研究預期結果不符，推測原因同上所述，因為非黑色素瘤的種類太繁雜，因此不規則程度對於辨別黑色素瘤可能不是一個適合的特徵變數。

三、顏色

明度平均值為負向影響，在模型開發中為重要性偏高的變數，表示明度平均值越低，即越靠近黑色，辨別為黑色素瘤的機率越高；飽和度平均值亦為負向影響，在模型開發中為重要性偏高的變數，表示飽和度平均值越低，即色彩越不鮮豔，辨別為黑色素瘤的機率越高。

四、紋理

同質性為負向影響，在模型開發中為重要性偏高的變數，表示同質性越低，即紋理變化較大、平滑性較低，辨別為黑色素瘤機率越高。

五、病徵占原圖比例

病徵占原圖比例為正向影響，在模型開發中為較重要的變數，即病徵占比越高，辨別為黑色素瘤的機率越高，表示病徵在圖片中的比例會影響到黑色素瘤的辨別，後續研究應將此

因素做為控制變量列入考量。

綜合上述，明度平均值、飽和度平均值和同質性為與預期結果相符合的特徵變數，因此本研究認為顏色與紋理是辨別黑色素瘤的關鍵特徵變數。不對稱性在各模型中的重要性皆低，不規則程度為與預期結果不符之特徵變數，因此本研究認為不對稱性和不規則程度較不適合做為分辨黑色素瘤的特徵變數。

4.3.2 與其他研究結果之比較

大多研究對於黑色素瘤的分類皆是將大量的原始圖片直接套入深度學習模型中做分類，如 (Edubirdie , 2022) [14]，但這樣的作法不僅要花費許多醫療成本來取得圖片，並且還需要有較高規格的電腦設備才能進行模型訓練，否則會花費較多執行時間。再者，由於深度學習的特徵解釋不易，因此即便使用深度學習模型訓練出了很好的成果，也依然沒有辦法找出足夠清楚判別黑色素瘤的特徵。

部分研究使用深度學習模型以外的分類器做分類，使用 MED-NODE 數據集 (Sultana,)，並用 SVM 進行分類的最高準確度為 77.1%，但此研究用來訓練模型的圖片皆為專業儀器所拍攝之清晰圖片，如圖 31，但若使用較差的圖片進行分辨，如圖 3 和圖 7，其分類效果有限。

圖 31： 專業儀器所拍攝之黑色素瘤

本次研究將雜訊較多的原始圖片經過預處理後擷取出病徵，並以 ABCDE 法則為基礎提取特徵，並且同樣使用 SVM 進行分類的狀況下，得到 76.5% 的準確度。如此一來，不僅可以利用少量、病徵不明顯且雜訊較多的圖片進行分析，還能降低醫療成本及執行時間，並對特徵進行解釋，進而找出分辨黑色素瘤的關鍵特徵。除此之外，本研究還針對數據平衡以及抽樣的穩定性進行處理，使本研究之結果更有參考價值。

5. 結論

在經過各種分析以及嘗試之後，本次研究加入了自動化裁切圖片方法，精確抓出各種圖片中的病徵位置，並以 ABCDE 法則為基礎提取特徵，對特徵進行解釋。利用測試集所得到準確度最高的模型是 SVM 的 76.5%，雖然準確度不及其他利用 CNN 進行分類的研究，但本研究找出了辨別黑色素瘤的關鍵特徵，即顏色與紋理，不僅縮小了後續黑色素瘤的研究範圍，還能提供民眾更明確的黑色素瘤分辨依據，以便及早治療，提高存活率。未來的研究可以針對顏色及紋理進行更進一步的探討，尋找更多相關特徵，增加成功分辨黑色素瘤的機率。期許未來自動診斷能完全取代現今的臨床皮膚檢測，成為更好的檢測依據。

6. 附錄

A)　ABCDE 法則

一、不對稱性 (Asymmetry)

看特徵中心兩邊的對稱性。良性的色素性病徵多是對稱的，也就是說，如果在病徵中央劃一條線，左半和右半應該是對稱的。

二、邊緣 (Border)

為特徵的邊緣，觀察邊緣有無不規則的突起或凹陷。黑色素瘤的邊緣大多呈現不規則狀，如圖 32，而一般皮膚上痣的邊緣則多為平滑的圓形或橢圓曲線，如圖 33。

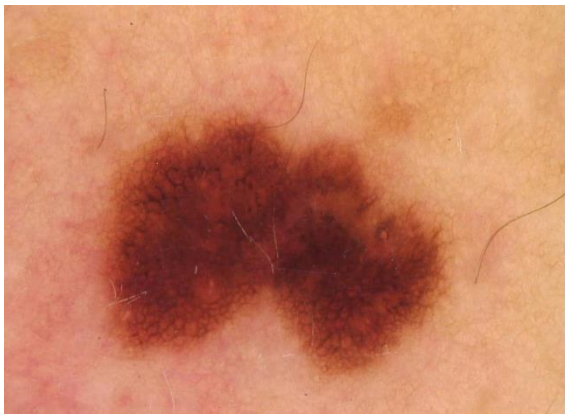

圖 32 :邊緣呈現不規則狀的黑色素瘤

圖 33： 一般皮膚上的痣

三、顏色 (Color)

為特徵的顏色，以特徵的顏色是否均勻分布作為判斷依據。黑色素瘤顏色分佈較不均勻，在黑褐色的基礎上容易有藍色或紅色的斑點，斑點可能或大或小，也沒有特定分佈區域，如圖 34；痣的顏色則大多分佈平均，呈現整塊的黑褐色，如圖 35。

圖 34: 黑色素瘤顏色分佈

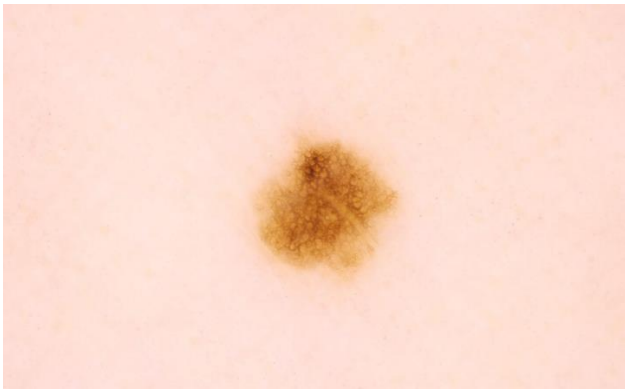

圖 35： 痣的顏色分佈

四、直徑 (Diameter)

為特徵的直徑大小，若大於 6 mm 則高機率為黑色素瘤。

五、擴展 (Enlargement)

特徵的變化，包含形狀變大、體積增加、顏色改變等。

B) hsv 色彩三屬性

hsv 色彩三屬性圖，參考圖 36。

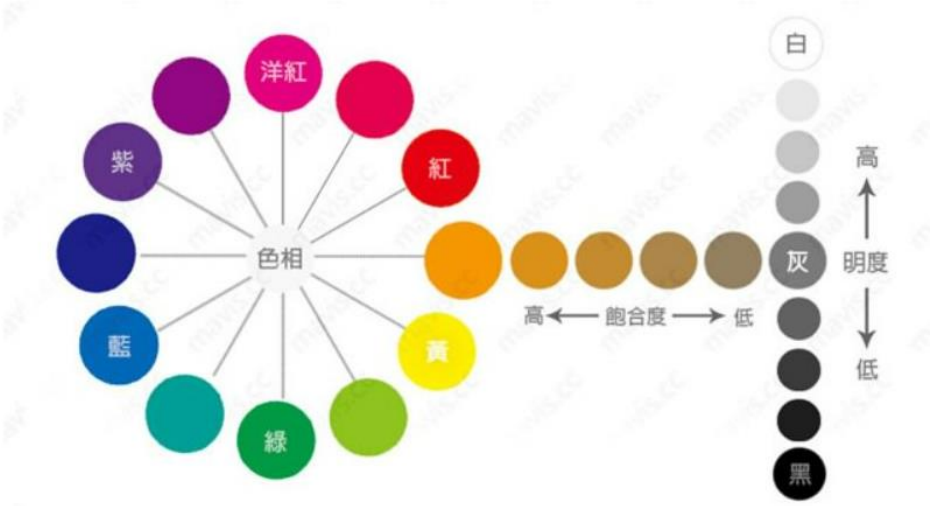

圖 36 : hsv 色彩三屬性圖

C) 紋理

黑色素瘤除了上述的顏色分佈不均，其表面由於病變還更容易有不規則的高低起伏。如圖 37，黑色素瘤的表面大多有不規則的突起；一般痣的表面則較平滑，如圖 38。而顏色不均和表面凹凸不平，都可能造成病徵表面的紋理多而複雜。相較於顏色統一且表面光滑的痣，黑色素瘤的紋理普遍較明顯且複雜。

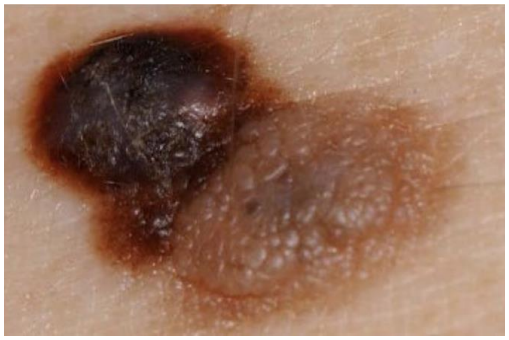

圖 37： 黑色素瘤表面

圖 38： 痣的表面

為了分析紋理分佈的情況，我們計算圖片的灰度共生矩陣來做後續分析，而關於灰度共生矩陣的簡短介紹如下：

灰度共生矩陣兩軸的單位皆為像素值，範圍是 0-255。(0 代表黑，255 代表白）而 GLCM（i , j）的值，代表原彩色圖中像素為 (i、j) 組合的成對點有幾組，而相鄰成對會根據 0 度、45 度、90 度、135 度共四個方向計算，所以一張圖最後會有四個角度的灰度共生矩陣。

此外，如果原圖轉成灰度圖後相鄰像素點相同的組合過多，即 (i , i) 的組合過多，則灰度共生矩陣的對角元素會有比較大的值；如果原圖轉成灰度圖後在局部變化較大，則灰度共生矩陣的對角元素會有較小的值。

觀察圖 39 圖 40，圖 40 是將圖 39 模糊化的同一張照片 ，因為模糊化導致圖 40 的紋理沒有圖 39 清晰。因此圖 42 的對角線比圖 41 更加明顯。

圖 39： 舉例用的原始圖片

圖 40： 模糊後的圖 39

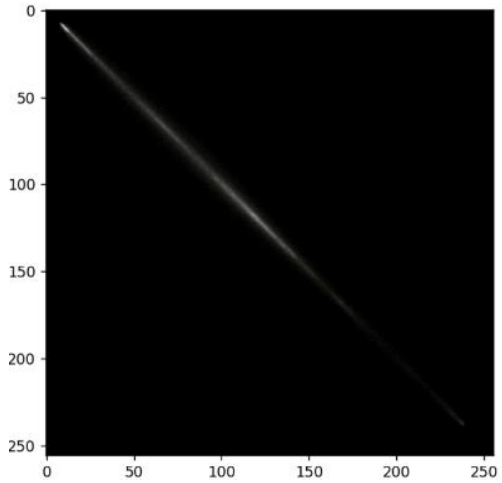

圖 41： 圖 39 的灰階共生矩陣

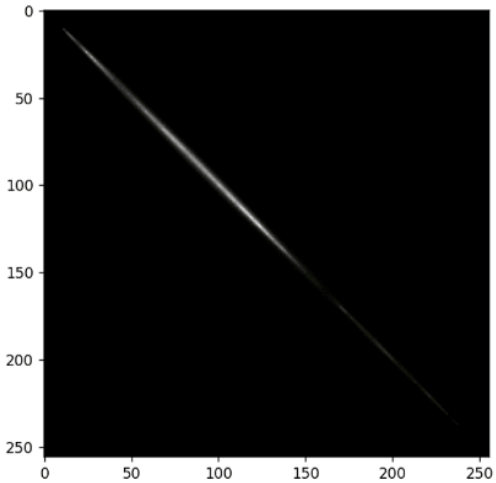

圖 42： 圖 40 的灰階共生矩陣

做出灰度共生矩陣後，可以透過它計算五種統計量幫助我們提取紋理特徵，先令 p (i , j)為 (i , j) 像素對應其灰度共生矩陣之值。

一、對比度 (contrast)

反應圖像的清晰度以及紋理溝紋的深淺。紋理越深，對比度越大，視覺效果越清晰；反之，對比度越小，紋理溝紋越淺，視覺效果越模糊。

$$Contrast = \sum_{i,j}(i-j)^2 p(i,j) \ \ ; i,j = 0,1\ldots255$$

二、相關性 (correlation)

計算圖像灰度值在行和列方向的相似程度，相似程度越大，視覺較清晰，則相關性越大，灰度矩陣中的灰度值相差大，則相關值越小。

$$Corr = \sum_{i,j}\frac{(i-u_i)(j-u_j)p(i,j)}{\sigma_i\sigma_j} \ \ ; i,j = 0,1\ldots255$$

三、能量 (energy)

反應圖片灰度分佈的均勻程度和紋理粗細度。若灰度共生矩陣中的元素之值分佈均勻，則能量較小，表示紋理細緻；若其元素之值分布不均且差異大，則能量較大，表示原圖的紋理有一定程度的規則變化。

$$Energy = \sum_{i,j} p(i,j)^2 \ ; i,j = 0,1\ldots255$$

四、差異性(Dissimilarity)

感興趣區域（像素）之間距離的度量，用於評估紋理不規則程度，差異性越高代表有明顯的不規則紋理區域；差異性越低則紋理區域較均勻且相似。

$$Dissimilarity = \sum_{i,j} p(i,j) * |i-j| \ ; i,j = 0,1\ldots255$$

五、同質性(homogeneity)

衡量紋理的同質程度，可以反應紋理局部變化之大小，值大代表紋理不同區域較少變化且均勻，即紋理平滑性較高；同質性低，則紋理局部變化較大，且紋理平滑性低。

$$Homogeneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \ ; i,j = 0,1\ldots255$$

由於有四個角度所以一張圖會有四個灰度共生矩陣，即每一角度一個，故共會有 20 個統計量。

D)　維度縮減前後所有變數的相關係數圖

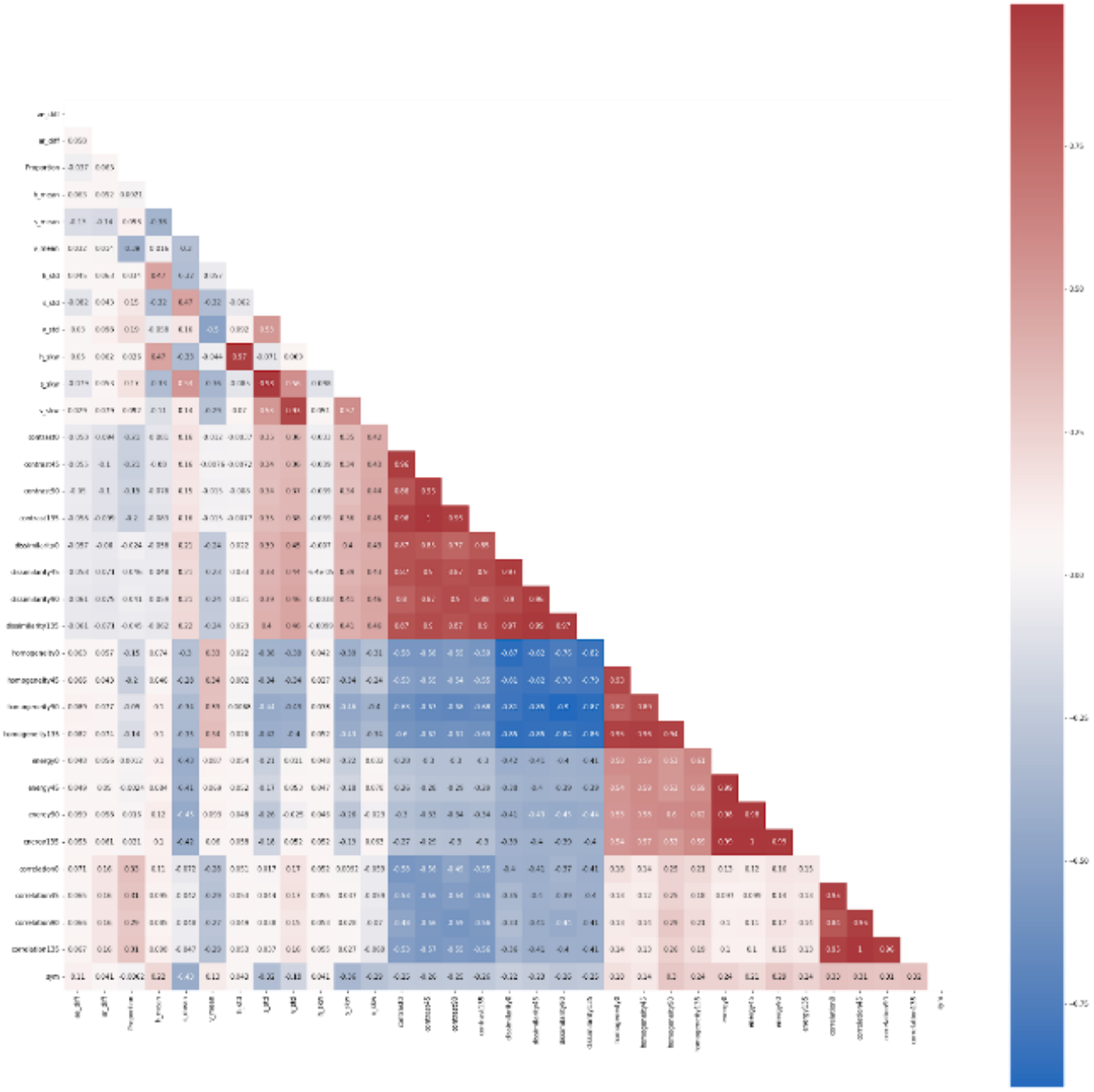

圖 43： 維度縮減前所有變數的相關係數圖

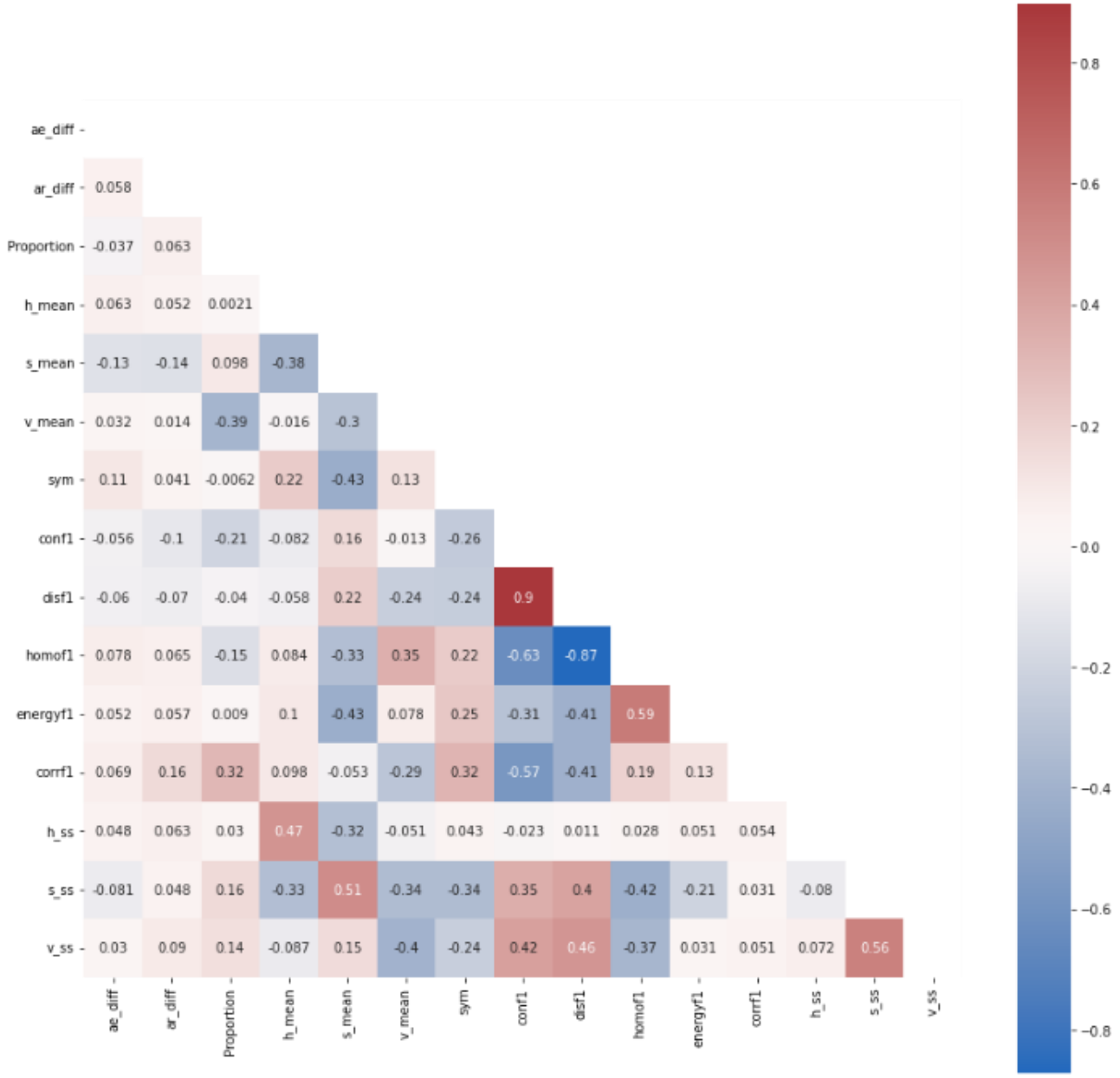

圖 44： 維度縮減後所有變數的相關係數圖

E) 模型介紹

一、支援向量機 (SVM)

是一種監督式學習，透過將資料點映射在高維空間中並建構超平面，以進行分類、迴歸或其他任務。給定一組訓練集，每個資料點會被標記為屬於兩個類別中的其中一類，SVM 假設存在一個超平面 $\vec{w}^{\mathrm{T}}\vec{x} + \vec{b} = 0$ 可以完美分割兩組資料，透過計 $\vec{w}$、 $\vec{b}$ 使得兩類之間的邊界最大化，並將資料重新分類為兩個類別之一。值得注意的是，需要先對資料的每個特徵進行標準化，使資料點的每個特徵在 SVM 裡同等重要。

二、隨機森林 (Random forest)

屬於集成學習 (Ensemble learning) 法。訓練分類器時，建造多個分類樹 (Decision Tree)，再以拔靴法 (Bootstrap) 抽樣進行訓練，最後採多數決決定分類結果。樹的分支則是利用熵 (Entropy) 或吉尼不純度 (Gini impurity)，尋找特徵子集內擁有最大資訊增益 (Information Gain, IG) 的變數，進而停止分割節點。與單一分類樹相比，此方法較能降低偏誤及變異，且不易過度擬合。

三、單純貝氏 (Naive bayes)

為一種構建分類器的方法。它不是訓練這種分類器的單一演算法，而是一系列基於相同原理的演算法：所有單純貝氏分類器都假定樣本特徵間彼此獨立。在許多實際應用中，單純貝氏模型參數估計使用最大概似估計方法，換言之，在不用到貝氏機率或者任何貝氏模型的情況下，單純貝氏模型也能奏效。儘管假設很簡單，但單純貝氏分類器在很多複雜的現實情形中仍能夠取得相當好的效果。優勢在於只需根據少量的訓練資料，即能估計出必要的參數（變數的均值和變異數），且由於變數獨立假設，所以不需要確定整個共變異數矩陣。

四、K-近鄰演算法 (KNN)

又稱 KNN 演算法，屬監督式學習。訓練資料為具有標籤的資料，計算每個樣本點與目標點之歐式距離，選 k 個最近點並做類別判斷。判別方式分為兩種：投票法及加權投票法，分類結果為類別最多的組別，兩者差別為後者考慮近鄰之距離遠近進行加權，再進行分類。優點是精度高且無資料輸入假定、資料型態不受限，缺點是時間空間複雜度高，樣本平衡度依賴高。

五、羅吉斯迴歸 (Logistic Regression)

主要在探討依變數與自變數之間的關係，與一般線性迴歸不同處在：依變數為類別變數，特別是分成兩類的變數。優點是方便計算與訓練，且自變數不需常態分佈的假設，故限制較少。

六、神經網路 (NN)

屬於非線性的統計模型，用神經元傳遞信號，將資訊從輸入層匯入人工神經網路，中間經過隱藏層的非線性轉換，在經過層層的隱藏層後，最終輸出層給出人工神經網路預測

的反應變數。

七、深度卷積神經網路 (CNN)

CNN 相較於神經網路有權值共享的優勢，利用相同幾個神經元組成的卷積核去學習相同的特徵，進而達到節省參數的效果。CNN 首先透過卷積核(Kernels)滑動對圖像做訊息提取，並藉由步長(Strides)與填充 (Padding)控制圖像的長寬，再經由池化層壓縮圖片並保留重要資訊，之後使用攤平(Flatten)來銜接 CNN 層與全連接層，最後在全連接層做特徵提取。

7. 參考文獻

1. 美國癌症協會 (https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html)
2. 美國皮膚科學會 (https://www.aad.org/media/stats-skin-cancer)
3. Majumder, S. and Ullah, M. A. (2019). Feature extraction from dermoscopy images for melanoma diagnosis., SN Applied Science, vol. 1, 753.
4. 余佳蓉 (2021)。智慧醫療領域對黑色素瘤的生物醫學影像分析。明志科技大學電機工程系碩士論文。
5. Gessert, N., Nielsen, M., Shaikh, M., Werner, R. and Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX, 7, 100864.
6. Rehman, M., Ali M., Obayya M., Asghar J., Hussain L., K. Nour M., et al. (2022). Machine learning based skin lesion segmentation method with novel borders and hair removal techniques. PLoS ONE, 17 (11), e0275781.
7. ADDI Project. (2013). PH² Database. [Data file]. (https://www.fc.up.pt/addi/ph2%20database.html)
8. Giotis, N. Molders, S. Land, M. Biehl, M.F. Jonkman and N. Petkov: "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images", Expert Systems with Applications, 42 (2015), 6578-6585.
9. ISIC Challenge. (2018). ISIC Challenge Datasets. [Data file]. (https://challenge.isic-archive.com/data/#2018)
10. Kaggle. (2020). SIIM-ISIC Melanoma Classification. [Data file]. (https://www.kaggle.com/competitions/siim-isic-melanoma-classification/data)
11. OpenCV. (n.d.). Retrieved May 27, 2023, from https://docs.opencv.org/3.4/db/df6/tutorial_erosion_dilatation.html
12. OpenCV. (n.d.). Retrieved May 27, 2023, from

https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html

13. Murzova, A. and Seth, S. (2020). Otsu's Thresholding with OpenCV. Retrieved May 27, 2023, from https://learnopencv.com/otsu-thresholding-with-opencv/
14. Edubirdie. (2022). Skin Cancer: Convolutional Neural Network Based Skin Lesion. Retrieved May 28, 2023, from https://edubirdie.com/examples/skin-cancer-convolutional-neural-network-based-skin-lesion/

專題競賽報告分工表

學號	姓名	分工內容
410978004	郭依璇	資料與文獻搜集、流程構想與安排、程式構想與製作、報告撰寫、PPT 製作
410978011	譚靖蓉	資料與文獻搜集、流程構想與安排、程式構想與製作、報告撰寫、PPT 製作
410978044	謝瑋芸	資料與文獻搜集、流程構想與安排、程式構想與製作、報告撰寫、PPT 製作
410978056	黃名揚	資料與文獻搜集、流程構想與安排、程式構想與製作、報告撰寫、PPT 製作
410978058	李博業	資料與文獻搜集、流程構想與安排、程式構想與製作、報告撰寫、PPT 製作