科技部補助

大專學生研究計畫研究成果報告

計　畫
名　稱：　喉內視鏡影像物件切割且具評估聲帶病灶指標產生系統

報 告 類 別 ：成果報告
執行計畫學生：陳怡妙
學生計畫編號：MOST 110-2813-C-040-097-E
研 究 期 間 ：110年07月01日至111年02月28日止，計8個月
指 導 教 授 ：秦群立

處 理 方 式 ：本計畫可公開查詢

執 行 單 位 ：中山醫學大學醫學資訊學系

中 華 民 國　111年03月25日

(一) 摘要

現今耳鼻喉科醫師在臨床上診斷聲帶相關病症時，會使用喉內視鏡攝影儀觀測病患病灶表徵與聲帶振動狀況，但在拍攝時容易受到晃動或聲帶振動的影響，導致醫師無法在拍攝的同時觀測到喉內視鏡影像中重要的細部病理特徵，並即時進行診斷。而醫師目前皆透過自身經驗進行聲帶相關病症的診斷，多數喉部相關研究並未提供客觀的喉部物件數據給醫師作為參考依據，導致醫師看診時間過長。因此，本研究先透過 3D VOSNet 模型抽取喉內視鏡視訊檔中空間特徵，並利用此模型物件遮擋不變性以及平移不變性的特性，來分類喉內視鏡序列影像中各像素是否為目標物件，根據實驗結果顯示，本研究在左聲帶、右聲帶及聲門的分類準確度分別為 93.48%、94.63%以及 89.91%，表示本研究能有效從序列影像中切割出聲帶及聲門區域。最後，會利用本研究提出的聲帶指標產生演算法計算出具評估聲帶病灶的各項指標數值，包含聲帶長度、聲帶與聲門面積、聲帶長度與面積偏差、聲帶曲率以及左右側聲帶震動的對稱性，透過提供即時且客觀數據給醫師進行聲帶相關病症的診斷，提升整體聲帶治療成效與醫病品質。未來希望可以將此方法應用於其他領域，以利醫師臨床醫學診斷。

關鍵字：喉內視鏡影像、聲帶振動、3D VOSNet 模型、空間特徵、指標數值。

(二)研究動機與研究問題

人們經常透過聲音來與他人進行溝通與傳達內心的情緒，然而大多數的人對於聲帶的保養並不會特別的注重，再加上生活壓力大、睡眠不足以及抽菸酗酒習慣，都有可能會加重聲帶的受損，甚至是聲帶的病變。然而聲帶受損對於人們日常生活中會帶來諸多不便，像是無法正常發聲導致溝通表達上的問題，甚至病患可能會因聲帶萎縮或麻痺造成聲帶閉合不全而產生吞嚥困難[1]，嚴重者可能因此喪命。正常聲帶組織、聲帶瘜肉與閉合不全之聲門影像如圖 1 所示。

(a) (b) (c)

圖 1、(a)為正常聲帶組織，(b)為聲帶瘜肉，(c)為閉合不全之聲門影像

病患若出現喉部相關病灶到醫院求診時，耳鼻喉科醫師會透過喉內視鏡攝影儀觀測患者喉部內部的聲帶[2]，檢查聲帶的運動狀況以及觀察聲門閉合程度等。然而，由於人眼對於動態物體的辨識力較為不足，且可能會因拍攝視訊檔晃動導致聲帶被其他組織遮擋，因此，醫師難以透過喉內視鏡影像直接觀測到重要的細部紋理特徵。而引起聲帶受損的病因有許多種，包含聲帶瘜肉、聲帶囊腫以及聲帶麻痺[3]等疾病，有時還需搭配侵入性的喉肌電圖進行診療[4]，然而仍會出現以下狀況：

1. **拍攝喉內視鏡影像時可能會因晃動導致聲帶物件被喉部組織遮擋，影響診斷**
2. **由於尚未有具聲帶病灶評估的相關指標數據供耳鼻喉科醫師參考，醫師只能透過人眼觀察喉內視鏡拍攝影像並加以診斷**
3. **侵入性的喉肌電圖檢測，對於病患而言相當痛苦且接受度不高**
4. **目前尚未有任何測量工具及指標提供給醫師參考，醫師為了更精準的檢測病灶，經常需花費較長時間觀測喉內視鏡拍攝的影像，就診時間長**

為了輔助醫師於臨床上有更精確的診斷，使病患可獲得最佳且即時的治療，本計畫將著重於如何透過喉頻閃攝影視訊檔計算出有效的醫療指標，協助醫師診斷病症，提升整體治療成效與醫療品質。此外，為了使本計畫所開發聲帶指標方法能更具臨床應用價值，研究中將會參考目前耳鼻喉科醫師使用的醫療系統介面以建置出喉內視鏡影像指標產生系統介面，希望本計畫除了能輔助醫師醫療診斷，還能符合耳鼻喉科醫師實際使用需求，期望達到以下目標：

1. **透過 3D VOSNet 具上下文資訊的模型將視訊檔中的聲帶物件進行切割，此模型不因物件遮擋而影響其切割準確率**
2. **自動將喉內視鏡拍攝到的影像進行聲帶的切割，左聲帶區域、右聲帶區域及聲門區域，並以視覺化的方式顯示來輔助醫師診斷。**
3. **利用大量分析資料及強大的神經網路自動計算出可靠聲帶病灶評估的相關指標，包含聲帶曲率、聲帶長度偏差以及聲門區域，並將計算出的數據提供給醫師參考，透過數據分析病灶原因，以減少侵入性喉肌電圖的檢測**
4. **利用人工智慧技術快速分析及精準辨識，提供醫師參考資訊，協助醫師加速問診的時間，並有效減少病患等待的時間**

(三)文獻回顧與探討

近年來人工智慧技術日益蓬勃且廣泛應用於醫療領域，許多學者紛紛投入發展判斷醫學影像的電腦輔助診斷，以提升廣大民眾的醫療品質。然而，目前深度學習於聲帶指標之技術尚未有學者針對這方面進行研究，本計畫希望參考多篇文獻所提出之想法，加強或改善其作法，使本計畫提出之方法更加完善且精準。本計畫將探討的文獻分為四大類，以下將逐一進行介紹。

1. 現今對於聲帶辨識的相關研究

在 2020 年，Matava 學者等人[5]透過 ResNet、Inception 和 MobileNet 等卷積類神經網路(CNN)即時將喉內視鏡及支氣管鏡拍攝到影像中聲帶及氣管進行分類，並進行三種模型的比較，發現表現最好的為 ResNet 及 Inception，其特異度分別達到 0.985 及 0.971，靈敏度分別為 0.865 及 0.892，但兩者模型皆只能分類出大致位置，較無法將實際聲帶輪廓進行描繪標記。

在 2020 年，Ren 學者等人[6]利用 ResNet-101 模型對拍攝到的影像進行分類，分別為正常聲帶組織、聲帶結節、聲帶白斑、良性息肉和惡性腫瘤，在辨識正常聲帶組織與不同類型的聲帶病症，其準確率高達 90%以上。然而，對於影像被遮擋時病灶辨識效果不佳。

2. 序列影像切割相關研究

在 2018 年，Xu 學者等人[7]認為模型長期學習時間與空間特徵對於許多影片分析是相當重要的，因此他們使用 LSTM 架構對序列影像中的物件進行切割，以此方法抽取影片中的時空特徵進行影像物件切割，根據他們的研究數據顯示，此方法準確率比一般只針對單張影像進行物件切割的方法高，但此模型在標物與背景顏色相近的情況表現不佳，容易切割到非目標物的區域。

在 2019 年，Duarte 學者等人[8]提出 CapsuleVOS 半監督式學習模型進行影片物件切割，他們利用影片為序列影像的特性，透過提取影片中上下文資訊，切割出影片物件的位置，根據研究顯示，他們提出的影片物件切割在小物件的切割以及物件被遮擋等情況表現佳。然而，此模型在物體移動或偏移時，對於物體邊緣的切割效果不佳，且容易出現目標物未被完整切割。

同年，Kao 學者等人[9]提出一種新穎且有效的方法來進行腦腫瘤的切割。此論文在腦部 MRI 所建立的 MNI 空間座標系統中，將不同類型的腦腫瘤病變生成熱圖(heat map)，再透過熱圖建立出感興趣體積圖(Volumes-Of-Interest map, VOI map)，最後 VOI map 與多模態 MRI 影像一併輸入至 3D UNet 模型進行腦腫瘤的切割，其實驗結果顯示 Kao 學者等人提出的方法可有效切割腦腫瘤，證明若能將病灶的特徵融合並使用 3D UNet 模型可有效從序列影像中切割出病灶，但此方法僅適用於有拍攝 MRI 影像之病灶。

在 2020 年，Xiao 學者等人[10]提出一種基於 3D UNet 和 Res2Net 的 3D Res2UNet，用於切割 CT 影像中的肺結節。此論文在實驗結果的部分提到 3D Res2UNet 具有強大的提取多尺度特徵之能力，使 3D Res2UNet 能找出 CT 影像中更細微的肺結節特徵，其召回率達到 99.1%、Dice 係數指數達到 95.30%，證明在 CT 影像中使用 3D Res2UNet 偵測與切割肺結節的結果比單獨使用 UNet 或是 3D UNet 的準確率更好，但此論文僅進行 3D UNet 結合 Res2Net 的實驗，並未探討結合其他殘差網路來進行特徵抽取的實驗。

在 2021 年，Yang 學者等人[11]提出 MSDS-UNet 模型用於自動分割 CT 中的肺腫瘤，他們利用此模型將肺片切片序列影像進行切割，透過切片的前後文資訊，有效的切割出肺部腫瘤，已經證明了它們相對於傳統醫學圖像分割算法的有效性和優越性，且不會因目標物與背景顏色相近或是目標的的移動而影響其效果。

3. AI 模型應用於識別及分類的相關研究

在 2017 年 Xie 學者等人[12]提出一種用於影像分類的 ResNeXt 網路架構。此網路是建構出一個模組，並將依據此模組建構出多個相同架構的模組，再聚合成一個同架構、多分支的網路架構，並將分支的數量稱為基數。此論文在實驗結果的部分除了說明增加基數能提高模型分類的準確率外，還比較 ResNeXt 與 ResNet 在 ImageNet-5K 資料集與 COCO 資料集上的結果，證明 ResNeXt 在自然影像上的分類結果比 ResNet 優異，但此論文並未探討 ResNeXt 及 ResNet 於醫學影像上的分類結果。

在 2021 年 Zhou 學者等人[13]研究了在不同環境與不同長度的語句中，利用 ResNeXt、Res2Net 和 ResNet 驗證說話者身分的結果差異。此論文在 VoxCeleb 資料集上的實驗除了證明 ResNeXt 和 Res2Net 驗證說話者身分的準確率高於傳統的 ResNet 模型之外，還證明 ResNeXt 和 Res2Net 在嘈雜環境中仍能有效識別說話者的身分，並且不會受到語句長度的差異而影響辨識結果，但此論文是建立在驗證三位不同說話者身分的基礎上進行實驗，並未探討 ResNet、ResNeXt 和 Res2Net 於更多不同說話者的環境中，識別說話者身分之結果。

4. 聲帶指標相關研究

在 18 及 19 世紀，有諸多學者提出可用於進行聲帶相關研究，以及作為病灶與病情變化的重要分析指標。Omori 學者等人[14]提出測量聲門面積，且透過聲門面積除以聲帶長度的平方，用來取得標準化後的聲門面積，並提出標準化的聲門面積可作為聲帶的相關指標。根據 Woodson 學者等人[15]的研究中提出測量聲門角度來反映單聲帶麻痺麻痺側聲帶的偏差比偏向化，最後他們還提出計算聲門角度可作為聲帶診斷的依據。此外，在 Casiano 學者等人[16]的研究中顯示，若聲帶腫瘤範圍擴大會提高後續手術的失敗率，因此可透過聲帶區域面積來作為推定聲帶腫瘤手術是否成功。

在 2020 年，Cho 學者等人[17]為了避免不同觀察者對於喉內視鏡影像中聲帶運動狀況診斷上有所不同，因此利用 CNN 網路架構對喉內視鏡拍攝到的影像進行二元分類，透過深度學習模型抽取出影像中細微且重要的資訊，來輔助醫師診斷。同年，Zhang 學者等人[18]由於目前對於聲音障礙的診斷在很大程度上取決於醫師的經驗，因此他們希望透過機器學習方法，計算出聲帶的幾何形狀、韌性、位置與聲門下壓力，透過計算出的數據，可有效幫助耳鼻喉科醫師於臨床上對於聲帶障礙的診斷。

綜合上述文獻回顧，發現目前提出的方法只能針對一張影像進行聲帶相關病灶的辨識且對於物件遮擋時成效不佳。而 CapsuleVOS 架構在物件被遮擋時也有相當不錯的表現，但其在物件邊緣上的切割效果不佳，而 3D UNet 架構可以利用序列影像中的上下文資訊，學習連續序列影像中的目標物件特徵，以準確地切割出物件區域，並不會受到物件位置改變而影響模型的物件切割能力，因此符合本計畫對喉內視鏡視訊檔聲帶物件切割之需求。此外，根據上述文獻回顧，我們發現 ResNeXt 無論是在自然影像還是時間序列影像上，其分類目標物件之結果皆非常優異，並且 ResNeXt 結合殘差網路以及 Inception 結構作為特徵抽取網路，除了能夠避免過擬合及提升效率外，還能具有提取多尺度特徵之能力。因此本計畫參考 3D UNet 及 ResNeXt 網路架構，在 3D UNet 的 encoder 網路中使用 ResNeXt，以提取喉內視鏡視訊檔之序列影像中的細微特徵，提升 3D UNet 切割聲帶及聲門等物件之準確率。最後由於目前臨床醫療上尚未出現喉內視鏡視訊檔影像聲帶相關的客觀數據分析系統，因此本計畫期望提出視訊檔之分析演算法，輔助耳鼻喉科醫師作出醫療診斷，提升醫療品質。

(四)研究方法及步驟

本計畫利用 3D VOSNet (Video Object Segmentation through 3D UNet)方法建置出「喉內視鏡影像物件切割且具評估聲帶病灶指標產生系統」，其中分為訓練、測試以及聲帶指標分析等三大部分，計畫流程圖如圖 2 所示。

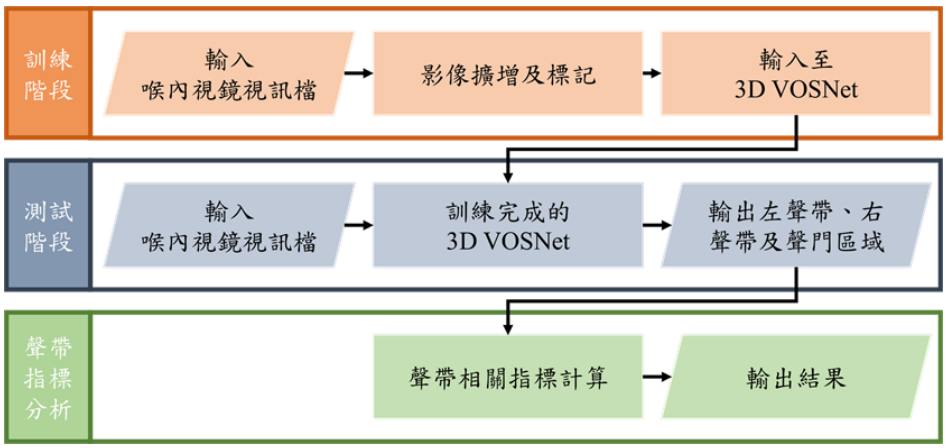


圖 2、計畫流程圖

1. 輸入喉內視鏡視訊檔資料集

本計畫中使用的喉內視鏡視訊資料集是由台中榮民總醫院耳鼻喉科醫師提供，而為了提高模型的泛化能力，本計畫也蒐集多個線上開放資料集，包括[15]研究中所開放提供的「Quantitative Laryngoscopy」的大型公開喉內視鏡視訊資料集，以及 YouTube 上經影片持有者授權的喉內視鏡視訊檔，而這些視訊檔來自於世界各地的醫療中心或醫院，在此資料集中共包含 50 個喉內視鏡視訊檔。為了避免因訓練資料不足導致模型標記準確率效果不佳，再加上耳鼻喉科醫師在診斷病灶時，會透過喉內攝影儀觀測病患聲帶的振動情形，所拍攝出的視訊檔會因拍攝角度不同或醫師拍攝時所造成拍攝儀器的晃動而導致亮度不同或模糊。因此，本計畫中會將影像進行銳利度及對比度的調整，以此來進行喉內視鏡視訊檔擴增，擴增完後的視訊檔共有 250 個，每個視訊檔片長皆約為 10 秒，且以 fps 為 30 進行擷取影像，每個影片皆擷取出 256 張影像，共有 64000 張影像。最後，本計畫會將擴增後的資料集進行模型的訓練與測試，其中 80%的資料用於模型的訓練，其餘 20%的資料用以評估該模型的準確率。

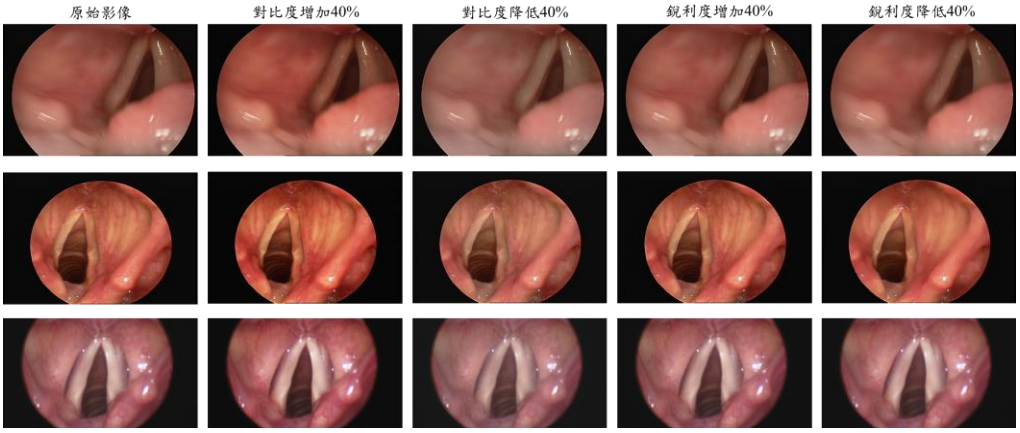


圖 3、對比度及銳利度調整後之影像

2. 手動標記左右聲帶位置及聲門區域

為了標記喉內視鏡視訊影像，本計畫請專業耳鼻喉科醫師協助標記轉換為灰階影像之喉內視鏡視訊影像。本計畫使用 APEER 來標記喉內視鏡視訊影像中的左右側聲帶以及聲門區域。

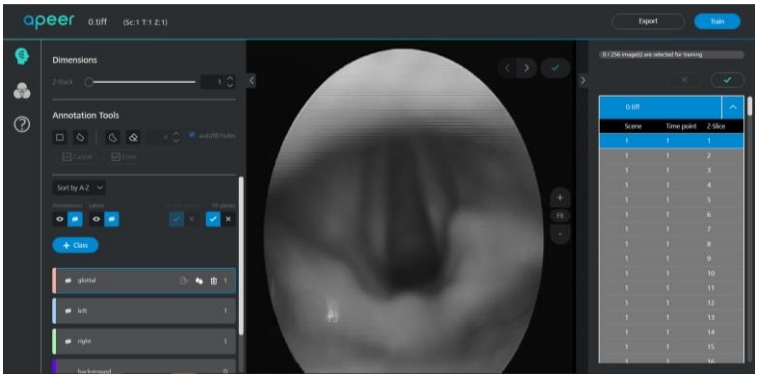
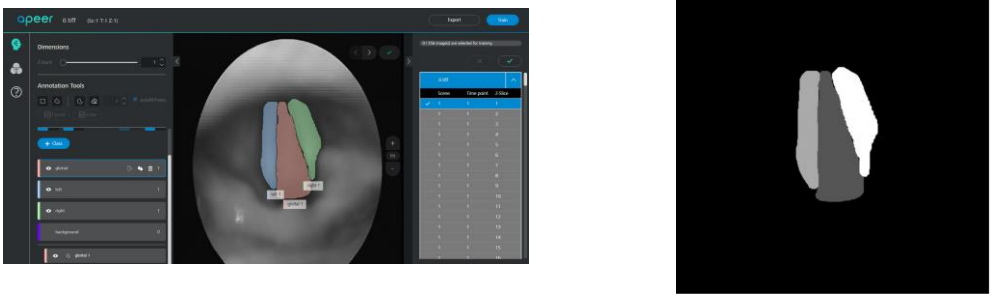


圖 4、使用 APEER 標記影像

在本計畫中，將喉內視鏡視訊影像中的物件分為左側聲帶、右側聲帶以及聲門共三種類別，如圖 5(a)所示。經過 APEER 標記後，會產生與原始影像大小相同的遮罩影像，其會以不同灰階值標示不同類別的物件，如圖 3.4(b)所示。



(a) (b)

圖 5、(a)為已使用 APEER 標記之影像，(b)為 APEER 產生之遮罩影像

3. 3D VOSNet 切割左右聲帶及聲門區域

由於透過喉內視鏡攝影儀所拍攝到的喉內視鏡視訊檔是由連續影像所組成，其完整包含聲帶過程的狀態。而為了將影像中的左聲帶區域、右聲帶區域以及聲門區域切割出來，且不因聲帶運動所造成的偏移，而產生嚴重的誤判。此外，在拍攝視訊檔時，容易因為醫師操作儀器的角度或聲帶運動導致影像曝光或模糊。因此，本計畫提出 3D VOSNet 模型，其包含 encoder 及 decoder 兩大部分，在 encoder 以及 decoder 的部分皆含有 1 個 7×7×7 的卷積層、4 個 Conv Block 以及 16 個 Identity Block，其中 Conv Block 為一個 Inception 結構，其中包含 65 個 1×1×1 的卷積層以及 32 個 3×3×3 的卷積層；而 Identity Block 則為 Inception 結構結合殘差區塊，其中包含 64 個 1×1×1 的卷積層以及 32 個 3×3×3 的卷積層，此模型可用於物件切割，且保留影像與影像之間的上下文資訊，而為了獲取多尺度的特徵，我們採用 ResNeXt 作為模型的 Backbone，其結合 Inception 結構以及 ResNet 的殘差塊，可在避免梯度消失的狀態下抽取多尺度的特徵。3D VOSNety 架構圖如圖 6 所示。

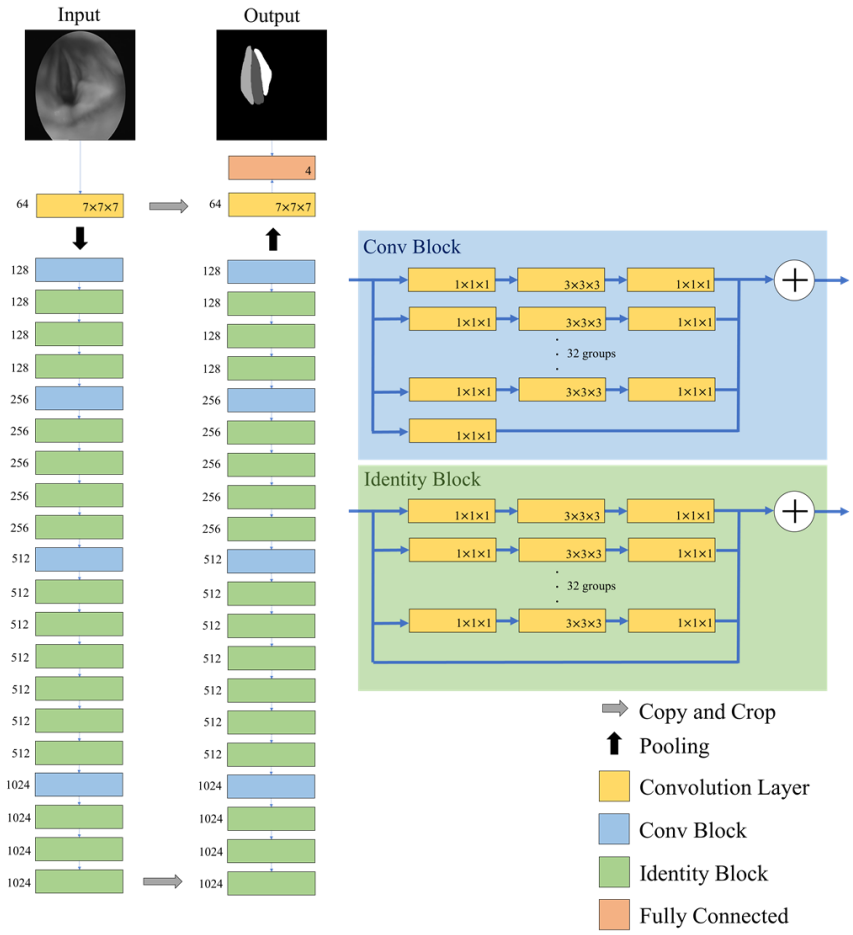


圖 6、3D VOSNet 架構圖

為了評估模型訓練過程中的成效，我們採用兩個損失函式，分別為 Dice Loss 和 Categorical Focal Loss。Dice Loss 主要是針對目標區域進行評估，也就是說未考量背景因素，喉內視鏡攝影儀拍攝聲帶時的位置以及距離可能不同，會使得聲帶在喉內視鏡視訊檔中的比例過小，因此本計畫採用 Dice Loss 作為損失函數，Dice Loss 藉由計算 Ground Truth 與 3D VOSNet 的預測結果中左聲帶區域、右聲帶區域以及聲門區域的相似度。但由於未考量背景因素，若模型在訓練過程中出現一部份預測錯誤，即會導致 Dice Loss 產生變動，進而使梯度劇烈變化。因此本計畫將 Dice Loss 結合 Categorical Focal Loss 作為損失函數，其會將喉內視鏡視訊影像中較複雜背景區域設定較大的權重，以此加強訓練該部分。相較於 Dice Loss，Categorical Focal Loss 不僅考慮了左聲帶區域、右聲帶區域以及聲門區域，還考慮了背景，因此可以使 Loss 的變動較為平衡。

4. 聲帶相關指標計算

為了有效輔助耳鼻喉科醫師在臨床上的診斷，本計畫將 3D VOSNet 切割後的聲帶經由自建演算法計算出聲帶的相關指標，以利後續進行病症或治療等分析。在本計畫中，我們根據聲帶對稱性以及聲帶振動模式提出了 6 種指標，分別為左右側聲帶曲率、聲帶長度偏差、聲帶面積偏差、聲門面積、聲門角度以及聲帶振動的對稱性。

4.1左右側聲帶曲率

為了得知聲帶是否有萎縮或彎曲的情形以及了解喉部的神經狀態，本計畫會找出切割後的聲帶區域的上頂點、下頂點和中心點，並取得此三點所構成的外接圓，而聲帶曲率即為此外接圓從上頂點連接至下頂點的弧長，除以外接圓半徑的值。若喉內視鏡視訊檔中的聲帶曲率被計算出皆趨近於 0，則代表該側聲帶有萎縮或麻痺之狀況；反之，則代表聲帶正常運動，聲帶曲率如公式(1)所示。

$$C_{VC} = \frac{Arc\ (P_u,\ P_c,\ P_d)}{r_{cir}} \tag{1}$$

其中，P_u為聲帶區域的上頂點，P_d為聲帶區域的下頂點，P_c為聲帶的中心點，Arc為P_u至 P_d的外接圓弧長，r_{cir}則為外接圓半徑。

4.2聲帶長度偏差

為了瞭解聲帶是否有麻痺或息肉等病症的狀況產生，可藉由計算聲帶長度的對稱性得知，因此本計畫會先計算出左聲帶與右聲帶差值，再分別將結果與左聲帶以及右聲帶長度的比值相加取絕對值，最後輸出平均聲帶長度偏差值，當此數值越接近 0，代表兩側聲帶對稱性越高，聲帶長度偏差如公式(2)所示。

$$D_{len} = \frac{\left|\frac{Arc_R - Arc_L}{Arc_R}\right| + \left|\frac{Arc_L - Arc_R}{Arc_L}\right|}{2} \tag{2}$$

其中，Arc_R為右側聲帶長度，Arc_L為左側聲帶長度。

4.3聲帶面積偏差

為了瞭解喉部病灶區域是否有擴大，如喉癌區域擴張，其可藉由計算聲帶面積的對稱性得知。因此本計畫會先計算出左右聲帶面積差值，再分別將結果與左聲帶面積以及右聲帶面積的比值相加取絕對值，最後輸出平均聲帶面積偏差值，當此數值越接近 0，代表兩側聲帶對稱性越高，聲帶面積偏差如公式(3)所示。

$$D_A = \frac{\left|\frac{A_L - A_R}{A_L}\right| + \left|\frac{A_R - A_L}{A_R}\right|}{2} \tag{3}$$

其中A_R與A_L分別為左側與右側聲帶的面積。

4.4聲門面積

為了瞭解聲帶是否有閉合不全的狀況，如聲帶麻痺或聲帶萎縮時，可藉由計算聲門區域的大小得知，而本計畫透過格林公式[19]計算模型切割出的聲門面積，若計算出聲門面積為 0 時，則代表聲帶有完全閉合；反之，則代表聲帶處於運動狀態，當喉內視鏡視訊檔中的聲帶面積皆未被計算為 0，則代表此病患有聲帶閉合不全的情況，聲門面積公式如公式(4)所示。

$$Area\_g = \frac{1}{2}\left|\sum_{i=1}^{n}(x_i y_{i+1} - x_{i+1} y_i)\right| \tag{4}$$

其中 n 代表構成聲門區域的頂點數，x 和 y 為每個點的坐標。

4.5聲門角度

為了觀測聲帶振動的運動狀況以及聲門是否有閉合不全的狀況，可藉由計算聲門角度來得知，本計畫將聲門角度定義為兩側聲帶的上頂點和聲門下頂點所形成的夾角大小，若喉內視鏡視訊檔中的聲門角度被計算出的最小值大於 0，則代表聲帶閉合不全；反之，則代表聲帶閉合完全，聲門角度如公式(5)所示。

$$Ang_g = \cos^{-1}\left(\frac{{O_l}^2 - {L_{left}}^2 - {L_{right}}^2}{-2 \times L_{left} \times L_{right}}\right) \tag{5}$$

其中O_l代表左右側聲帶底點的連線距離，L_{left}代表夾角與左側聲帶相鄰邊的長度，L_{right}代表夾角與右側聲帶相鄰邊的長度。

4.6聲帶震動的對稱性

為了瞭解聲帶振動的運動狀態以及對稱性，我們根據左右側聲帶中心點與聲門中線位置的距離進行評估，如公式(6) 所示，當喉內視鏡視訊檔中被計算出的聲帶振動對稱性數值大多趨近於 0，則代表聲帶可完全閉合；反之，則代表聲帶閉合不全。

$$Sym = \left\|P_{c\_left} - M_g\right\| - \left\|P_{c\_right} - M_g\right\| \tag{6}$$

P_{c_left}代表左側聲帶之中心點位置，P_{c_right}代表右側聲帶之中心點位置，M_g代表聲門中線位置

4.7輸出聲帶相關指標數據

最後，輸出左右側聲帶曲率、聲帶長度偏差、聲帶面積偏差、聲帶面積、聲帶角度、以及聲帶振動的對稱性等聲帶相關數據，並利用視覺化的方式呈現於系統介面上以供臨床醫療診斷上。

(五)實驗結果與探討

在實驗結果中，本計畫會先介紹系統的開發環境，接著會將實驗結果進行討論，將會介紹本計畫實作的流程，且以系統介面來做輔助說明。

1. 開發環境

本計畫在每次的實驗中都會在相同的硬體配置下進行訓練及測試，其電腦硬體配置的規格為 16GB RAM、i9-9900X CPU 和 NVIDIA Quadro RTX5000。本計畫在 Anaconda3 的虛擬環境中建立 Keras 框架，使用 Python 程式語言開發本計畫所使用的深度學習網路模型，系統的介面是使用 Python 內建的 GUI 函式庫 Tkinter 進行設計及開發。

表 1、軟硬體設備

軟體	作業系統	Linux Ubuntu 18.04
	深度學習開發環境	Anaconda 3
	開發程式語言	Python 3.8
	深度學習函式庫	Keras
	GUI	Tkinter
硬體	CPU 處理器	Intel Core i9-9900X
	GPU 顯卡	NVIDIA Quadro RTX 5000

2. 實驗結果討論

本計畫利用 3D VOSNet 模型切割喉內視鏡視訊檔內物件之結果圖如表 2 所示。從切割之結果圖可以觀察到，無論聲帶開合的狀態以及遮擋狀況，3D VOSNet 模型皆可有效切割出物件，但若因喉部組織皺褶而造成影像上呈現黑影，或因拍攝時造成影像有雜訊時，則容易有標記錯誤的狀況，如圖 7 中紅框處所示。

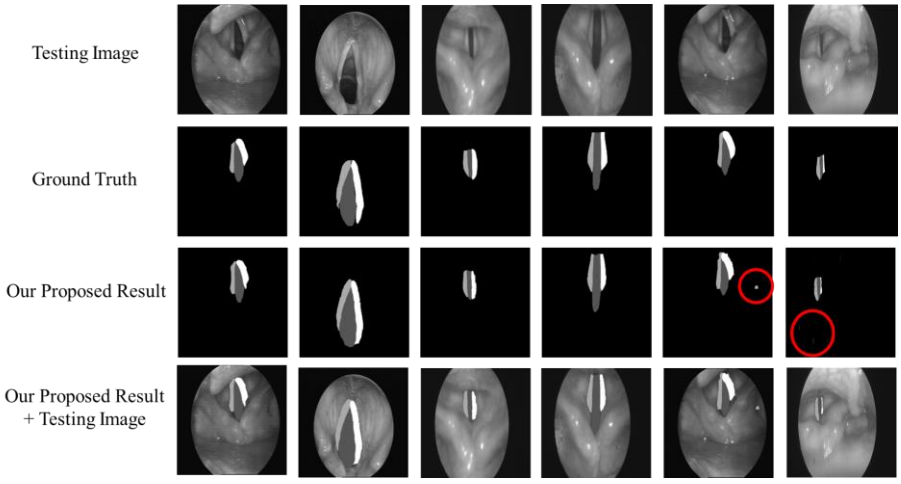


圖 7、3D VOSNet 切割結果圖

而為了驗證本計畫所提出的 3D VOSNet 之物件切割準確率，本計畫也將資料集中的 50 個喉內視鏡視訊檔視為測試資料，共擷取出 12800 張影像，並計算其混淆矩陣的相關指標以作為評估本計畫之指標，其中測試準確率達 92.67%，表 2 為本計畫在喉內視鏡視訊檔切割之混淆矩陣。

表 2 物件切割之混淆矩陣

Confusion Matrix of Testing		*Actual*			
		Background	*Left*	*Right*	*Glottal*
Prediction	*Background*	11880	626	537	930
	Left	470	11966	60	245
	Right	405	117	12112	166
	Glottal	110	192	145	11509

本計畫透過上表混淆矩陣中計算出相關指標來作為本實驗結果的評估指標，如圖 8 所示，3D VOSNet 模型切割出左聲帶、右聲帶及聲門的測試準確率分別可達到 93.48%、94.63%%以及 89.91%，得以證實 3D VOSNet 模型在多數情況下皆可準確切割，其中聲門的切割效果相較於左右聲帶略顯遜色，猜測可能原因為聲帶振動時，會將聲門遮擋，因此訓練資料較少所導致；在精確度的部分，左聲帶、右聲帶及聲門分別為 94.73%、94.09%以及 89.86%，則表示 3D VOSNet 模型可有效獲得聲帶物件的位置以及其對應的類別，其中 3D VOSNet 模型對於聲帶區域的表現不佳，主要是因為若聲帶完全閉合時，則不會有聲門區域，造成模型在切割時誤判聲帶位置；在敏感度的部分，左聲帶、右聲帶以及聲門分別為 93.85%、95.42%以及 90.86%，表示 3D VOSNet 模型對於聲帶物件區域有高靈敏度因此可正確將其進行標記；最後，特異度的部分，左聲帶、右聲帶以及聲門，分別為 92.99%、93.80%以及 88.89%，表示實際為非喉部物件區域，正確被標記的比例都約為 90%以上。由此可知，3D VOSNet 模型切割出喉部物件具有相當不錯的成效。

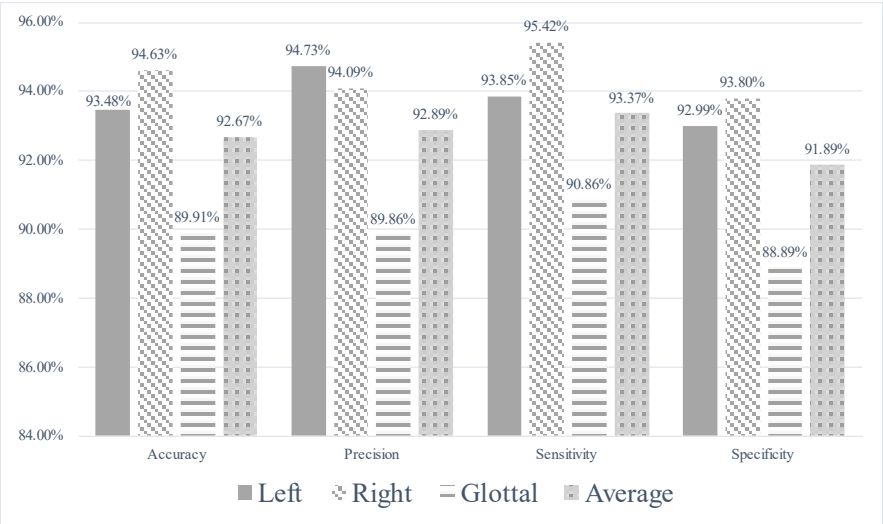


圖 8、混淆矩陣計算出的指標之比較長條圖

3. 系統實作介面

本計畫之系統採用簡潔明瞭的介面風格，期望有效輔助醫師診斷，以下會詳細介紹本系統的使用者介面以及使用流程。圖 9 為系統登入介面。為了確保病患隱私權不受侵犯，當醫師欲使用本系統進行喉內視鏡視訊檔的診斷及檢查時，醫師需先透過帳號密碼登入才可使用本系統。

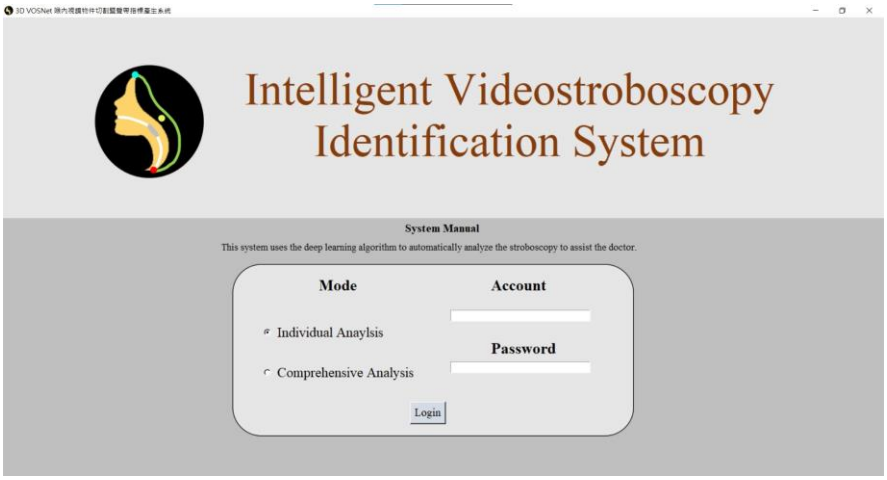


圖 9、系統登入介面圖

當醫師登入系統主介面，介面上會顯示該病患的基本資料供醫師參考，醫師可點選「Open File」選擇該病患過往拍攝的喉內視鏡視訊檔。接著點選「Analysis」按鈕，系統會自動將喉內視鏡視訊檔輸入至 3D VOSNet 中進行物件切割，將左聲帶區域、右聲帶區域以及聲門的區域切割出來，分析完成後，會將結果顯示於介面的右方區域，醫師可點選上一張或下一張按鈕，查看聲帶狀況。此外，介面上會顯示分析後的指標數據，提供給醫師作為診斷的依據。系統分析介面圖如圖 10 所示。

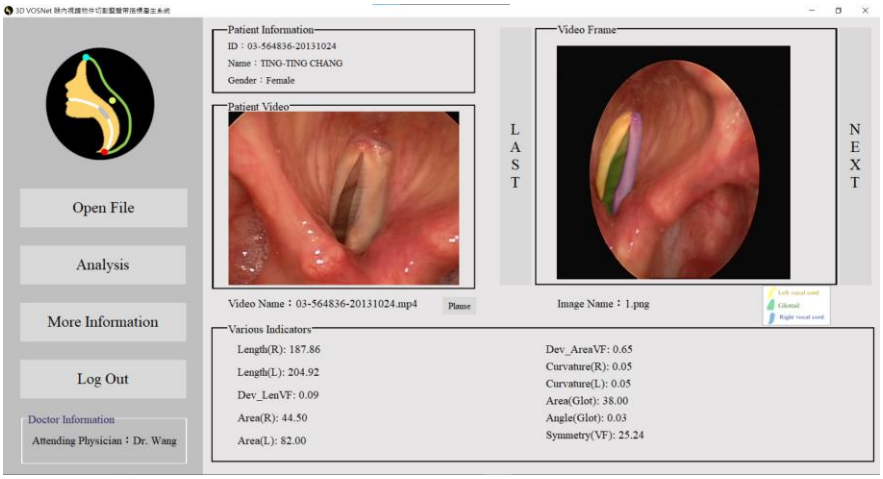


圖 10、系統主介面圖

最後，醫師可點選「More Information」按鈕，系統則會將計算出的各項指標顯示於介面上，此外，醫師還可點選左側按鈕，系統會根據不同指標數值大小進行排序，有效幫助醫師診斷病患，系統分析結果介面圖如圖 11 所示。

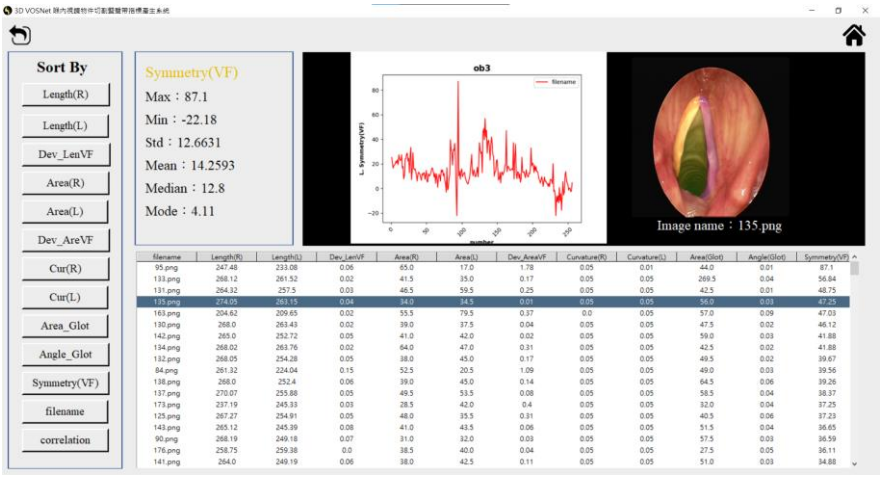


圖 11、系統分析結果介面圖

4. 討論與改進

除了使用 3D VOSNet 進行喉內視鏡視訊檔切割，本計畫亦比較了 3D UNet 以及可用於實例切割的 Mask R-CNN[20]模型在喉內視鏡視訊檔的切割效果，可由圖 12 得知，Mask R-CNN 於序列影像的切割準確率較低於 3D UNet 以及本計畫提出的方法，由於喉內視鏡視訊檔為序列影像且聲帶振動快速，而 3D UNet 以及本計畫提出的方法皆可保留前後文資訊且具有平移不變性，反觀 Mask R-CNN 模型不具前後文資訊的特性，物體快速移動則容易切割效果不佳。然而採用 3D UNet 其切割效果亞於本計畫提出的方法，推測可能原因為本計畫採用之 Backbone 架構為 ResNeXt，其可用於抽取出多尺度的特徵，因此對於快速震動而造成的聲帶邊緣模糊，可更有效切割出其邊緣。

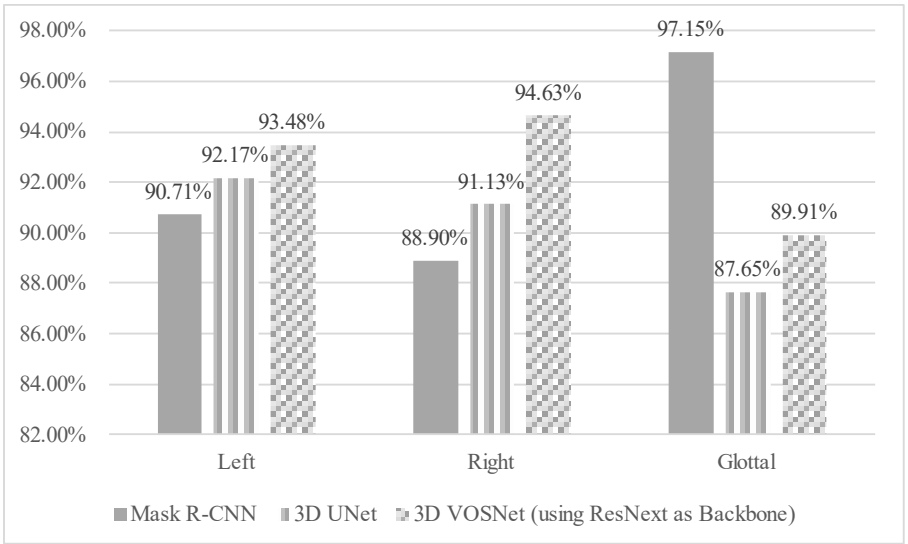


圖 12、各模型切割效果比較圖

此外，在本計畫中目前完成喉內視鏡聲帶物件切割以及各項指標的分析及排序，尚未針對病灶進行研究，未來希望將本計畫計算出的指標輸入至 1D 訊號分析的模型中，分析出可能患有疾病的機率，更有效的幫助醫師診斷病情。

(六)結論

目前在診斷喉部相關病症，耳鼻喉科醫師通常會藉由喉內視鏡攝影儀檢測病患聲帶的運動情形，然而由於人眼對於動態影像辨識率不佳，在診斷上需花費較長時間觀測才能進行病症的診斷，再加上有些病症還需搭配具有侵入性的喉肌電圖檢測，才可精準了解病患之病症，而此診斷方式對於病患而言接受度不高。且有鑑於目前國內耳鼻喉科門診尚未有關聲帶相關指標之醫學影像輔助診斷軟體，因此，本計畫提出「喉內視鏡影像物件切割且具評估聲帶病灶指標產生系統」，協助醫師於臨床醫療診斷，發揮實際應用價值。本計畫的準確率為 92.67%，可見本計畫可準確的切割出左聲帶區域、右聲帶區域以及聲門區域。期望透過本計畫自動化及智慧化的功能，有效提高臨床診斷效率與醫療品質。

（七）參考文獻

[1] L. N. Onwordi and C. A. Yaghchi. (2021). Airway Glottic Insufficiency. Treasure Island (FL): StatPearls Publishing.

[2] H. Hosono, C. Katada, T. Okamoto, M. Ichinoe, Y. Sakamoto, et. al., “Usefulness of narrow band imaging with magnifying endoscopy for the differential diagnosis of cancerous and noncancerous laryngeal lesions,” Head & Neck., vol. 41, pp. 2555–2560, 2019.

[3] S. J. Seyed Toutounchi, M. Eydi, S. E. Golzari, M. R. Ghaffari, N. Parvizian, “Vocal cord paralysis and its etiologies: a prospective study,” J Cardiovasc Thorac Res, vol. 6, pp. 47-50, 2014.

[4] R. T. Sataloff, P. Praneetvatakul, R. J. Heuer, M. J. Hawkshaw and Y. D. Heman-Ackah, “Laryngeal Electromyography: Clinical Application,” Journal of Voice, vol. 24, pp. 228-234.

[5] C. Matava, E. Pankiv, S. Raisbeck, and M. Caldeira, et al., “ A Convolutional Neural Network for Real Time Classification, Identification, and Labelling of Vocal Cord and Tracheal Using Laryngoscopy and Bronchoscopy Video,” Journal of Medical Systems, vol. 44, No. 2, 2020.

[6] J. Ren, X. Jing, J. Wang, X. Ren, Y. Xu, and Q. Yang, et al., “Automatic recognition of laryngoscopy images using a deep-learning technique,” Laryngoscope, vol. 28539, 2020.

[7] N. Xu, L. Yang, Y. Fan, and J. Yang, et. al., “YouTube-VOS: Sequence-to-Sequence Video Object Segmentation,” Proceedings of the European Conference on Computer Vision (ECCV), pp. 585-601, 2018.

[8] K. Duarte, Y. S. Rawat, and M. Shah, “CapsuleVOS: Semi-Supervised Video Object Segmentation Using Capsule Routing,” Computer Vision and Pattern Recognition, pp. 8479-8488, 2019.

[9] P.-Y. Kao, J. W. Chen, and B. S. Manjunath, “Improving 3D U-Net for Brain Tumor Segmentation by Utilizing Lesion Prior,” Computer Vision and Pattern Recognition, 2020.

[10] Z. Xiao, B. Liu, L. Geng, F. Zhang, and Y. Liu, “Segmentation of Lung Nodules Using Improved 3D-UNet Neural Network,” Symmetry, vol. 12, 2020.

[11] J. Yang, B. Wu, L. Li, P. Cao, and O. Zaiane, “MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT,” Computerized Medical Imaging and Graphics, vol. 92, 2021.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987-5995, 2017.

[13] T. Zhou, Y. Zhao and J. Wu, “ResNeXt and Res2Net Structures for Speaker Verification,” 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 301-307, 2021.

[14] K. Omori, D. H. Slavit, A. Kacker, and S. M. Blaugrund, “Quantitative videostroboscopic measurement of glottal gap and vocal function: An analysis of thyroplasty type I,” Annals of Otology, Rhinology & Laryngology, vol. 105, pp. 280-285, 1996.

[15] G. E. Woodson, “Configuration of the glottis in laryngeal paralysis. I: Clinical study,” Laryngoscope, vol. 103, pp. 1227-1234, 1993.

[16] R. R. Casiano, J. D. Cooper, D. S. Lundy, and J. R. Chandler, “Laser cordectomy for T1 glottic carcinoma: A 10-year experience and videostroboscopic findings,” Otolaryngology–Head and Neck Surgery, vol. 104, pp. 831, 1991.

[17] W. K. Cho, and S. H. Choi, “Comparison of convolutional neural network models for determination of vocal fold normality in laryngoscopy images,” Journal of Voice, 2020.

[18] Z. Zhang, “Estimation of vocal fold physiology from voice acoustics using machine learning,” Acoustical Society of America Journal, vol. 147, pp. EL264-EL270, 2020.

[19] L. Challis, and F. Sheard, “The Green of Green Functions,” Physics Today, vol. 56, pp. 41, 2003.

[20] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, Proc IEEE Comput Soc Conf Comput Vis, pp. 2961-2969, 2017.