國立勤益科技大學資訊工程系學生實務專題成果報告

※※※※※※※※※※※※※※※※※※※※※※※

※ ※

# ※ 情緒聲音辨識處理系統 ※

※ ※

※※※※※※※※※※※※※※※※※※※※※※※

執 行 期 間 ：112 學年度上學期至 113 學年度下學期

專題參與人員：

部別：日間部

組別：第 組  
 班別：四訊延 座號： 10 姓名：翁凱霖

班別：四訊延 座號： 28 姓名：曾義翔

指導老師： 楊勝智

中 華 民 國 113 年 4 月 14 日

目錄

[一、緒論 6](#_Toc194431930)

(一)、前言 6

(二)、動機與目的 6

[二、專題簡介 7](#_Toc194431931)

(一)、製作目的 7

(二)、方法 7

(三)、結果 7

1.特徵萃取

2.模型訓練

3.易操作的介面

4.流程圖

[三、製作理論探討 9](#_Toc194431932)

(一)、MFCC簡介 9

1.預強調

2.音框化

3.漢明窗

4.快速傅立葉轉換

5.三角濾波器

6.對數轉換公式

7.離散餘弦轉換

(二)、深度學習介紹 11

(三)、數據處理 13

(四)、神經網路建置 15

(五)、APP開發 17

[四、軟硬體分析 2](#_Toc194431934)0

(一)、環境開發 20

(二)、軟體 20

[五、測試結果](#_Toc194431935) 21

六、結論 23

[七、參考文獻 23](#_Toc194431936)

## 圖目錄

圖一、(流程圖) **8**

圖二、(MFCC示意圖) **9**

圖三、(預強調波型變化) **9**

圖四、(預強調原理) **9**

圖五、(音框化示意圖) 10

圖六、(漢明窗) 10

圖七、(三角濾波器) 11

圖八、(神經網路收斂流程) 12

圖九、fear聲音梅爾頻譜圖 13

圖十、Pleasant\_surprise聲音梅爾頻譜圖 13

圖十一、sad聲音梅爾頻譜圖 14

圖十二、happy聲音梅爾頻譜圖 14

圖十三、angry聲音梅爾頻譜圖 15

圖十四、神經網路架構圖 16

圖十五、js程式碼 17

圖十六、upload\_file程式碼 17

圖十七、前端示意圖 18

圖十八、judge程式碼 18

圖十九、輸出結果之模板程式碼 19

圖二十、最終顯示結果介面 19

圖二十一、Loss 隨著訓練次數變化 21

圖二十二、準確率隨著訓練次數變化 21

圖二十三、Loss 隨著訓練次數變化 23

圖二十四、準確率隨著訓練次數變化 23

圖二十五、五類測試資料預測混淆矩陣 23

# 情緒聲音辨識處理系統

Emotion Recognition System

組長:翁凱霖 組員:曾義翔 指導老師:楊勝智老師

Group Leader: Kai-Lin Wong Group Member: Yi-Siang Zeng

Advisor Professor: Sheng-Jhih Yang

國立勤益科技大學資訊工程系

Department of Computer Science and Information Engineering National Chin-Yi University of Technology

### 摘要

本研究旨在探討人的情緒聲音辨識處理系統，設計並實現了一個基於自然語言處理和機器學習的情緒聲音辨識系統，該系統能夠高效地將人員的語音輸入後經過模型辨識情緒類別，我們的系統集成了先進的語音辨識模型，並使用了多種微調技術，如改善學習率或是改變模型結構以優化其性能。系統功能包括以MFCC萃取聲音之特徵再進行模型訓練，最終可以達到99%以上之準確率辨識人的情緒類別。

### Abstract

This study aims to explore the emotional voice recognition processing system. A system based on natural language processing and machine learning for emotional voice recognition is designed and implemented. This system efficiently recognizes emotional categories from voice input using a model. Our system integrates advanced speech recognition models and employs various fine-tuning techniques, such as learning rate or changing structure of NN, to optimize its performance. The system's functionality includes extracting sound features using MFCC and then training the model, ultimately achieving over 99% accuracy in recognition of emotional.

## 一、緒論

#### (一)、前言

隨著現在人工智慧發展得如火如荼的背景下，我們的專題對其進行了深度研究，探討如何利用深度學習模型對語音進行分類，會先進行類神經網路建立一個模型後，再開發了一個包含客戶端和伺服器端的系統，可以讓用戶錄製一個聲音，透過我們在伺服器端運行事先訓練好的模型，進行分類。

在這個報告中，我們將詳細介紹我們的系統該如何運作，以及如何運用深度學習的技術來實現語音分類，我們希望透過這份報告能給人們有關人工智慧(AI)在情緒聲音的分類上， 期待未來能應用在客服上，輔助人員更精準判斷客戶之情緒。

#### (二)、製作動機與目的

情緒聲音辨識技術在客服領域的研究動機和目的主要包括以下幾方面：首先，通過快速分析客戶的語音，提高服務效率和準確性，特別是在面對大量使用時。其次，可以減少人力資源負擔。

更進一步提升使用者體驗，通過便捷的語音操作減少等待時間和不便。情緒聲音辨識技術還能增強情緒狀況監測，對於需要情緒支持的個體提供持續的情緒變化追蹤。特別是在特殊狀況時，無接觸的情緒篩查有助於減少直接接觸的風險。研究目的包括開發高效的情緒識別工具、提升遠程服務技術、改善使用者與服務提供者之間的互動、推動數據分析與管理、促進個性化服務以及支持心理健康和社會研究。

本研究的最終目標，是打造一套高效、準確的語音情緒辨識系統，推動智慧客服的應用發展，促進個人化服務與心理健康支持，進而提升整體社會的服務品質與生活福祉。

## 二、專題簡介

#### (一)、製作目的

我們希望能夠用更製能的方式，進行語音分類，協助客服人員能更即時偵測出客戶之情緒。

#### (二)、方法

使用MFCC萃取出聲音之特徵，再用類神經網路進行模型訓練後，最後使用自己的聲音進行模型驗證，再將此模型以flask製作一個APP，以達到即時錄音後即時預測。

#### (三)、結果

##### 特徵萃取

使用MFCC將聲音波轉化成一個特徵向量，後續進行CNN模型訓練

##### 模型訓練

我們使用keras的模組，進行類神經網路模型訓練，以2D CNN為主架構，優化內部神經元節點數或是增加神經層以達到最佳準確率，再使用自己錄製聲音進行預測驗證

##### 易操作的介面

我們為情緒聲音分類系統設計了一個直觀且易操作的介面，使用者無需擁有專業的知識， 也可以輕鬆簡單的使用我們的系統辨識聲音類別。

##### 流程圖

一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片

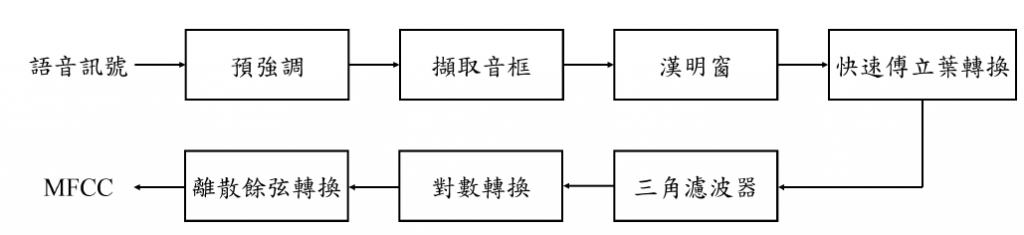
AI 產生的內容可能不正確。

圖一 (流程圖)

## 三、製作理論探討

#### (一)、MFCC簡介

MFCC全名為Mel-Frequency Cepstral Coefficients，以下表示其流程圖。



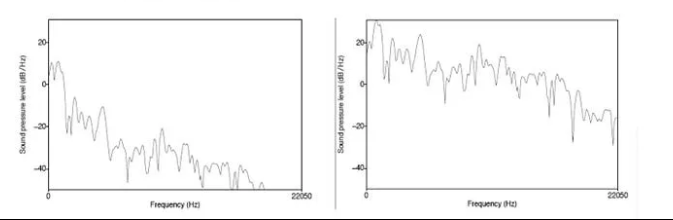
圖二 MFCC流程示意圖

1. 預強調(Pre-Emphasis

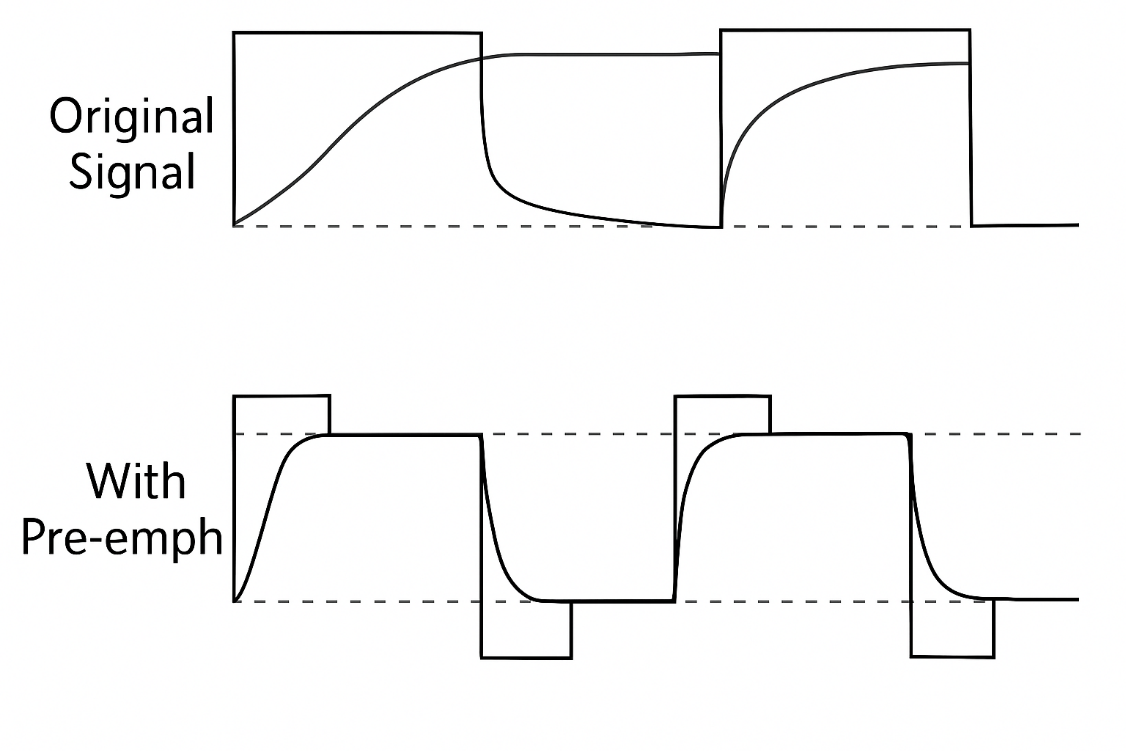
聲當音收集完成之後，訊號處理時頻率較高的部分會產生能量衰減的情形，因此在進行辨識或比對前，會將聲音訊號先經過一個高通濾波器，目的是為了提高高頻的訊號，避免後續進行轉換時發生問題，如公式(2)：

y(𝑡) = 𝑥(𝑡) – 𝛼-𝑥(𝑡 − 1) , 0.9 ≤ 𝛼 ≤ 1.0 (2)

其中y(t): 預加重後的信號，𝑥(𝑡)為原先的波型訊號α𝑥(𝑡−1)為經過高通濾波器訊號，通常α介於0.9到1之間，此方步驟主要是將聲音訊號的高頻進行放大，放大後的聲音訊號變得較尖銳清脆，但音量會降低。其變化可以參照圖2. 5。



圖三 預強調波形變化



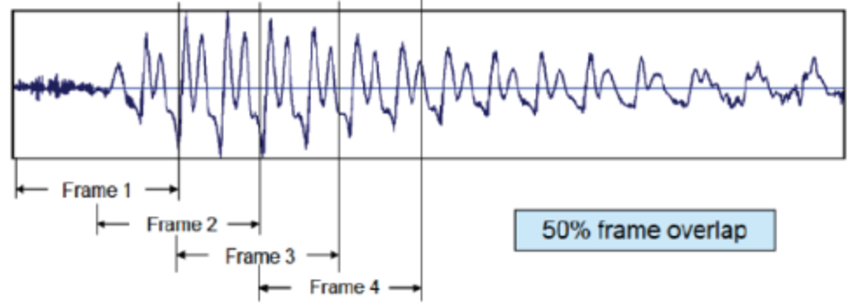
圖四 預強調原理

1. 音框化(Frame Blocking)

連續的完整聲音訊號資料因為時間過長導致快速變化，因此在分析訊號時，通常會將聲音訊號以 N 個取樣點為切割數量，將連續訊號切割分塊，我們將這些區塊稱為音框(Frame)，標準音框大小為 25ms，舉例說明，取樣頻率 44.1kHz 的聲音資料，因此取樣點為 1103 個，而在各音框間有一段重疊區域，目的在於增加各音框間的連續性，為 1/2 到 2/3 不等。其算式如下：

x\_m[n] = x[n + m \* hop\_size]

x[n]為原始信號，x\_m[n]表示第m幀



圖五 音框化示意圖

1. 漢明窗(Hamming Window)

將分割的音框乘上一個漢明窗，目的在於增加各音框之間的連續性，避免後續傅立葉轉換時產生訊號不連續而產生雜訊，其公式如下：

W[n]=α-β×cos(2πn/(N-1))n=0~N

其中N是窗的長度α=0.54 和 𝛽=0.46 是窗的常數

一張含有 行, 繪圖, 圖表, 斜率、斜坡 的圖片

自動產生的描述

圖六 漢明窗

1. 快速傅立葉轉換

由於在傳統的傅立葉轉換時，我們必須將連續的聲音資料進行積分，因此處理上非常繁瑣，而本文使用的是快速傅立葉轉換(Fast Fourier Transform, FFT)，頻域相較於時域能夠依據能量的分佈來辨識不同聲音的特性。

X[k]=∑〖w[n]×y[n]×e^(-j2πkn/N) 〗

N 是信號框的長度 k 是頻率索引，表示從 0 到𝑁−1

N−1 的頻率成分

1. 三角濾波器

由於各個音框為短時間訊號，因此處理上較快速且簡單，提高轉換的效率，接著將轉換完的頻譜能量乘上一組三角濾波器，並求得每一個濾波器所輸出的對數能量(Log Energy)，其中這三角帶通濾波器，必須依據「梅爾頻率」（Mel Frequency），就有如人的聽力系統，模擬人耳在對於不同的頻率資料時有著不同的敏感度，將各頻率間的資料進行過濾，只提取所需的特徵，各組濾波器並沒有平均分布的在符合梅爾頻率的三角濾波器的數量 M 通常設定為 20~40，而濾波器之間的間隔也隨著頻率的提高而減少變成稀疏，如下圖所示

一張含有 文字, 螢幕擷取畫面, 行, Rectangle 的圖片

自動產生的描述

圖七 三角濾波器

其中橫軸為頻率，縱軸為頻率響應，在各個濾波器中，頻率響應都為 1，梅爾頻率是依據人耳對於頻率的感受程度而計算的。依據人耳的特性，當聲音頻率越低時，三角濾波器越密集，則能保留人耳可分辨的聲音資料；反之當聲音頻率越高時，三角濾波器越稀疏。在語音識別系統常使用 MFCCs 做為特徵提取的方法，會去考量到音調的變化，而在本研究並不會受到輸入語音的音調不同而有所影響，其最主要目的，是根據不同的三角濾波器的個數所產生的頻譜圖之間的差異，做為聲音分類的特徵，並探討不同三角濾波器數量之間的差異。

1. 對數轉換其公式

S'[m] = ln(|S[m]|)

S[m] 是經過梅爾濾波器組處理後的能量譜

S'[m] 是對數轉換後的結果

ln() 是自然對數函數

1. 離散餘弦轉換(Discrete cosine transform, DCT)

經常用於訊號處理和圖像處理，對訊號和圖像進行壓縮，實際就是對每幀數據在進行一次降維，由三角濾波器的梅爾頻譜加權濾波器相乘而獲得，而m是第m個三角形濾波器係數，最後得到我們整個流程所要提取的梅爾倒頻譜係數。公式如下

C[n]=α(n)\*∑▒〖S^' [m]\*cos⁡(πn(2m+1)/2M)〗

C[n] 是第n個MFCC係數

S'[m] 是對數梅爾頻譜

M 是濾波器組的數量

m代表梅爾濾波器的索引

n 是MFCC係數的索引

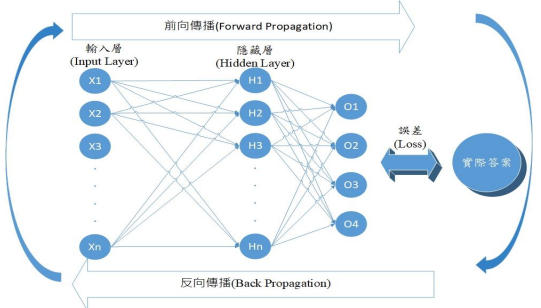
α(n) 是歸一化因子

綜合以上簡言之，梅爾頻率倒譜係數是一種模擬人耳對不同頻率敏感度的頻率尺度。在這種尺度下，低頻部分的變化較為敏感，而高頻部分的變化較為不敏感。因而適合用於語音辨識。語音訊號經過預強調 (Pre-emphasis)，目的是加強語音訊號中的高頻成分，因為語音信號的高頻部分通常較弱。這樣可以平衡語音中不同頻率範圍的能量，使後續的處理過程能夠更好地捕捉到語音的特徵，以此來突顯高頻的部份，再將訊號轉變為音框(Frame)，並對每一個音框乘上一漢明窗(Hamming Window) 來增加音框的連續性，接下來經由快速傅立葉轉換(Fast Fourier Transform,FFT) 將訊號從時域轉換到頻域上，再將得到的能量頻譜乘上M個三角帶通濾波器(Triangular Bandpass Filters)，獲得每一個濾波器輸出的對數能量(Log energy)，將上述的M個對數能量透過離散餘弦轉換(Discrete Cosine Transform, DCT) 最終算得梅爾頻率倒譜係數。

#### (二)、深度學習介紹

深度學習為本研究會使用得核心技術，這項技術在2006年，因為深度學習之父 Geoffrey Hinton 提出深度信念網路(Deep Belief Networks, DBN)，使用非監督的提前訓練方法來優化網路權重的初始值，接著進行權值的些微調整(Fine-Tune)後讓多層神經網路能夠真正的被實現，而當時的深度神經網路則被稱之為深度學習。在 2012 年Hinton 學生使用圖形處理器(Graphics Processing Unit，GPU)運算深度學習模型，使其運算效率高於原本中央處理器(Central Processing Unit，CPU)運算的 70 倍以上。此後深度學習技術應用於各大領域，包含金融股市預測、醫療保健用品、氣象預測、客戶分析、醫療影像辨識和智慧交通等產生十分深遠且巨大的影響。其最具代表性的深度學習為AlphaGo，AlphaGo 在圍棋領域擊敗各路的高手。現在深度學習架構應用的領域十分廣泛普及，例如音訊識別、語音識別、影像識別、自然語音處理、生物資訊學與生活眾多相關領域。

深度學習是模擬人類神經網路的運作方式。由於人腦結構非常複雜，預估具有 860 億個神經元及超過 100 兆條的神經相連，形成的網路比最先進的超級電腦還要強大。但為方便以電腦模擬，將神經元分為多層次，來模擬神經網路模型。在人工神經網路(Artificial Neural Network, ANN)實現中，每層裡頭各個神經元都有所謂的節點，而節點都相互的連結，神經元和節點通常具有隨著學習進行而調整的權重，權重增加或減小做為連接處訊號的強度。神經元可以具有閾值，使得僅當總訊號超過該閾值時才發送訊號。通常，神經元聚集成各層，不同的層可以對它們的輸入執行不同的變換。神經網路模型通常會有 1 個輸入層、1 個輸出層、隱藏層(可多層)，如下圖所示。

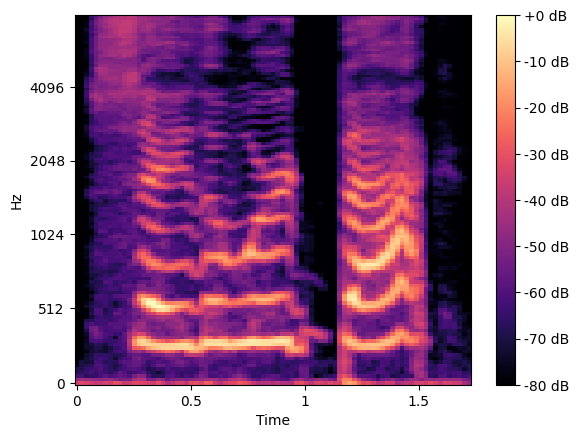


圖八 神經網路收斂流程

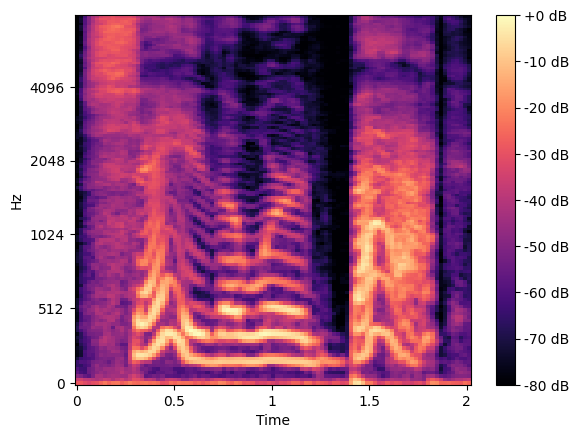
#### (三)、數據處理

##### 情緒聲音可視化

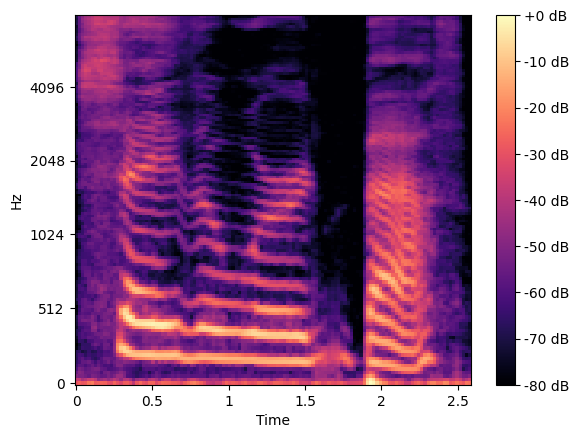
利用librosa套件，將時間域的訊號轉成頻率域，資料來源為https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess以下將各個類別聲音繪製成圖



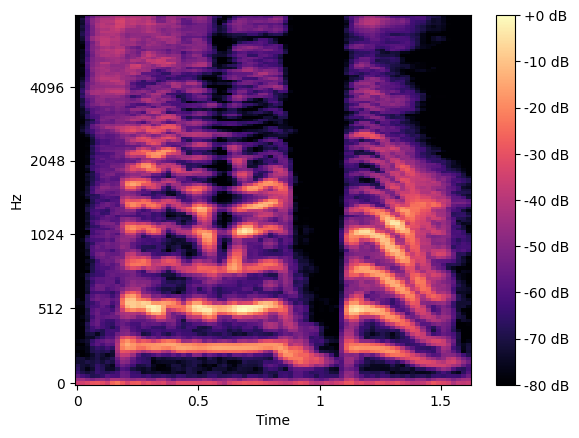
圖九 fear 聲音梅爾頻譜圖



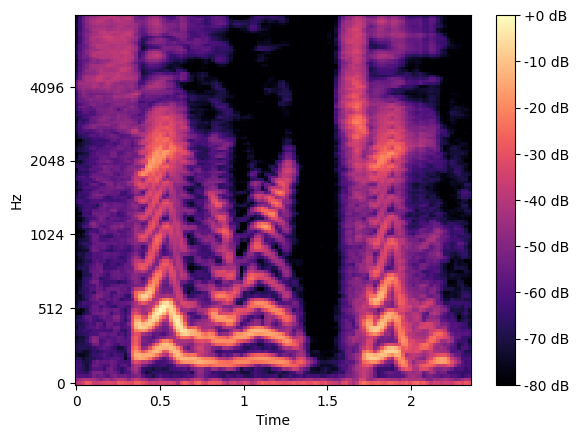
圖十 Pleasant\_surprise聲音梅爾頻譜圖



圖十一 sad聲音梅爾頻譜圖



圖十二 happy聲音梅爾頻譜圖



圖十三 angry聲音梅爾頻譜圖

由上述圖表可以看出不同類別存在一定差異，使用神經網路來訓練模型進行分類，具備一定可行性。

#### (四)、神經網路建立

1. 首先，我們使用的神經網路由卷積層(Convolutional Layers)、批次正規化(Batch Normalization)、激發函數(ReLU)、最大池化層(Max Pooling)、捨棄(Dropout)、全連接層(Fully Connected Layer)以及 Softmax 輸出層組成。
2. 輸入層(Input Layer)

輸入形狀為 (wave1\_c.shape[1], wave1\_c.shape[2], 1)，其中wave1\_c是將聲音產出的MFCC向量，形成一個2\*2的矩陣作為輸入的形狀。

1. 卷積層與池化層（Convolution + Pooling）

我們將測試使用一組或兩組 2D 卷積層 (Convolution2D)，每個卷積層使用 4 個 2×2 的過濾器（filters），padding="same" 用於確保輸出維度與輸入維度相同。

1. 批次正規化（BatchNormalization）

通過標準化每一層的輸入，使得訓練過程更加穩定，並且減少了對學習率和權重初始化的敏感性。

1. 激發函數(ReLU)

將負值設為 0，提升非線性特徵學習能力。

1. 最大池化層(MaxPooling2D)

降低特徵圖大小，減少計算量，並且讓模型處於過擬合和欠擬合之間，使模型能夠在訓練資料之外的測試集上進行更佳準確的預測。

1. Dropout(0.2)：為了防止過擬合，隨機丟棄 20% 的神經元。
2. 全連接層(Fully Connected Layer)

Flatten()：將多維輸出攤平成一維向量，以供全連接層處理。Dense(64\*16)：隱藏層，具有 1024 個神經元。經過批次正規化及 ReLU 激發，最終得到輸出層（Output Layer）

1. Adam 優化器 (Adam(0.001))

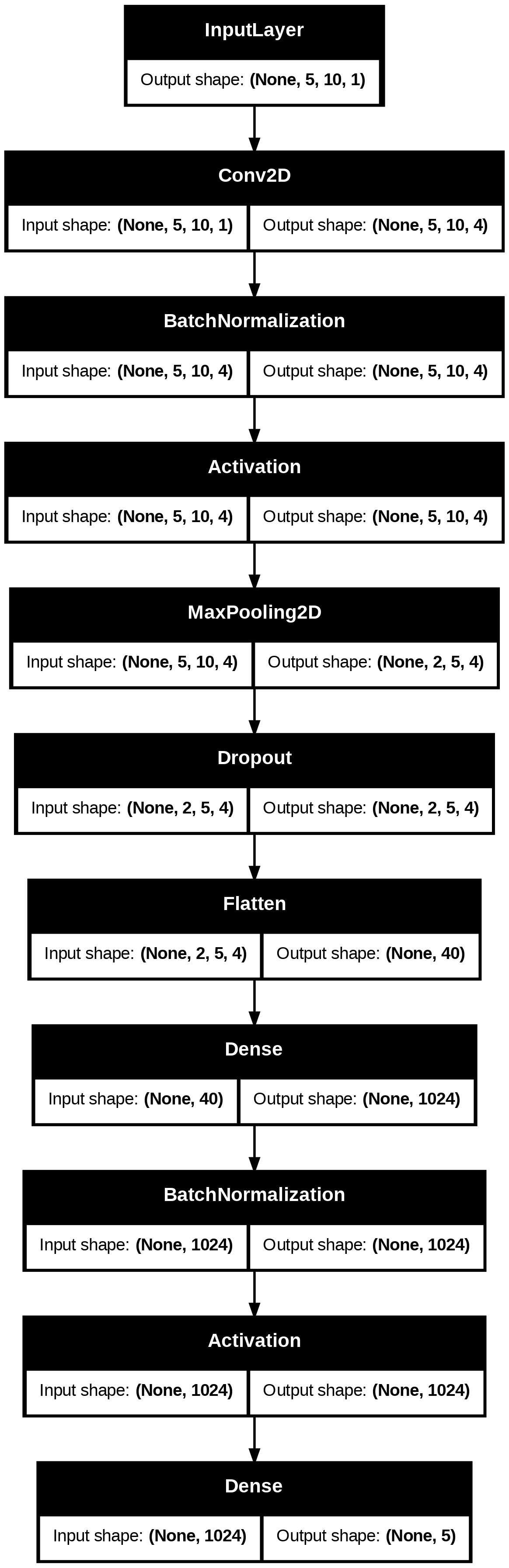
學習率 0.001，是適應性學習率的優化算法，有助於加快收斂。

1. 損失函數 (categorical\_crossentropy)

適用於多分類問題（One-hot 編碼的標籤）。

1. 評估指標

使用準確率(accuracy)作為評估標準。



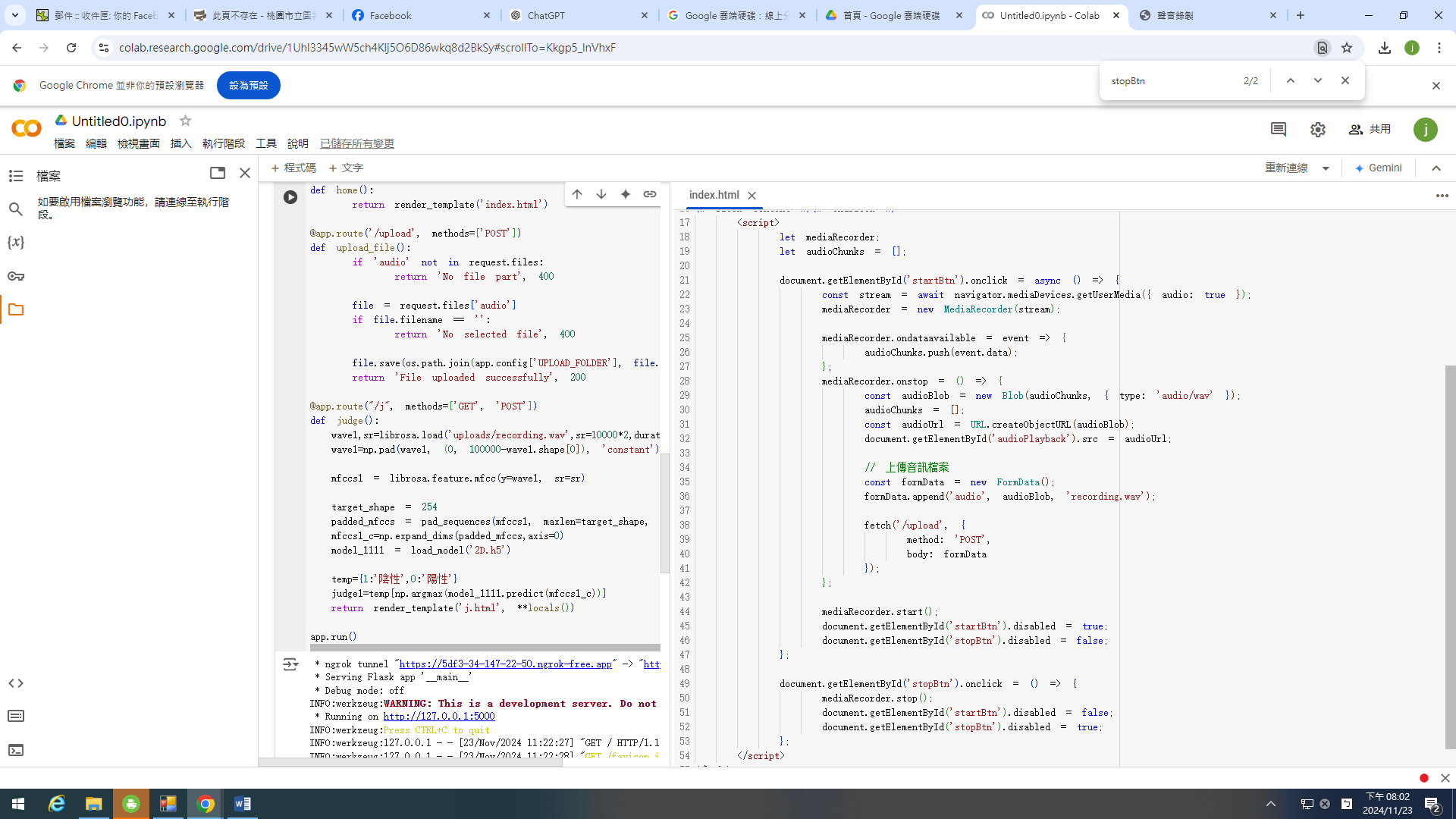
圖十四 神經網路架構圖

#### (五)、APP開發

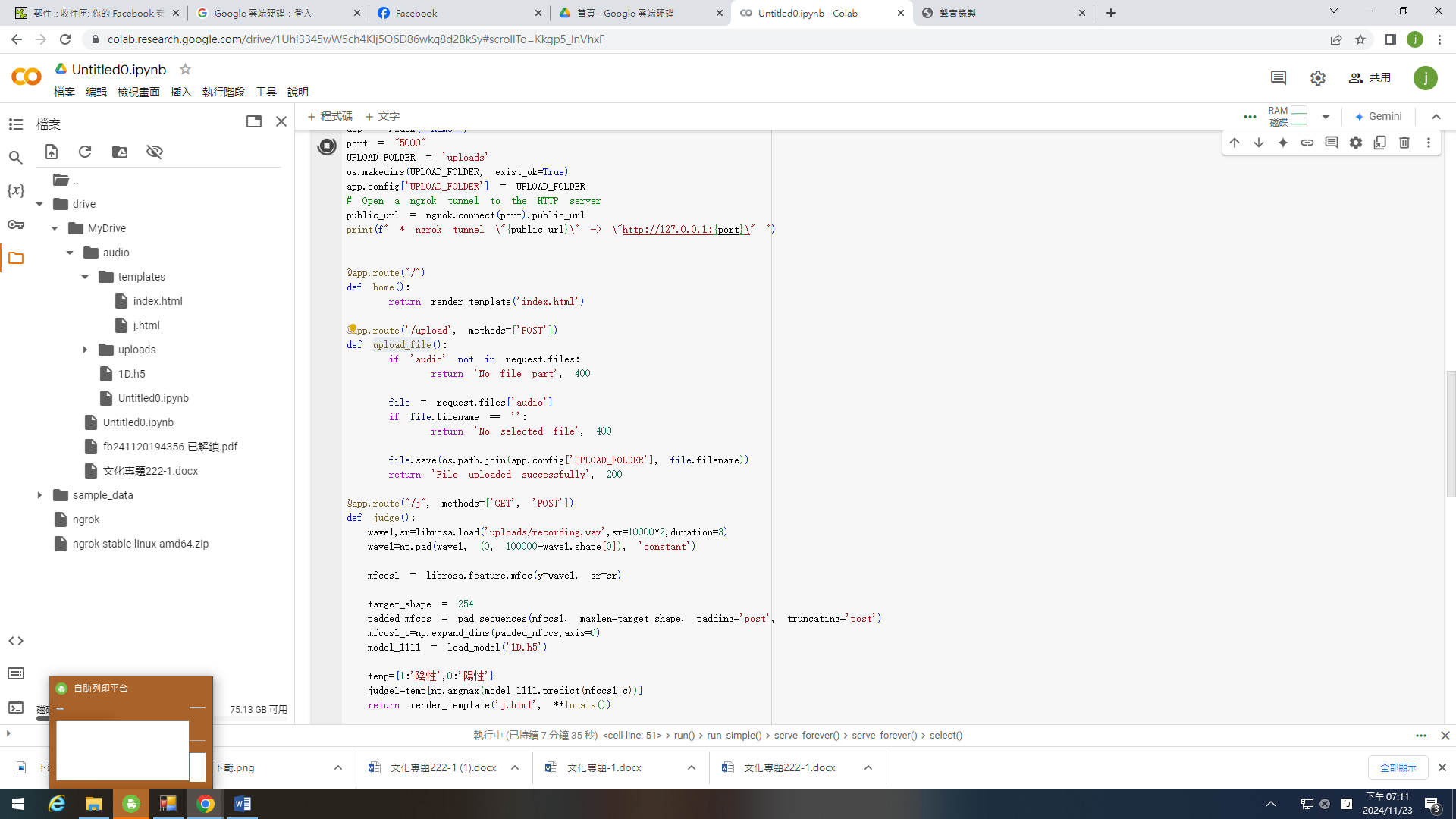
這裡的處理就好像是我們準備利用flask作為後端及html作為前端製作一個網頁應用程式，網站主要功能敘述如下：

1. 使用者錄製聲音，並存檔至後台。
2. 按下送出後，由後台進行數據處理，由模型計算出分類。
3. 由後台將資料拋到前台給使用者。

前端主要有三個按鈕，按下「開始錄製」則會觸發js內navigator.mediaDevices.getUserMedia這個Web API，它允許瀏覽器訪問用戶使用媒體設備（例如，攝像頭、麥克風），當使用者錄完聲音後按下「停止錄製」，此時會觸發js內onstop 事件，這時會利用fetch API來向後端伺服器發送一個 POST 請求，相關js程式碼如圖十四，之後後台的upload\_file這個函式如，此處POST是用來向服務器發送數據的請求方法，會偵測是否有一個audio的資料存在，若有就會將前端中的這個聲音檔案儲存到後台資料夾，另外右側一個播放介面可以即時放出剛剛錄製聲音如圖十六

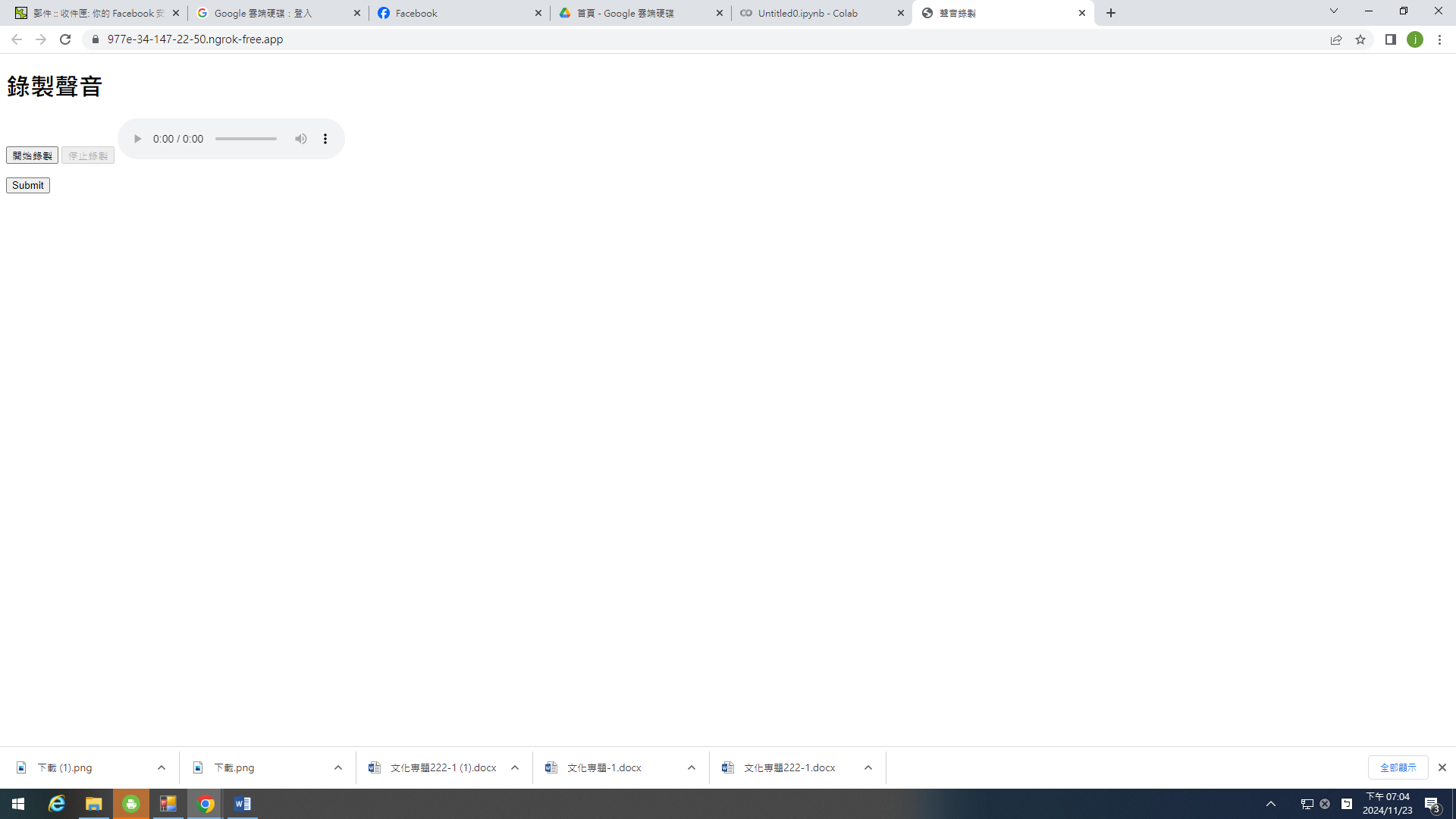


圖十五 js程式碼

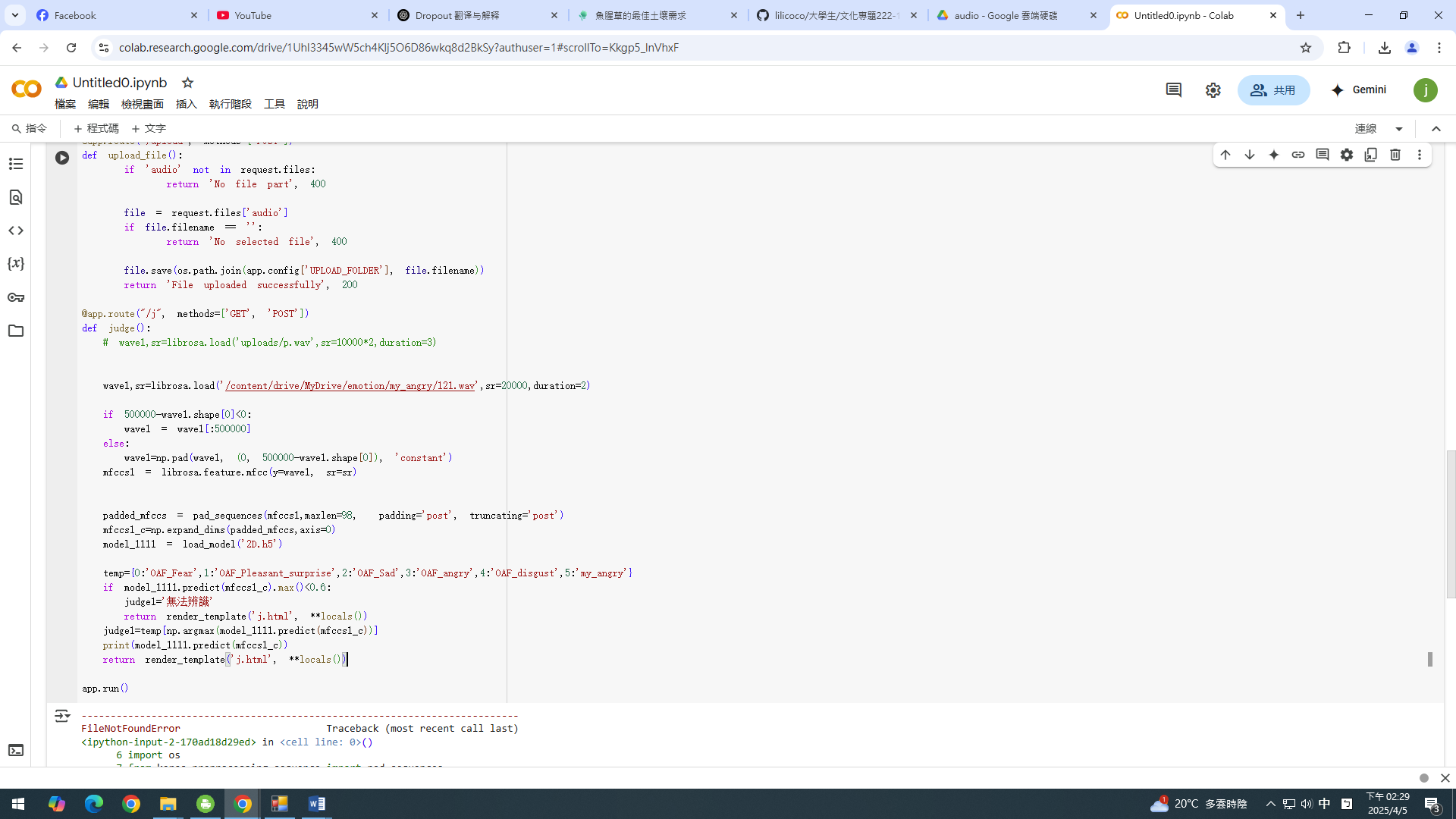


圖十六 upload\_file之程式碼

再來使用者需要按submit，在這個form tag內設定了action="{{ url\_for('judge') }}"，因此會呼叫後台名為judge的函式，此時就會依照之前數據前處理的模式將這個聲音從資料夾讀取並轉換成一個特徵向量，由模型進行預測類別。

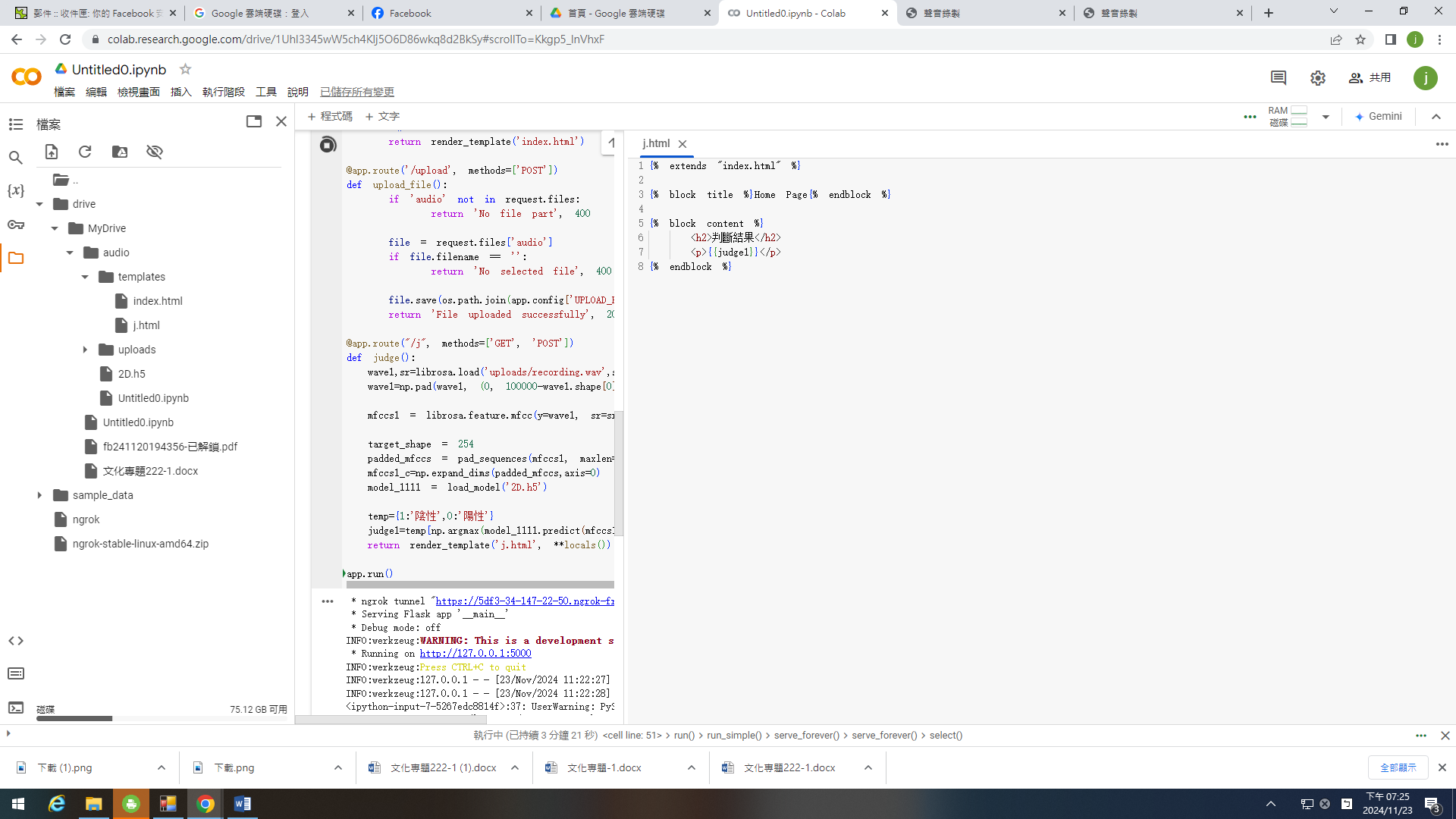


圖十七 前端示意圖

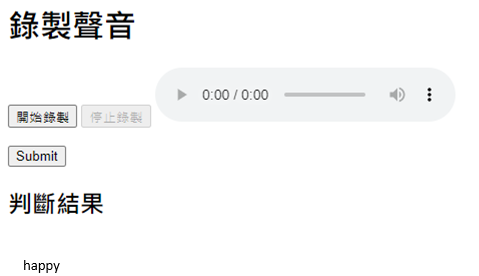


圖十八 judge之程式碼

由圖十七將預測好的類別存在變數judge1內部，然後會再發送出去，接收這個變數的模板為「j.html」，其程式碼參照圖十八，此處用到Jinja2模板語言，在Flask網頁框架的 HTML 模板設計非常實用，可以幫助開發者共用一些內容，{% extends "index.html" %}即是繼承上一層請使用者錄音的模板，最大優點，其一就是不需要再將index.html在j.html內再重複打一次，未來若是index.html有做更動，由於j.html是繼承也會跟著一起變動，而{{judge1}}這邊就負責接收後台傳過來的參數，此處會顯示使用者聲音判定為陰性或陽性，最終的介面就會如圖4. 20所示。



圖十九 輸出結果之模板程式碼



圖二十 最終顯示結果介面

此應用程式因為使用ngrok 開放本地端口，使用的套件為pyngrok，用來開啟一個ngrok隧道，將本地端口(此處預設5000)映射到公網。其他人可以通過一個公開的 URL 訪問本地 Flask 應用。也可以讓使用者透過手機直接操作這個應用程式。

#### 四、軟硬體分析

**(一)、開發環境**

Google colab

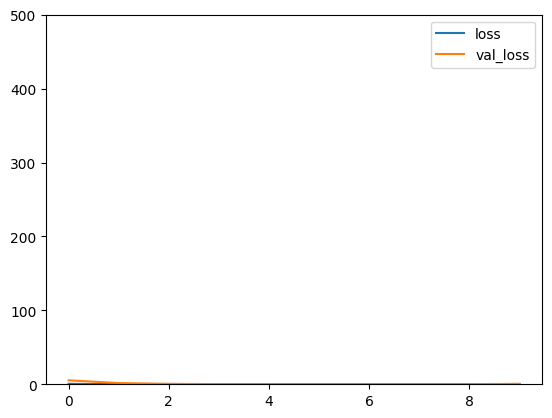
#### (二)、軟體

Python 3 Google Compute Engine 後端

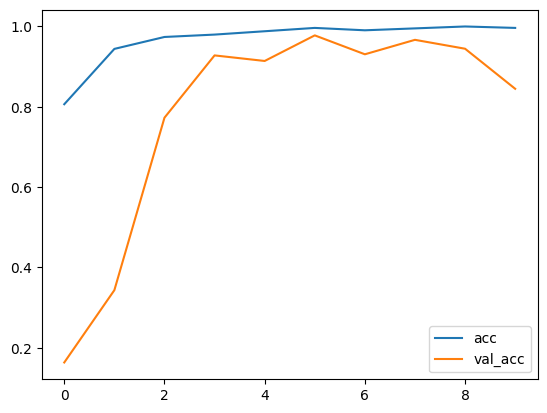
## 五、測試結果

#### (一)、訓練成果比較

使用兩種神經網路架構，第一種為一層另一個為二層，一層之訓練成效如下圖

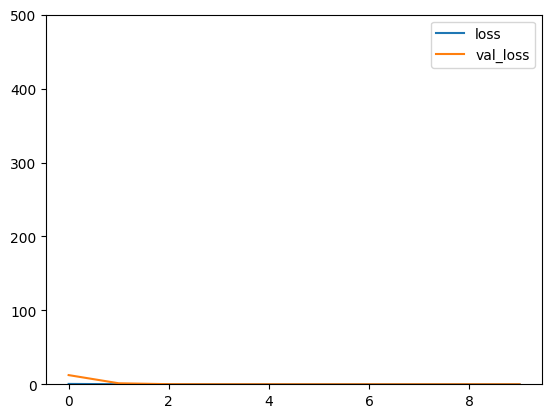


圖二十一 Loss 隨著訓練次數變化

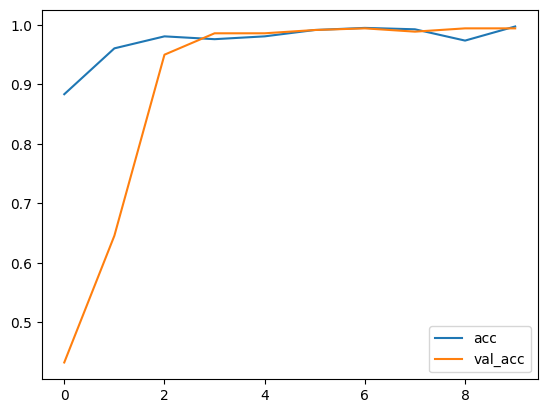


圖二十二 準確率隨著訓練次數變化

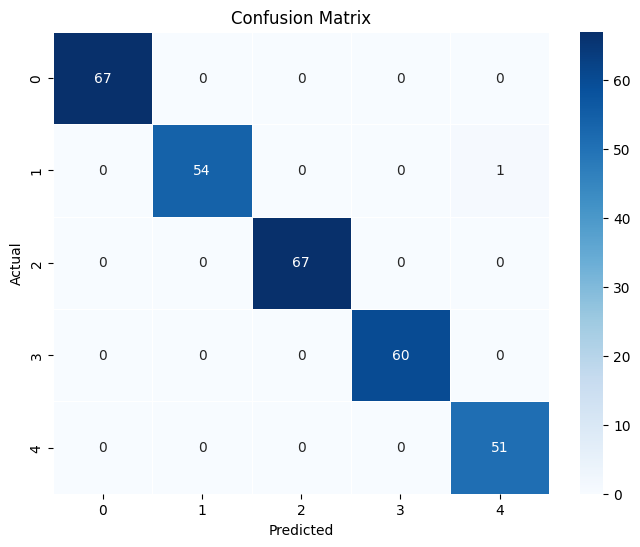
其中可以看到沒有過擬合狀況發生，但是驗證資料集之準確率並沒有一直提升，而二層之訓練成效如下圖



圖二十三 Loss 隨著訓練次數變化



圖二十四 準確率隨著訓練次數變化



圖二十五 五類測試資料預測混淆矩陣

依然沒有過擬合狀況發生，但是驗證資料集與一層神經網路比較可以越來越高，這樣的話對於未來未知的資料準確率會較好，以未知資料去預測其準確率達到90%以上，因此使用二層神經網路輸出成一個h5檔案，放置於APP後台以提供預測。

## 六、結論

二層建構的神經網路能更好預測訓練的資料以及驗證集資料，在製作專題時有發現使用js套件錄製的聲音，無法在windows環境由librosa讀取，因此轉去使用colab環境，因此才由一開始想在自己筆電建置APP改去使用colab環境啟動APP的服務。

## 七、參考文獻

[1] 劉佳鑫。「基於深層類神經網路之 聲音事件偵測系統」。碩士論文，國立臺北科技大學電子工程系研究所。

[2] 蔡佩樺。「機器學習聲音辨識應用於車禍偵測之研究」。碩士論文，逢甲大學資訊電機工程碩士在職學位學程

[3] 張永霖。「基於卷積神經網路的聲音分類機制」。碩士論文，朝陽科技大學資訊與通訊系，2021。

[4] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for FCC feature extraction," The 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, Gold Coast, QLD, Australia, 13-15 Dec. 2010.

[5] 林蓬榮, “規劃網站品牌的競爭策略,” 管理雜誌, vol. 326, pp. 127-129, 民90.

[6] M. Neely, “Optimal pricing in a free market wireless network,” Wireless Networks (10220038), vol. 15, no. 7, pp. 901-915, 2009.

[7] 吳克振, “非優質品牌應該如何應變,” 設計, vol. 99, pp. 110, 民90，6月.

[8] T. Wasserman, “Zsystems' Linux System Cracks $199 Price Tag,” Brandweek, vol. 41, pp. 20, 2000.

[9] T. Wasserman, “Zsystems' Linux System Cracks $199 Price Tag,” Brandweek, vol. 41, no. 18, pp. 20, 2000.

[10] 李杰修。「以聲音基頻與SVM分類器為基礎之人聲分類方法」。碩士論文，中央警察大學鑑識科學研究所，2018。