**中 國 文 化 大 學**

**資 訊 工 程 學 系**

**資 訊 系 統 專 題**

**醫療聲音辨識處理系統**

**學 生：方建智**

**康宇翔**

**曾敬峰**

**指導教授：劉仲鑫**

**中 華 民 國 113年5月**

**醫療聲音辨識處理系統**

**專題學生：方建智、康宇翔、曾敬峰**

**指導教授：劉仲鑫 博士**

**中國文化大學 資訊工程學系**

# 摘要

本研究旨在探討醫療聲音辨識處理系統在臨床應用中的潛力與實用性。本論文設計並實現了一個基於自然語言處理和機器學習的醫療聲音辨識系統，該系統能夠高效地將醫療專業人員的語音輸入轉換為文字記錄，從而提升醫療文書處理的效率和準確性。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LoRA和Prefixing，以優化其性能。系統功能包括語音轉錄、醫療術語識別、語境理解以及自動生成醫療記錄，這些功能旨在減輕醫療工作者的文書負擔，並提高患者資料的準確性和可讀性。

**關鍵詞：**AI 醫療聲音辨識、語音轉錄、MFCC

指導教授\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(簽名)

**Medical voice recognition processing system**

**Student: Yu-Xaing Ｋang, Jian-Zhi Fang, Jing-Fong Zeng**

**Advisor: prof.Chung-Hsin Liu**

**Department of Computer Science and Information Engineering**

**Chinese Culture University**

# Abstract

Medical voice recognition processing systemThis study aims to investigate the potential and practicality of medical voice recognition processing systems in clinical applications. The paper designs and implements a medical voice recognition system based on natural language processing and machine learning. This system efficiently converts verbal input from medical professionals into textual records, thereby enhancing the efficiency and accuracy of medical document processing. Our system integrates advanced speech recognition models and utilizes various fine-tuning techniques such as LoRA and Prefixing to optimize its performance. System functionalities include speech transcription, recognition of medical terminology, contextual understanding, and automatic generation of medical records. These functionalities are intended to alleviate the documentation burden on healthcare workers and improve the accuracy and readability of patient data.

**Keywords:** AI Medical Voice Recognition, Speech Transcription, MFCC

目 錄

[摘要 I](#_Toc183025895)

[Abstract II](#_Toc183025896)

[表 目 錄 IV](#_Toc183025897)

[圖 目 錄 V](#_Toc183025898)

[第 1 章 研究動機與目的 1](#_Toc183025899)

[1.1 研究動機 1](#_Toc183025900)

[1.2 研究目的 1](#_Toc183025901)

[第 2 章 文獻探討 3](#_Toc183025902)

[2.1 人工智慧在聲音辨識之應用 3](#_Toc183025903)

[2.2 聲音特徵提取 6](#_Toc183025904)

[2.3 機器學習(Deep Learning) 11](#_Toc183025905)

[2.4深度學習(Deep Learning) 12](#_Toc183025906)

[第 3 章 研究內容、方法與工作項目 18](#_Toc183025907)

[3.1 研究方法設計 18](#_Toc183025908)

[3.2 使用工具 19](#_Toc183025909)

[第 4 章 預期成果 20](#_Toc183025910)

[第 5 章 人力配置 25](#_Toc183025911)

[第6章 未來展望：醫療聲音辨識處理系統 26](#_Toc183025912)

[參考文獻 29](#_Toc183025913)

# 表 目 錄

[**表 2.1傅立葉形式的圖** 6](#_Toc168412424)

[**表 3.1方法技術工具表** 20](#_Toc168412425)

[**表 5.1工作人員工作分配圖** 25](#_Toc168412426)

# 圖 目 錄

[**圖 2.1研究流程示意圖** 3](#_Toc168412441)

[**圖 2.2研究流程示意圖** 4](#_Toc168412442)

[**圖 2.3研究流程示意圖** 5](#_Toc168412443)

[**圖 2.4 MFCC流程示意圖** 8](#_Toc168412444)

[**圖 2.5漢明窗** 9](#_Toc168412445)

[**圖 2.6三角濾波器** 10](#_Toc168412446)

[**圖 2.7神經網路收斂流程** 13](#_Toc168412447)

[**圖 2.8 CNN示意圖** 16](#_Toc168412448)

[**圖 2.9** 17](#_Toc168412449)

[**圖 2.10** 18](#_Toc168412450)

[**圖 3.1研究流程示意圖** 19](#_Toc168412451)

[**圖 4.1圖片之目錄示意圖** 21](#_Toc168412452)

[**圖 4.2 MFCC產出之圖** 22](#_Toc168412453)

[**圖 4.3** 23](#_Toc168412454)

[**圖 4.4** 23](#_Toc168412455)

[**圖 4.5 MFCC之數據作圖** 24](#_Toc168412456)

[**圖 4.6陽性的特徵圖** 24](file:///C:\Users\tom30\Downloads\文化專題-1%20(1).docx#_Toc168412457)

[**圖 4.7陰性的特徵圖** 24](#_Toc168412458)

# 研究動機與目的

## 研究動機

聲音辨識技術在醫療領域的研究動機和目的主要包括以下幾方面：首先，通過快速分析患者的語音、咳嗽聲和呼吸聲，提高診斷效率和準確性，特別是在面對大量患者時。其次，減少醫療資源負擔，利用自動化和遠程醫療減輕醫護人員工作壓力。再次，提升病患體驗，通過便捷的語音操作減少等待時間和不便。語音辨識技術還能增強疾病監測和管理，對慢性病和需要長期監測的病人提供持續的症狀監測。特別是在疫情期間，無接觸的篩查和診斷降低了病毒傳播風險。研究目的包括開發高效診斷工具、提升遠程醫療技術、改善患者與醫護人員的互動、推動醫療數據分析和管理、促進個性化醫療以及支持公共衛生和流行病學研究。這些目標旨在利用先進技術提升醫療服務效率、質量和覆蓋範圍，最終改善患者的健康和福祉。

## 研究目的

聲音辨識技術在醫療領域的研究目的包括以下幾個方面：首先，開發高效的診斷工具，通過快速分析和識別疾病（如COVID-19、哮喘、肺炎等）的特徵聲音，提供輔助診斷支持，提高診斷效率和準確性。其次，提升遠程醫療技術，開發基於語音辨識的遠程醫療應用，讓患者在家中或遠程地點獲得高質量的醫療咨詢和診斷服務，減少面對面接觸，降低病毒傳播風險。此外，改善患者與醫護人員的互動，創建智能化的醫療語音助手，理解和回應患者需求，提供自然的互動體驗。推動醫療數據分析和管理，利用語音辨識技術自動轉錄和分析醫療記錄，提高數據的準確性和可用性，支持大數據分析和臨床研究。促進個性化醫療，通過語音數據的個性化分析，提供針對每個患者的定制化治療方案。最後，支持公共衛生和流行病學研究，收集和分析大量語音數據，幫助預測和控制疾病傳播，提升公共衛生安全。

# 文獻探討

## 人工智慧在聲音辨識之應用

聲音辨識在現代社會應用廣泛，例如常見對著智慧音箱講「Hi Siri」即可以語音互動，或是像是智能客服等，在我們生活上該技術滲透率很高，過去較為常見且成熟的方法為隱含式馬可夫模型或是高斯混合模型，當今由於機器學習或深度學習很廣泛，並且因為開源，取得容易，適合各種規模的機構運用。以下簡介國內人工智慧在聲音辨識之應用研究：

1. 運用類神經網路做聲音之辨識

將時域之聲音檔案，這些檔案主要是生活中各種聲音，透過特徵萃取，取得其中特徵後並轉化成向量，再將這些參數運用類神經網路做訓練，計算屬於不同類別之機率。與傳統模型GMM相比，使用神經網路模型錯誤率較低，與傳統方式比較其錯誤率少0.05。

一張含有 文字, 螢幕擷取畫面, 字型, 圖表 的圖片

自動產生的描述

圖2. 1研究流程示意圖

1. 運用道路事故的聲音資料

蒐集道路車禍事故的聲音資料集，將音訊檔案初步篩選和分類，移除無效或是重複資料，音訊檔案主要分為碰撞聲及剎車聲兩大類。將蒐集到的聲音資料集進行前置處理，並切割為時間長度相同的音訊檔案。當音訊檔都切割完成後，使用Python 的 librosa 套件來分析聲音訊號，並使用短時距傅立葉變換產生頻譜圖，最後再用 YOLOv7-tiny 預設的配置來訓練車子滑行與撞

擊的頻譜圖。其方法架構流程圖如圖2. 2，研究中產出頻譜圖如圖2. 3，混淆矩陣結果如表2.1

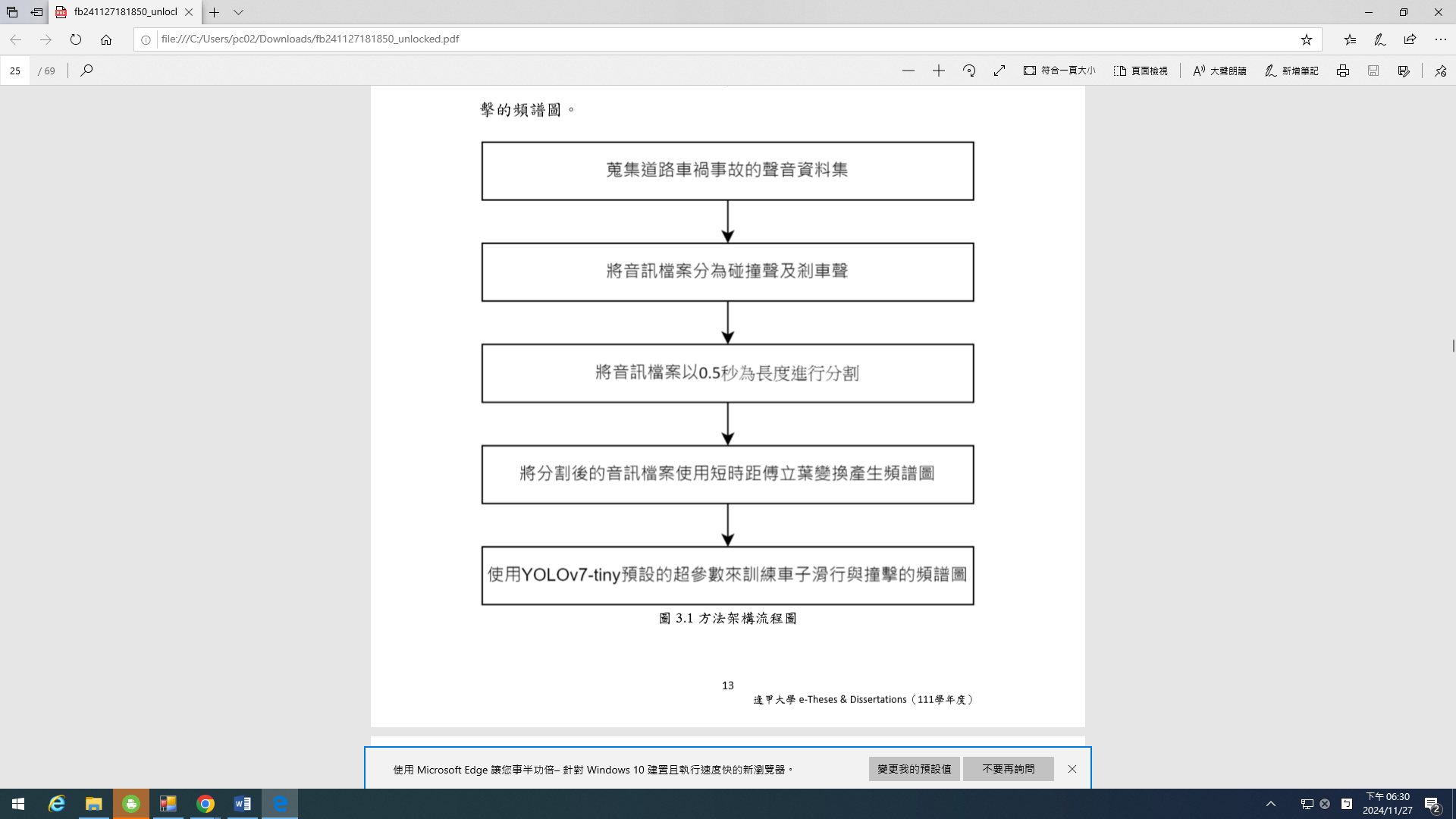


圖2. 2 方法架構流程圖

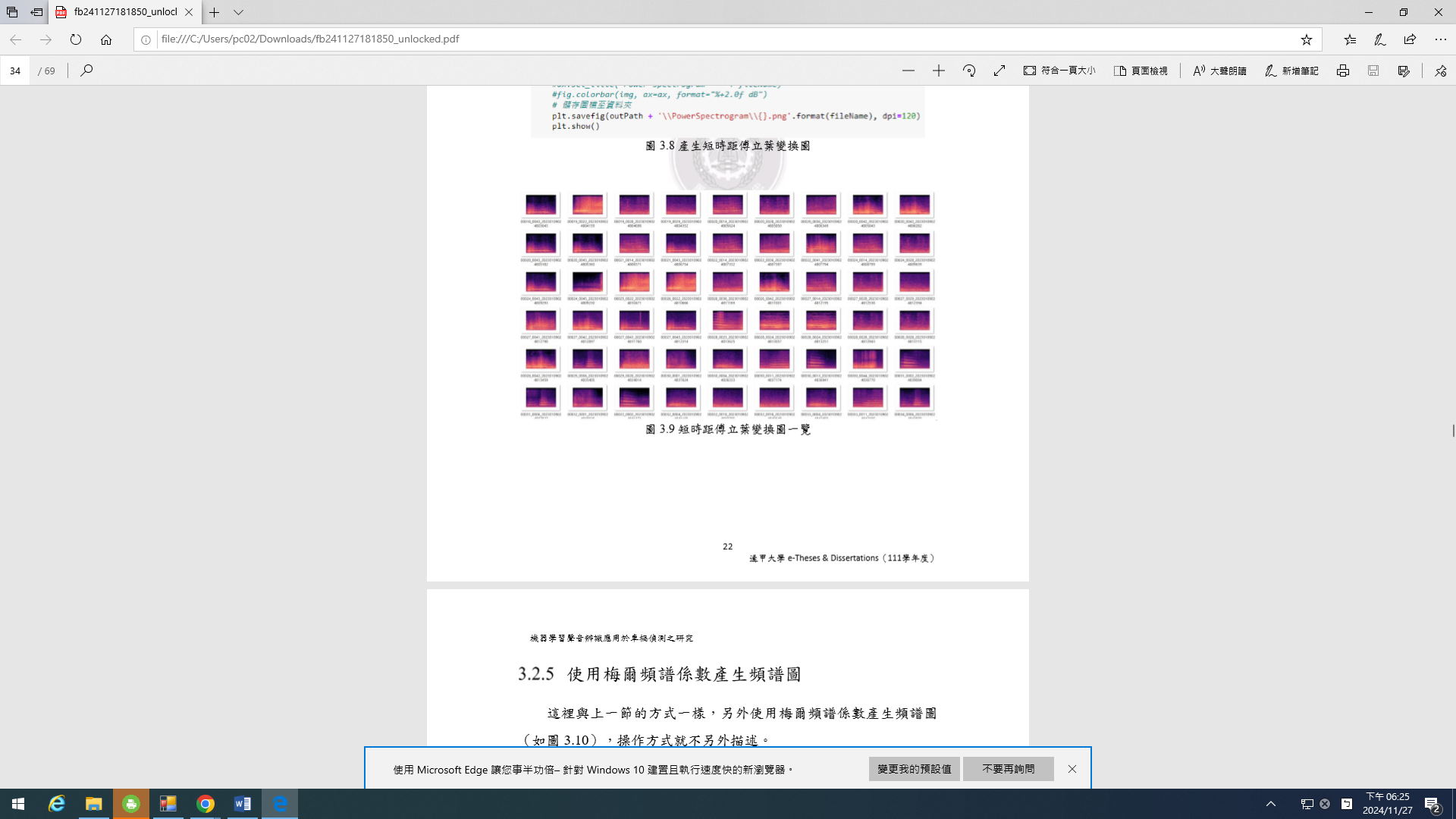


圖2. 3 頻譜圖

表2.1 混淆矩陣

|  |  |  |
| --- | --- | --- |
|  | 實際為accident | 實際為skidding |
| 預測為accident | 0.98 | 0.02 |
| 預測為skidding | 0.02 | 0.98 |

1. 利用收集的聲音資料

藉由訊號前處理轉換後找到相對應的特徵並加以分析，傳統之聲音訊號處理方式會用一些特徵如頻率、振幅來進行分類，但若是特徵無法找到有效的差異，使得分類準確率無法有效提升。因此該研究使用人工智慧中的卷積神經網路模型來建立聲音的分類模型。先運用梅爾倒頻譜係數原理將聲音訊號轉換成頻譜圖再匯入訓練模型進行訓練。其中卷積神經網路克服了傳統聲音識別只藉由單一特徵進行分類導致不準確問題。梅爾倒頻譜係數將數位的聲音訊號經由頻率之間的差異進行特徵提取。由於資料有些不平衡，因此藉由資料擴增的方式來改善分類模型之準確率，由實驗結果顯示能夠提升分類的準確性30%。

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

圖2. 3研究流程示意圖

## 聲音特徵提取

聲音可視為是一個連續的函數，若要進行訊號分析可以用以下方式進行特徵萃取：

1. 傅立葉轉換：

將聲音訊號以頻率域的方式呈現。轉換後聲音的本質特性還在，會將聲音訊號以「頻率」的方式來表達。傅立葉轉換本身為一種數學上的線性積分變換方式，藉由傅立葉轉換將週期函數轉換為另一種函數。傅立葉轉換根據輸入的訊號類型可以分成如表2.1：

**表 2.1傅立葉形式的圖**

|  |  |
| --- | --- |
| 方法轉換方法 | 圖形 |
| 非週期性連續訊號傅立葉變換 | 一張含有 設計 的圖片  自動產生的描述 |
| 週期性連續訊號傅立葉級數 | 一張含有 廚房用具, 餅乾成型切割刀 的圖片  自動產生的描述 |
| 非週期性離散訊號離散時域傅立葉變換 | 一張含有 扳手, 工具 的圖片  自動產生的描述 |
| 週 期 性 離散訊號離散傅立葉變換 |  |

在分析聲音時，會「短時距分析」（Short-term Analysis）

為主，在短時間內是相對穩定的。

詳細說明如下：

1. 音框

將聲音訊號切成數個音框，其中音框長度約為20 ms ~ 30ms之間，若音框太大，比較無法找到其中聲音特性隨時間的變化。反之，若太小會無法找出聲音本身的特徵，因此音框的用途主要是能夠切割成數段有週期性的聲音，各音框相鄰之間，可以保有重疊，藉此保有資料的連續性，一般來說重疊的部分約為音框長度的 1/2 到 2/3 不等，實務上亦可依照需求來做調整。

1. 特徵提取

將訊號音框化後，針對各個音框進行特徵提取，除了常見藉由聲音基本資訊，例如音量、音高或音質的方法，也有如過零率(zero-crossing rate, ZCR)、MFCCs 進行特徵提取參數…等方式做為特徵提取的方式。

1. 分析

根據音框的特徵提取後，針對各音框的特徵資訊，進行分析與辨識。

1. 梅爾倒頻譜係數：

MFCCs 在1980 年由 Davis 和 Mermelstein 共同所提出的，從那之後 MFCCs在語音識別領域被廣泛的利用，包括了語音識別，聲音識別，環境聲音識別等[4]，其流程參照下圖：

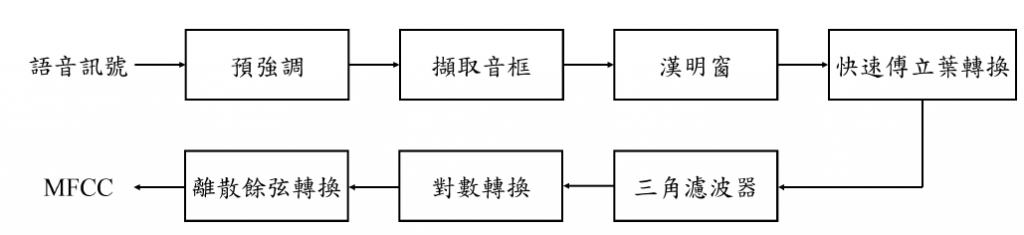


圖2. 4 MFCC流程示意圖

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。

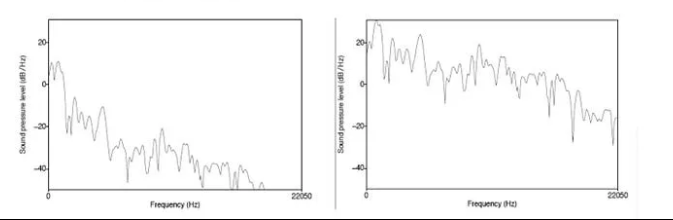


圖2. 5預強調波形變化

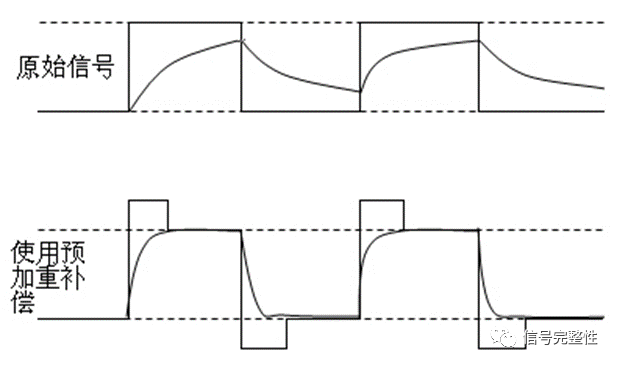


圖2. 6預強調原理

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

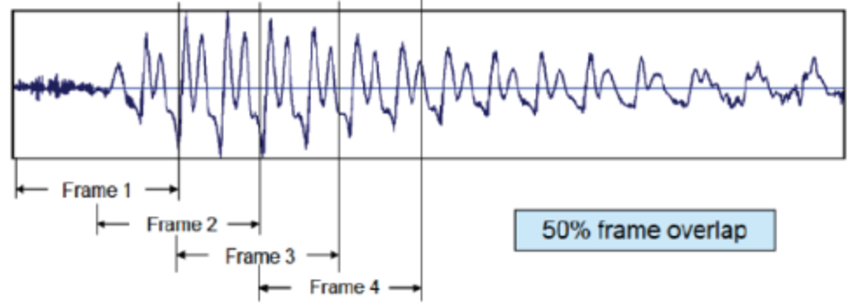


圖2. 7 音框化示意圖

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，

一張含有 行, 繪圖, 圖表, 斜率、斜坡 的圖片

自動產生的描述

圖2. 8漢明窗

其公式如下：

其中N是窗的長度α=0.54 和 𝛽=0.46 是窗的常數

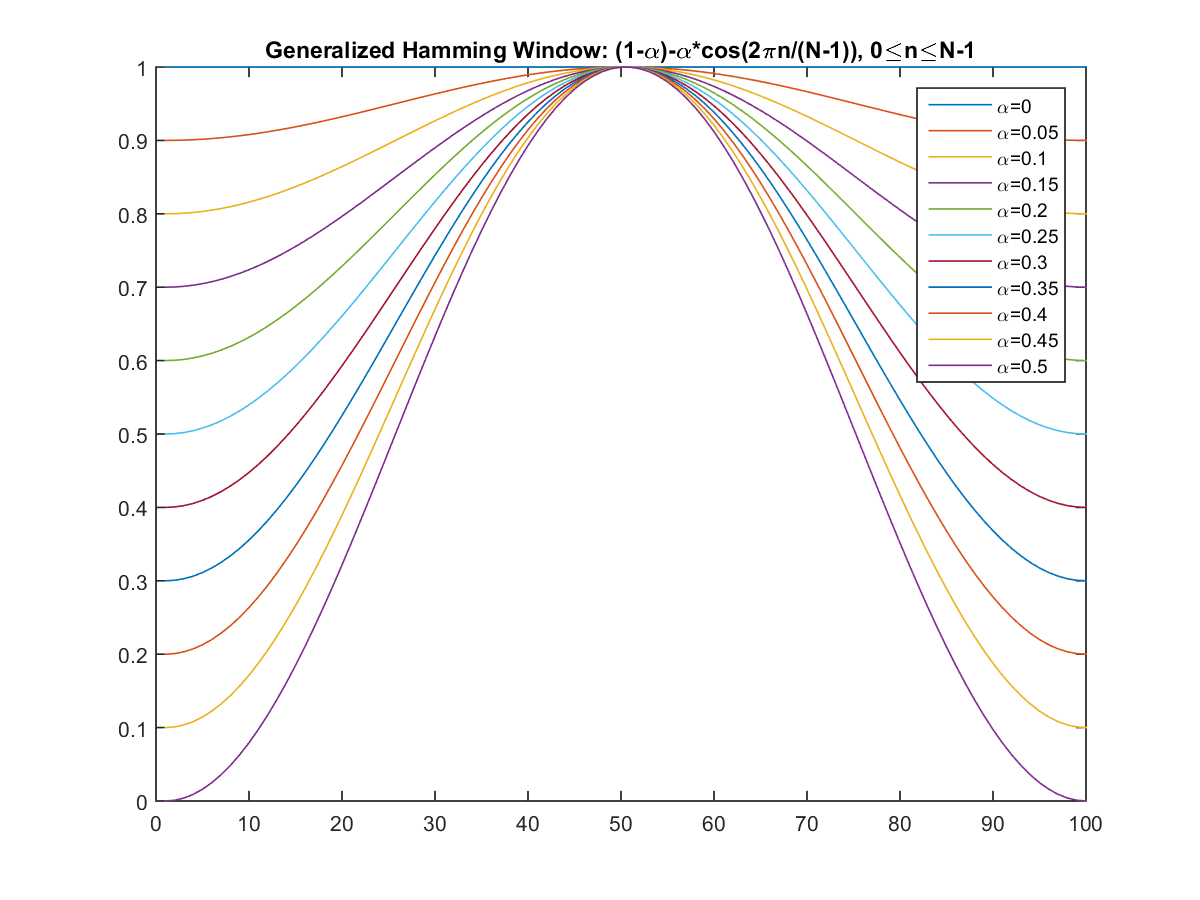


圖2. 9漢明窗改動不同α的圖

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。

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，如圖2.6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 2.6三角濾波器**

其公式為H[m,k] 表示第m個濾波器對第k個頻率分量的權重，其中：

1. 當 k < f[m-1] 時： H[m,k] = 0
2. 當 f[m-1] ≤ k ≤ f[m] 時： H[m,k] = (k - f[m-1]) / (f[m] - f[m-1])
3. 當 f[m] ≤ k ≤ f[m+1] 時： H[m,k] = (f[m+1] - k) / (f[m+1] - f[m])
4. 當 k > f[m+1] 時： H[m,k] = 0

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. 特徵提取參數(Feature Extraction)

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，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。

1. 對數轉換其公式如下

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] 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。

另外一種常見的聲音特徵萃取手法為非負矩陣分解(Non-negative Matrix Factorization)(NMF)，該方法可以達到數據降維和特徵學習之效果，將一個非負矩陣分解為兩個非負矩陣的乘積。這兩個矩陣的乘積能夠近似原始矩陣，並且這些分解後的矩陣通常有助於發現數據中的潛在結構和模式。其數學如下：

其中W 是一個 m×r 的矩陣，代表「基底」或特徵矩陣，H 是一個 r×n 的矩陣，代表在基底上加權的係數。其主要特色如下：

1. 非負約束：NMF要求所有矩陣元素都是非負的，即大於或等於零。這使得NMF能夠揭示數據中的積極關係或結構，並且在某些應用中（如圖像處理、文本分析等）這一點尤其有用，因為數據本身通常是非負的。
2. 解釋性強：由於NMF將原始數據矩陣分解成兩個低維矩陣，這兩個矩陣的元素通常可以解釋為某些隱含的因素或特徵。例如在文本分析中，這些特徵會對應於潛在的主題。
3. 降維：通過選擇較小的維度，NMF實現了數據的降維。這有助於去除噪音並突出數據中的核心結構。

NMF之應用如下：

1. 文本挖掘與主題建模：在文本分析中，NMF可以用來從文檔-詞語矩陣中提取主題。每個主題可以由一些詞語的加權組成，這些詞語通常反映了該主題的關鍵內容。
2. 聲音事件檢測：NMF還可以應用於聲音事件的檢測和分類。例如，對於環境噪聲監測系統或智能監控系統，NMF可以用來從音頻中提取出不同的聲音事件，如門鈴聲、車輛聲或人聲，並將它們與背景噪音分離開來，從而更準確地檢測和識別特定的聲音事件。
3. 音頻特徵萃取：NMF也用於音頻分類任務中，例如，對音樂片段、語音錄音或環境聲音進行分類。通過將音頻信號分解為基本成分，NMF可以提取有用的特徵，這些特徵可以用來進行後續的分類或聚類。例如基於音頻特徵的情感識別、語者識別等。
4. 語音分離和增強：在語音處理中，NMF被用來處理具有背景噪聲的語音信號。比如，當語音信號與背景噪音（如街道噪聲或其他對話）混合時，NMF可以用來提取和增強語音部分，並消除或減少噪音。

## 機器學習(Deep Learning)

機器學習範疇很多主要分監督式學習與無監督式學習，監督式學習就包含線性回歸、邏輯回歸、支持向量機(SVM)、決策樹、隨機森林，主要特點是需要有標註來進行訓練，無監督式學習有K-均值聚類、層次聚類、主成分分析，不需要標註即可以訓練模型，本研究參考李氏[10]使用SVM進行聲音的分類。

支持向量機於 1992 年由 Boser、Guyon及Vapnik等人所提出，主要的概念就是希望能在一群由 2 種不同類別混合而成的資料群集中，找到一個最佳的超平面(hyperplane)將這些資料區分開來，以 2 維資料為例，有兩群不同的資料(o 與×)分佈，可用1條直線將其分成2類如圖 2.7，若在3維甚至多維空間中這條直線則可想像成平面，故稱作超平面。此外，希望這條區分的超平面與兩群資料的距離(又稱邊距，margin)越遠越好，如此在進行分類時也就 越 容 易 判 別 ， 這 樣 的 超 平 面 則 稱 作 最 佳 超 平 面 (optimal hyperplane)，而與最佳超平面平行並且最靠近兩邊資料之平面稱為支持平面(support hyperplane)，在支持平面上的資料即稱為支持向量(support vector)。

一張含有 鮮豔, 圓形, 水, 設計 的圖片

自動產生的描述

**圖 2.7二維平面最佳超平面示意圖**

**一張含有 兒童藝術, 圖表, 螢幕擷取畫面, 設計 的圖片

自動產生的描述**

**圖 2.8三維空間最佳超平面示意圖**

在解題時，可以假設

X1=[2,3] y1=1

X2=[1,1] y2=1

X3=[-1,-2] y3=-1

假設平面為w1X+w2Y+b=0

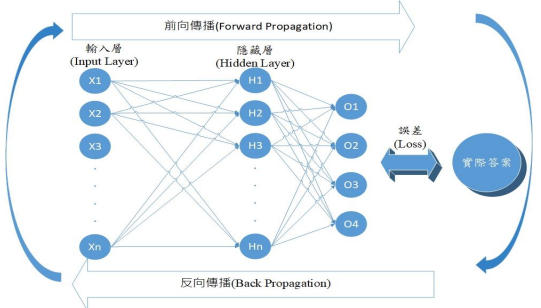
可以得出下列算式

來求最佳化問題，最終解決下列算式即可以得到w1及w2

2.4深度學習(Deep Learning)

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)，以下我們將逐一介紹各層，如圖 2.7所示。



**圖 2.8神經網路收斂流程**

1. 輸入層(Input Layer)

由多個神經元所組成，主要功能是接收輸入資訊，圖2.7中以 X1、X2…Xn為輸入層各輸入節點，一般為數值型的資料，但依據不同的神經網路模型，所輸入的型態也就不一樣，如數值型的輸入資料、以圖片做為輸入資料。

1. 隱藏層(Hidden Layer)

介於輸入層和輸出層之間的神經網路層，主要功能是透過非線性轉換函數運算，由神經元(Neuron)組成，透過前向傳播(Forward propagation) 計算出我們的輸出值，最終將值傳至輸出層，如圖 5 中以 H1、H2…Hn表示為隱藏層中各隱藏節點。通常隱藏層的層數不容易決定，須由多次嘗試運算找到最佳解。使用過多的層數會導致模型收斂過慢或無法收斂；而較少的層數則無法應付較困難的問題。最更新我們隱藏層的參數，經過運算所得到的預測值，我們藉由縮小預測值與實際標記值(Label)，更新我們隱藏層間的參數，最終可訓練出一組權重(weights)值，提供給輸出層進行評估。

1. 輸出層(Output Layer)

主要分析分類結果以及評估權重的資訊，如圖 5 中以 O1、O2、O3表示為輸出層中各輸出節點，節點數量也代表著分類的數量。

綜合上述各層介紹，神經網路透過前向傳播(Forward Propagation)透過神經元與權重的交叉運算，預測出解答與實際值比對，當預測與實際值相異時稱為相異值為損失(Loss)，依據 Loss 值進行反向傳播求導數，並進行更新神經網路的權重，反覆進行直到模型收斂。

在聲音分類中有許多可行的方法，如：基於機器學習中支持向量機(Support Vector Machine, SVM)進行呼吸聲音的分類，以決策樹進行聲音分類，皆為機器學習的方法。但是在對於環境聲音中如果類別單純時能夠有著準確的分類結果，但當分類種類呈倍數增加時，將大大影響了支持向量機與決策樹的分類效果，因此我們將考慮使用深度學習的方法，利用模型進行學習訓練，藉由學習特徵分類出生因資料。

常見的深度學習架構包含：多層感知器(Multilayer Perceptron, MLP)、人工神經網路(Artificial Neural Network, ANN)、深度神經網路(Deep Neural Network, DNN)、卷積神經網路 CNN(Convolutional Neural Network)、遞迴神經網路 RNN(Recurrent Neural Network)…等。

在深度學習裡，如基於傳統 ANN 的聲音分類方法[24]，基於 RNN的長短期記憶網路(Long short-term memory, LSTM)的城市聲音分類，對於聲音分類，使用各種分類模型進行訓練的技術已相當的純熟以集多元。而本論文選擇 CNN 做為分類模型的原因是由於影像識別模型的特性，已經證實了有很好的分類效果。我們能夠將轉換後的聲音頻譜圖以完整時間區間進行保留，並匯入到 CNN 模型裡進行模型訓練，在 CNN 模型中輸入為一個二維矩陣，圖型資料就是一個典型二維矩陣。因此能夠符合其模型的輸入，並將輸出的結果進行準確性的驗證，達到分類預測的能力。比起傳統利用聲音特性所建立的環境聲音分類預測模型，此方法大大提升了準確率。卷積神經網路:

在各種深度學習都有屬於自己的特性，CNN 為 DNN 之衍生，比起 DNN，CNN 多出了卷積層(Convolutional Layer)與池化層(Pooling Layer)兩層，這兩層與一個全連結層組合，此全連結層即為 DNN 的輸出層，最後並使用 Softmax 函數來輸出分類結果，如圖2.8 所示。

一張含有 圖表, 文字, 行, 方案 的圖片

自動產生的描述

**圖 2.9 CNN示意圖**

由於卷積神經網路的特性擅長於圖片的識別，透過上述的卷積層 與池化層可以將高維度的輸入劃分為多個低維度的特徵。根據此方式 學習特徵後，可做為辨識用的分類器。而若要處裡非圖像辨識有關的 資料時，則需將輸入資料轉換成一個二維矩陣，透過此方式讓卷積神 經網路可以廣泛的應用在許多領域，以藉此來分析環境聲音的分類。 CNN 各層運作的說明將於下列各小節中進行說明。

卷積層：在卷積運算(Convolution)就是將原始圖片的與特定的卷積核做卷積運算(以符號⊗表示)，如圖 7 的數值為例，其中中間的卷積核(3\*3)會隨機產生，並根據輸入圖像的大小確定評估值，如同掃描的方式，如圖2.9所示。將卷積核從圖片的最左上方的像素逐一掃描，我們可以發現使用此方式掃描矩陣，而在輸入圖片中卷積核無法完整的將聲音資料進行掃描，導致掃描出來的圖片找不到特徵，因此將每個像素皆為卷積核的正中心，當卷積核掃描到圖片的邊緣像素時，我們會以補零的方式進行，目的在於讓圖片保持原本輸入的大小，防止卷積後圖片越卷越小的情形

一張含有 圖表, 數字 的圖片

自動產生的描述

**圖 2.10**

池化層：在完成卷積層的特徵提取之後，將進入到 Pooling(池化層)，常見池化層為平均池化層與最大池化層，其中又以最大池化層最常見。其原理是透過特徵圖提取其最大值，一般取決於池化的大小，並設定它的步長，最大池化層主要用於提取重要特徵、加速收斂並縮小特徵圖尺寸。

全連接層：提取完特徵後，將特徵帶入到全連接層。基本上全連接層的部分就是將之前的結果平坦化之後接到最基本的分類，如圖2.10所示。

一張含有 文字, 行, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

**圖 2.11 映射示意圖**

# 研究內容、方法與工作項目

## 研究方法設計

本研究參考文獻中的做法，採納MFCC做數據的預處理，將聲音訊號換算成向量後，再分割成訓練及測試集資料，再帶入不同模型，以準確率來評估模型優劣。

語音輸入

語音輸入

NMF

MFCC

抽樣(oversampling, undersampling)

抽樣(oversampling, undersampling)

分成train/test

分成train/test

機器學習

機器學習

決定特徵萃取方式以及抽樣方式

決定特徵萃取方式以及抽樣方式

抽樣(選擇由上一步最佳之方法)

語音輸入

特徵萃取

分成train/test

類神經網路(1D/2D)

模型評估

**圖 3.1研究流程示意圖**

使用之數據來源為https://www.kaggle.com/datasets/lynnkong/covid19voicediagnosis-2/data，該網站提供很多公開且珍貴之數據，並且其數據有一定之根據性及可靠度。由於機器學習模型訓練較快，會先以SVM法做前測，決定使用MFCC或NMF其中一種特徵萃取方法，以及決定使用oversampling或under-sampling方法，取最佳之結果用在類神經網路。在訓練過程中會使用測試資料測試模型是否過擬合狀況發生，若有則會用dropout方式關閉部份節點，因為數據有咳嗽聲以及呼吸聲，預計會以其中一個數據集建立模型後，再以這份模型去套用在另一份數據集，試著不同參數例如學習率測試是否可以直接套用或是可以用較少的步驟訓練，達到遷移學習的效果。

## 使用工具

本研究使用google提供的colab之python編譯器，做深度學習演練。

系統主要分三個架構：

1. 數據前處理

將聲音訊號變成數字資訊。

1. 模型建置模組

利用不同深度學習架構

1. 利用flask製作成網頁APP

Flask 是一個輕量級、易於使用的 Python Web 框架，非常適合用於快速構建網頁應用程式。它以其簡單靈活的特性，讓開發者可以快速實現從簡單的靜態頁面到複雜的動態應用程式。其中內容包含讓使用者錄製一段聲音後，將聲音由瀏覽器提供的 Web Audio API 或 MediaRecorder API 進行錄音，並使用Flask來接收和存儲音訊檔案。之後再按送出，由後臺讀取聲音檔案後，利用訓練好的模型預測其類別，再將預測好的類別返回前台告知使用者

**表 3.1方法技術工具表**

|  |  |
| --- | --- |
| 套件 | 用途 |
| Keras | 實現深度學習 |
| Librosa | 處理聲音訊號 |

# 研究成果

本研究使用的資料集，陽性資料80筆，陰性資料877筆，可以利用套件將MFCC產出的圖繪出如圖4. 1及圖4. 2。

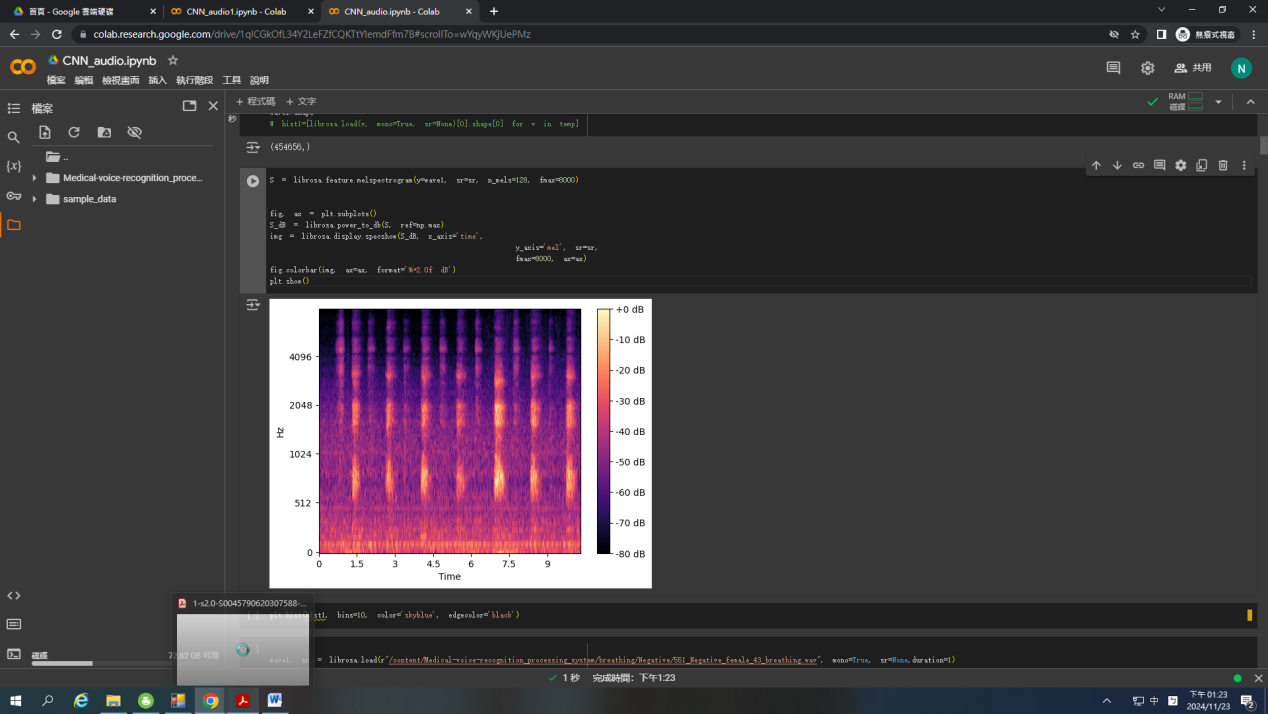


圖4. 1 MFCC產出陰性聲音圖及對應程式碼

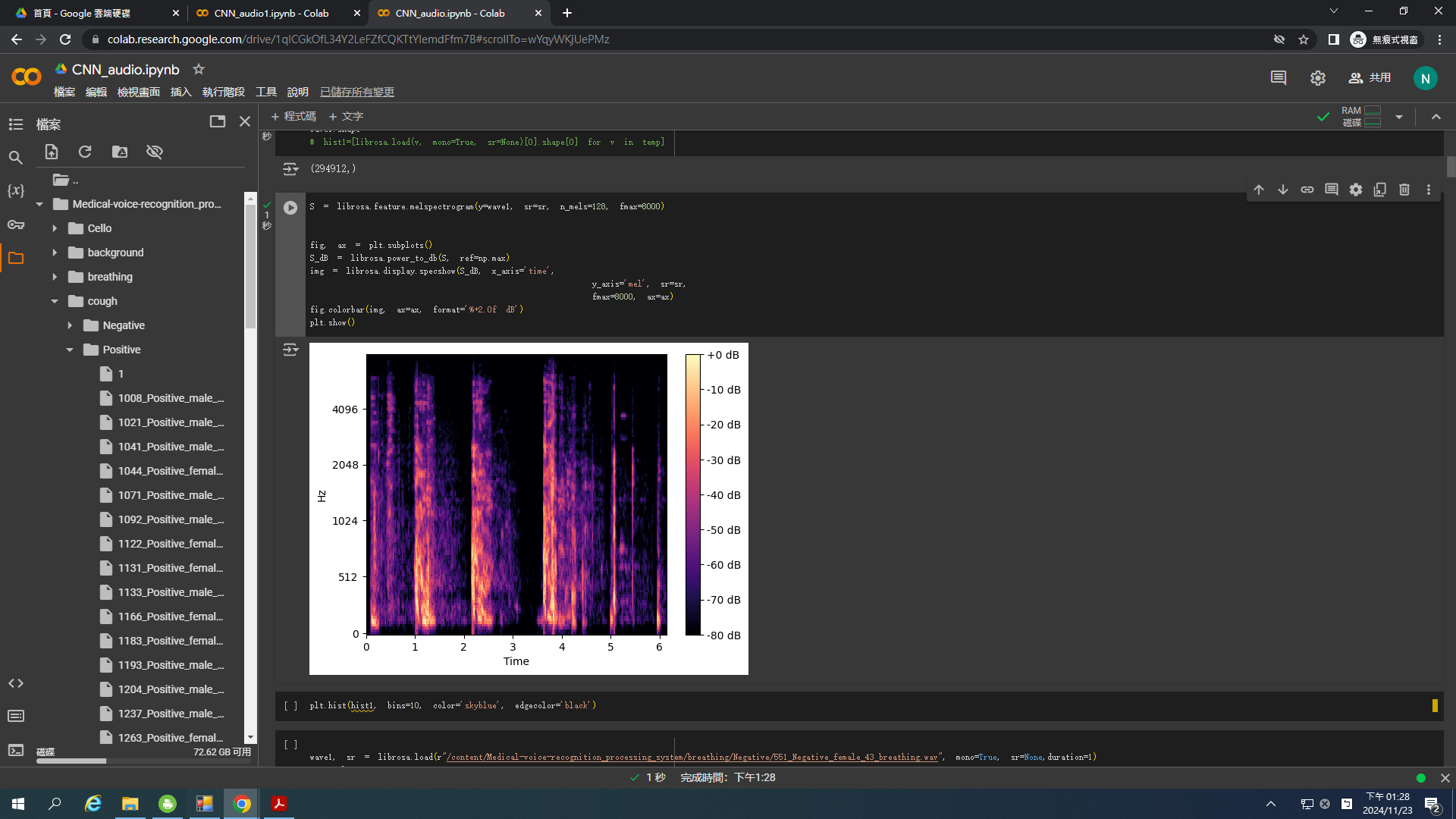


圖4. 2 MFCC產出陽性聲音圖及對應程式碼

因數據存在數據不平衡，會透過重複抽樣或是少抽樣達到數據平衡，運用librosa套件讀取聲音檔案，分別用MFCC以及NMF兩種方法做特徵萃取，其對應程式碼MFCC如圖4. 3，NMF為圖4. 4。

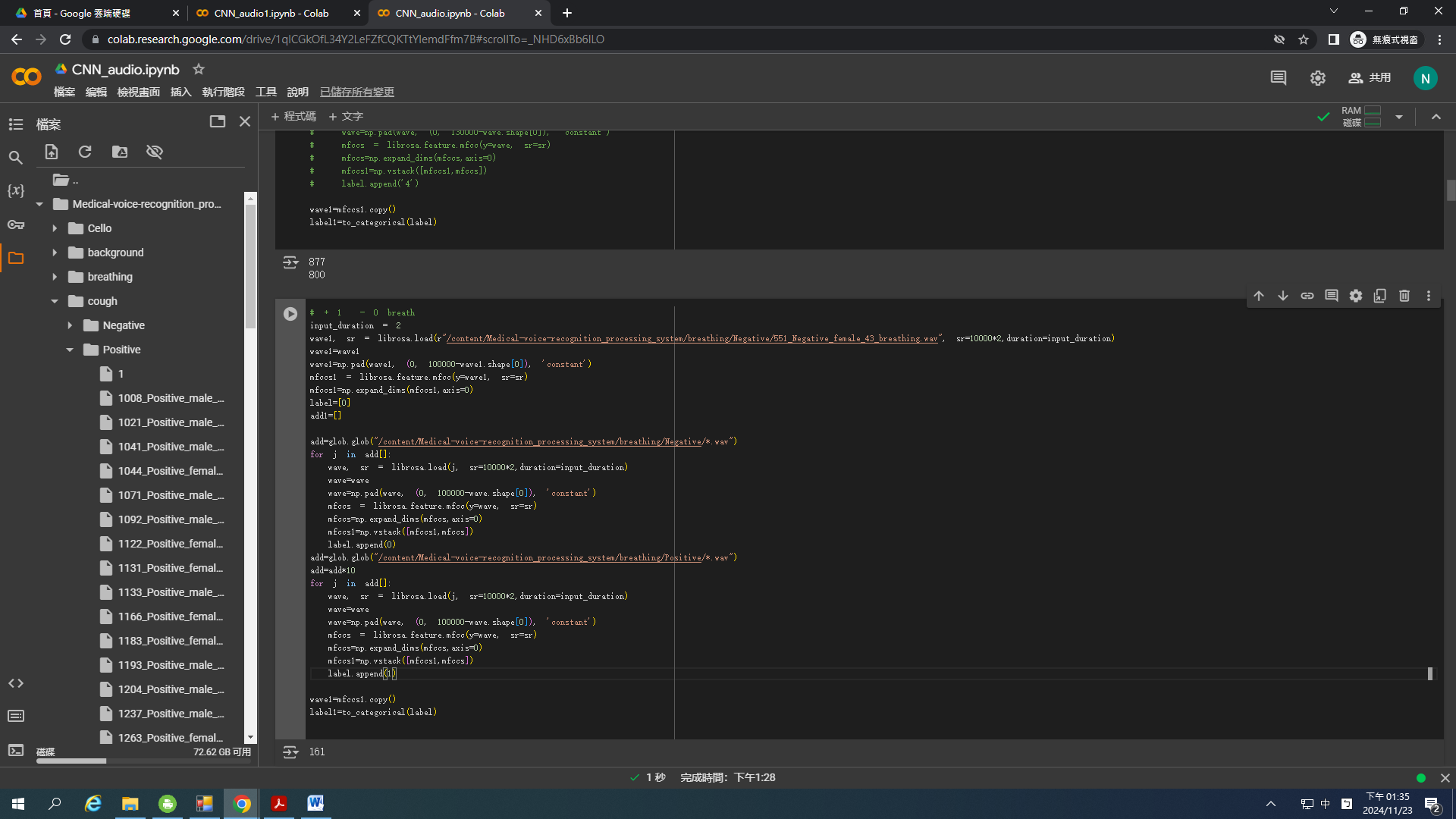


圖4. 3 MFCC處理數據之程式碼

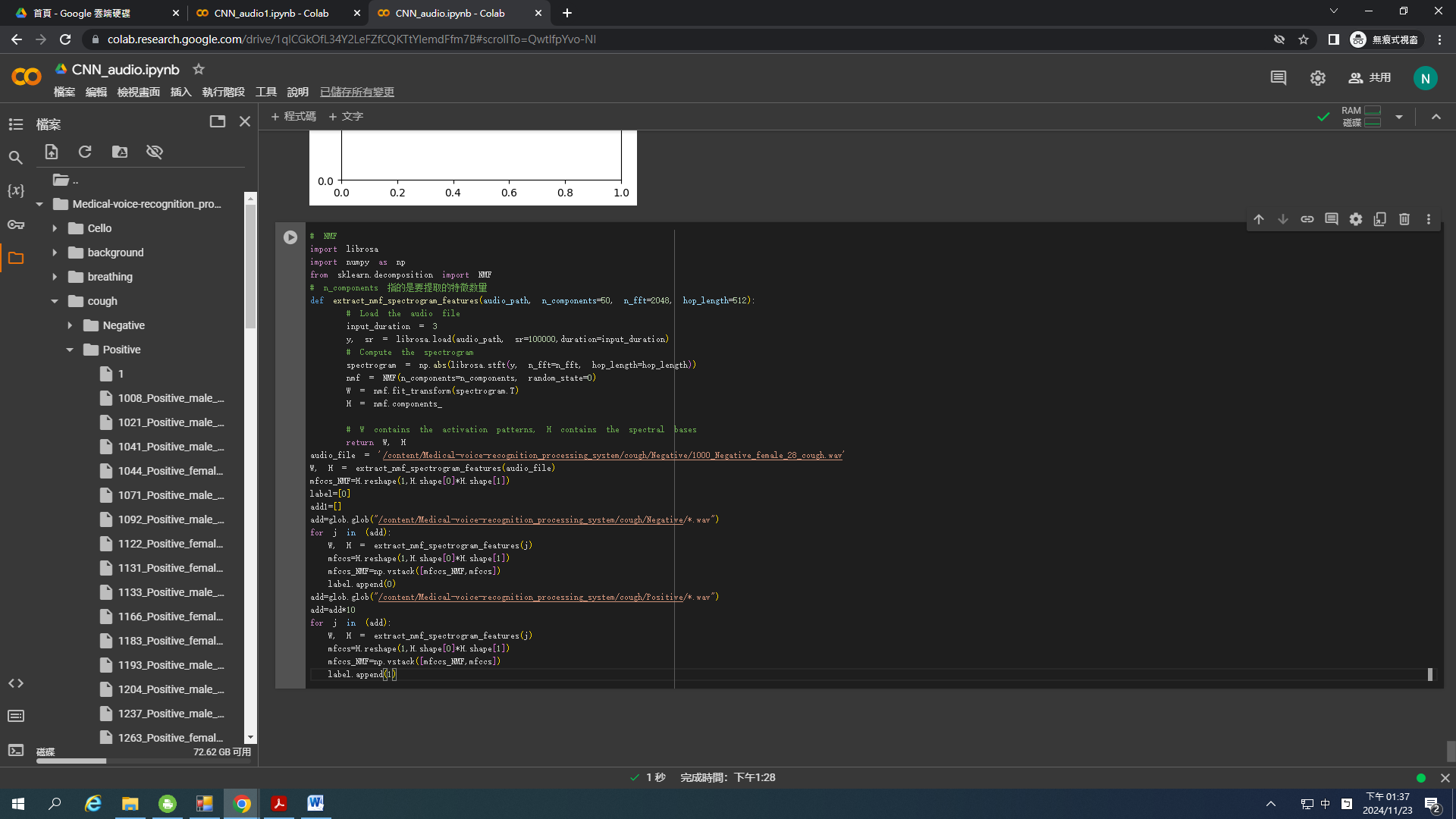


圖4. 4 NMF處理數據之程式碼

並且兩種方法再搭配重複抽樣或是少抽樣達到數據平衡，將實驗條件整理成表4. 1。SVM的kernel都是用linear，重複抽樣的方式因為陽性只有八十筆，因此所有資料重複取十次達到800筆進入資料集，而陰性則全取，若為少抽樣則陽性數據取80筆，陰性只取其中80筆進行訓練，實驗的成果如表4. 2，可以看出由於少抽樣學習的筆數太少不管是陰性或是陽性辨識正確度都很低，因此後續使用類神經網路會採納重複抽樣，若是同樣是重複抽樣可以看出使用MFCC法正確率高於NMF。而SVM本身可以透過調整其C值作準確率的優化，本研究將C值從0.1，每次增加0.1直到1進行一個實驗如表4. 2，可以看出最高就是C為1時準確率為94%。

表4. 1條件表

|  |  |
| --- | --- |
|  | 條件列表 |
| 特徵萃取方式 | MFCC、NMF |
| 抽樣方法 | 重複抽樣、少抽樣 |
| 機器學習方法 | SVM |

表4. 2實驗結果

|  |  |
| --- | --- |
|  | 準確率 |
| MFCC+重複抽樣 | 94% |
| MFCC+少抽樣 | 47% |
| NMF+重複抽樣 | 67% |
| NMF+少抽樣 | 50% |

表4. 2 使用MFCC+重複抽樣對SVM不同C值比較表

|  |  |
| --- | --- |
| C值 | 準確率 |
| 0.1 | 91% |
| 0.2 | 91% |
| 0.3 | 91% |
| 0.4 | 92% |
| 0.5 | 92% |
| 0.6 | 94% |
| 0.7 | 94% |
| 0.8 | 94% |
| 0.9 | 94% |
| 1 | 94% |

由上述成果得知特徵萃取方式使用MFCC及重複抽樣在SVM訓練過程中可以得到較好之結果，因此後續使用類神經網路依照這個條件做訓練。CNN模型使用1D及2D的結構進行模型建構，1D如圖4. 5，2D如圖4. 6。本研究預計要預測的聲音，用MFCC取得特徵向量後，若是畫成2D的圖可以參照圖4. 5。

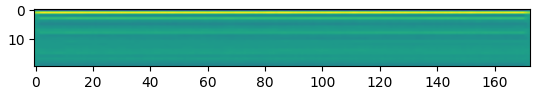


圖4. 5 MFCC之圖

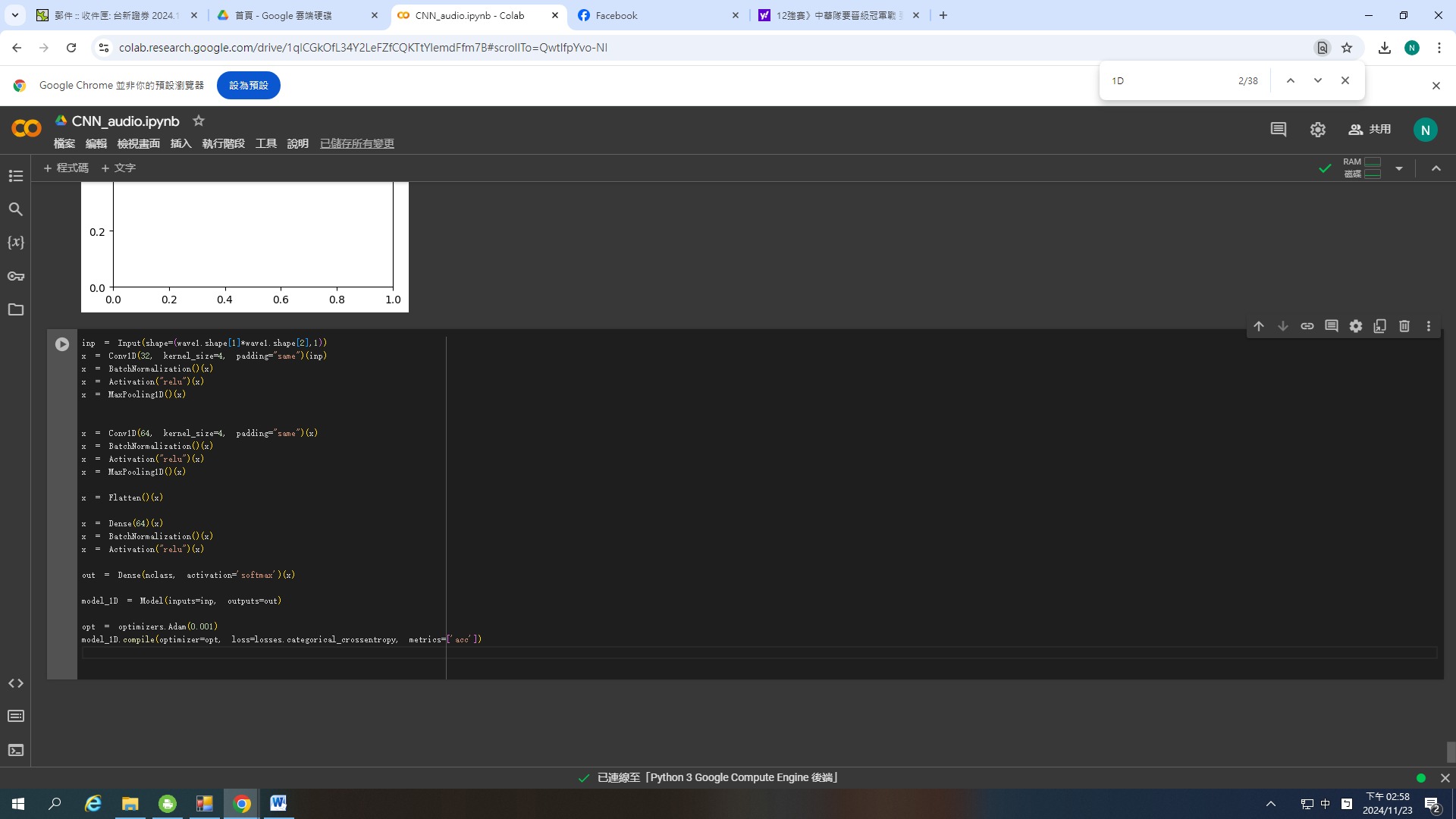


圖4. 6 CNN 1D之程式碼

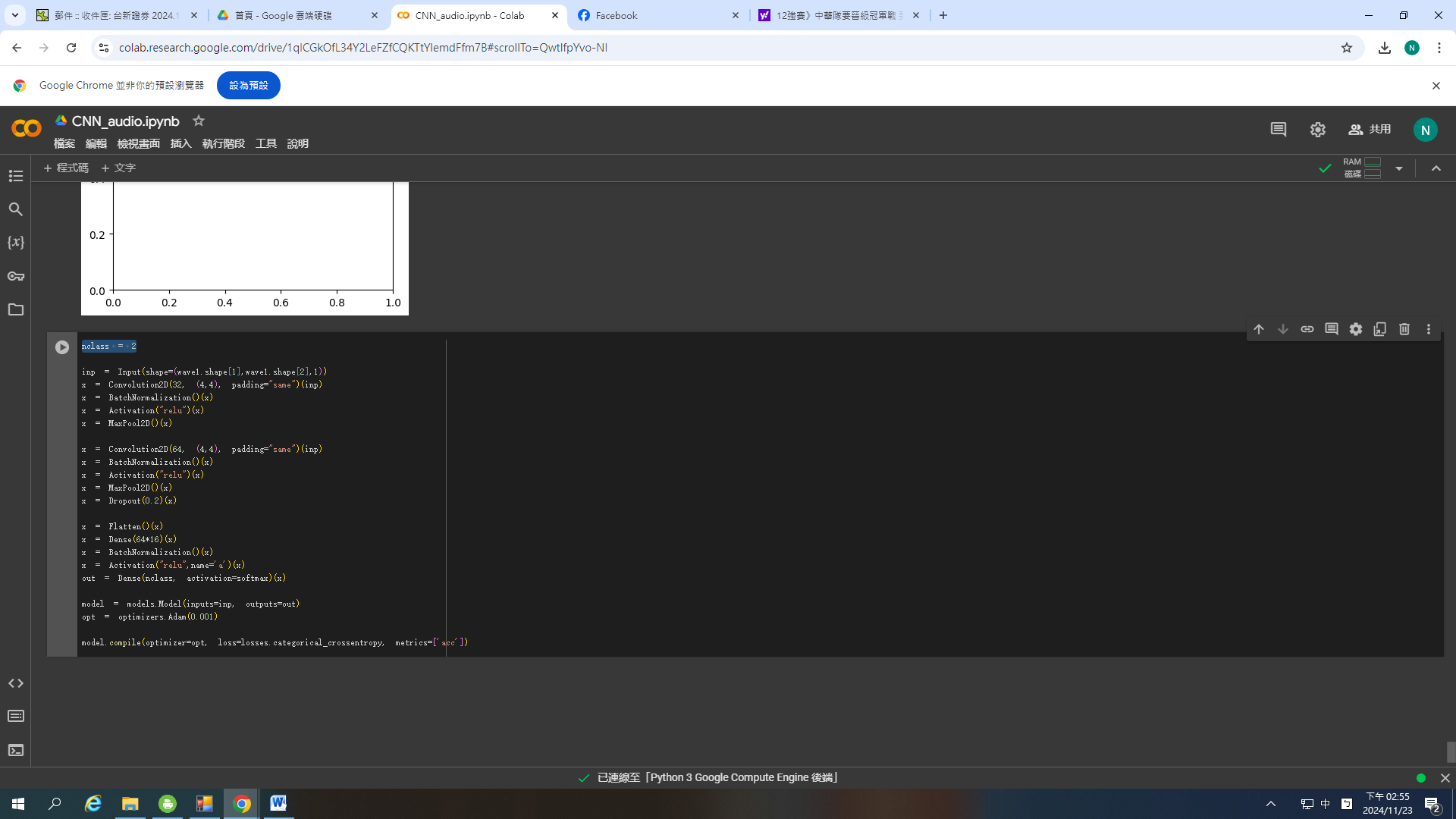


圖4. 7 CNN 2D之程式碼

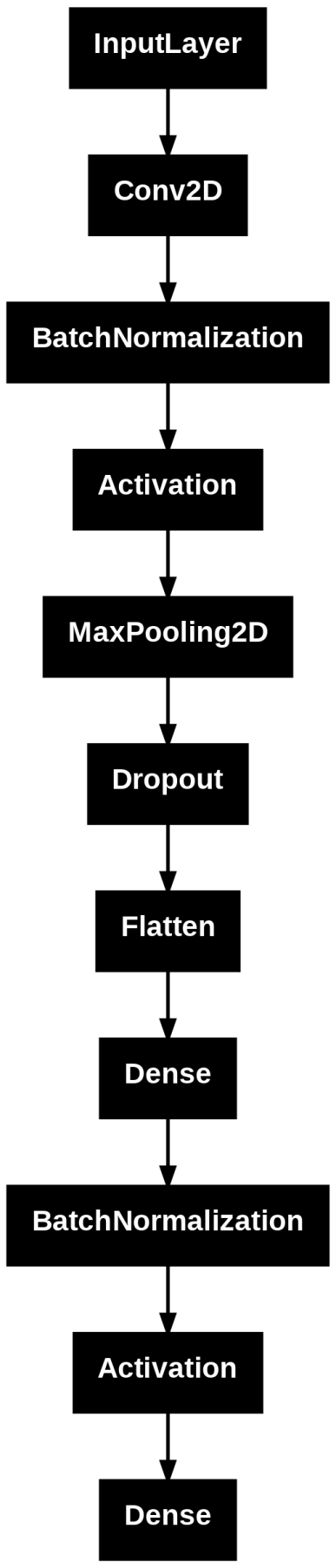
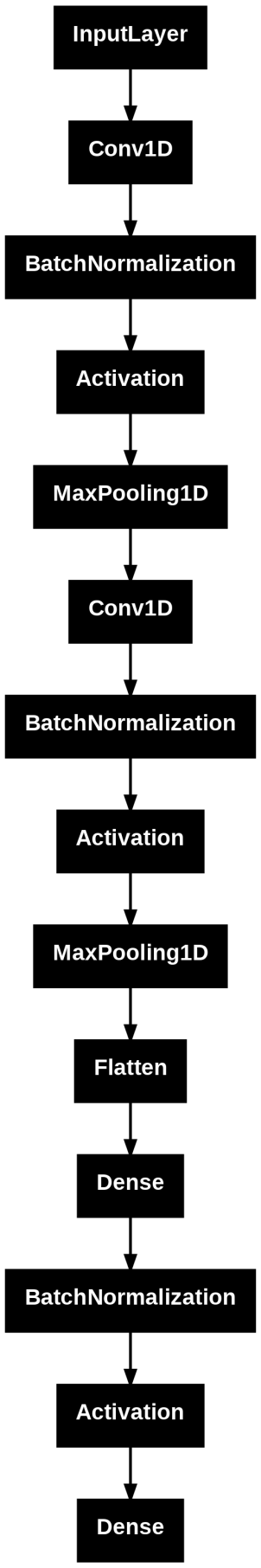


圖4. 8 CNN神經網路模型

建構的模型之流程示意圖參照圖4. 8，以此來進行聲音類別的預測，訓練的過程將準確率隨著epoch數增加繪製成折線圖，1D的部分如圖4. 9，2D的為圖4. 10，可以看出最終train資料即可以達到95%以上，並且沒有過擬合狀況發生。

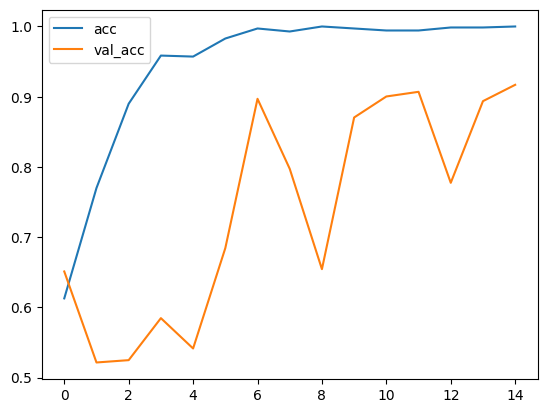


圖4. 9 1D CNN神經網路模型訓練過程

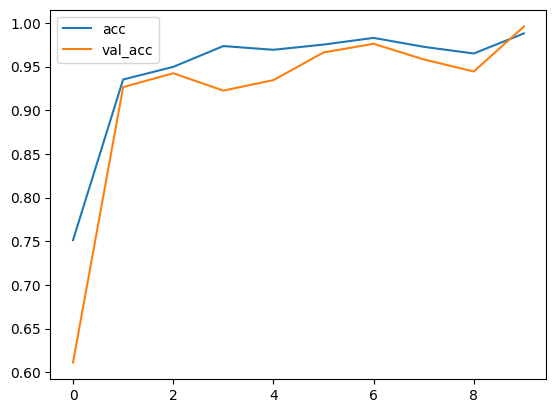


圖4. 10 2D CNN神經網路模型訓練過程

最終本研究以尚未訓練過的數據進行最終測試，以1D而言其混淆矩陣如圖4. 11，2D為圖4. 12。

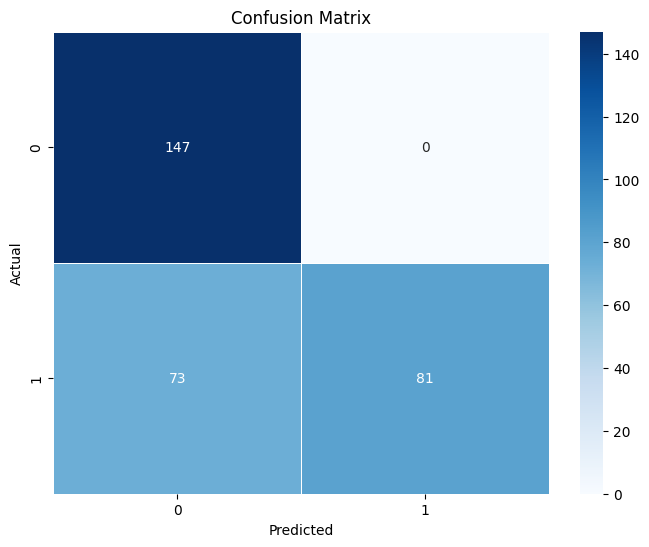


圖4. 11 1D CNN混淆矩陣



圖4. 12 2D CNN混淆矩陣

1D之CNN模型在訓練時就發現其驗證集準確率變動很大，而其混淆矩陣在陽性的辨識結果也不好，相較於2D，，其準確率為91%，後續要做成APP將採用2D CNN模型。

本研究想要了解經過激發層後的dense層，原本MFCC的數據轉換成什麼樣的數據，因此原本模型輸出為類別，將其改成輸出dense層權重值，分別用陰陽做區分參照圖4. 14及圖4. 13，直觀上可以看出差異很大，陽性的熱圖偏深色，也表示模型真的有學到其中差異以辨識出陰陽。

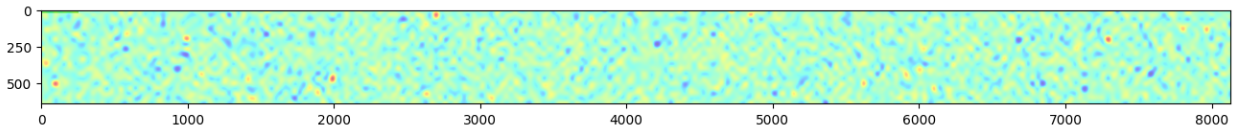
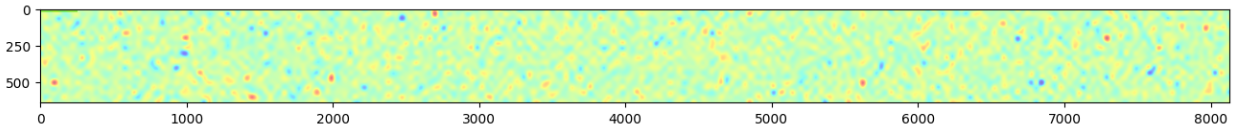
****

圖4. 13陽性的特徵圖

圖4. 14陰性的特徵圖

最後本研究將2D CNN模型輸出，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4. 15，之後後台的upload\_file這個函式如圖4. 16，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4. 17

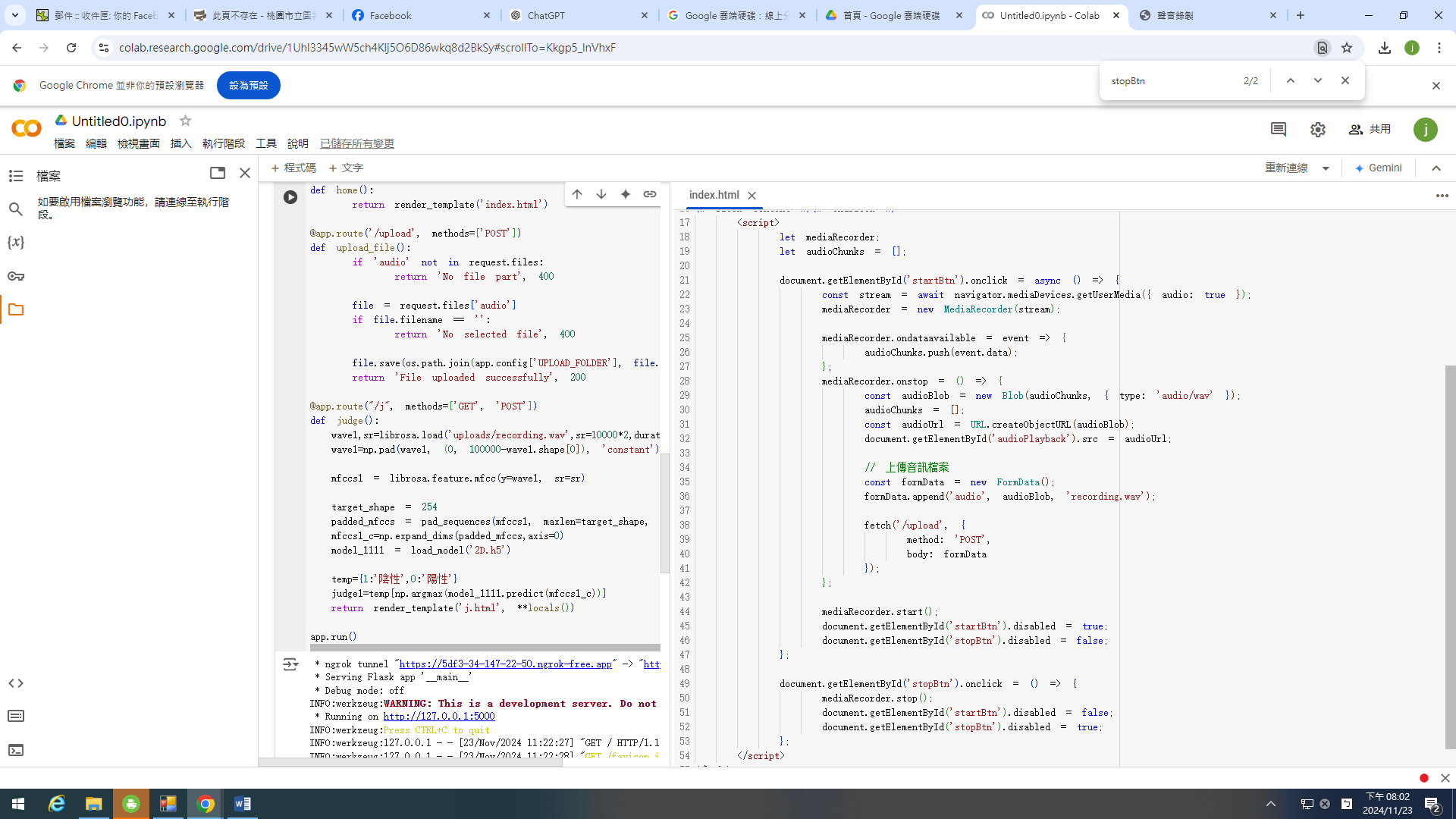


圖4. 15 錄製聲音的js程式碼

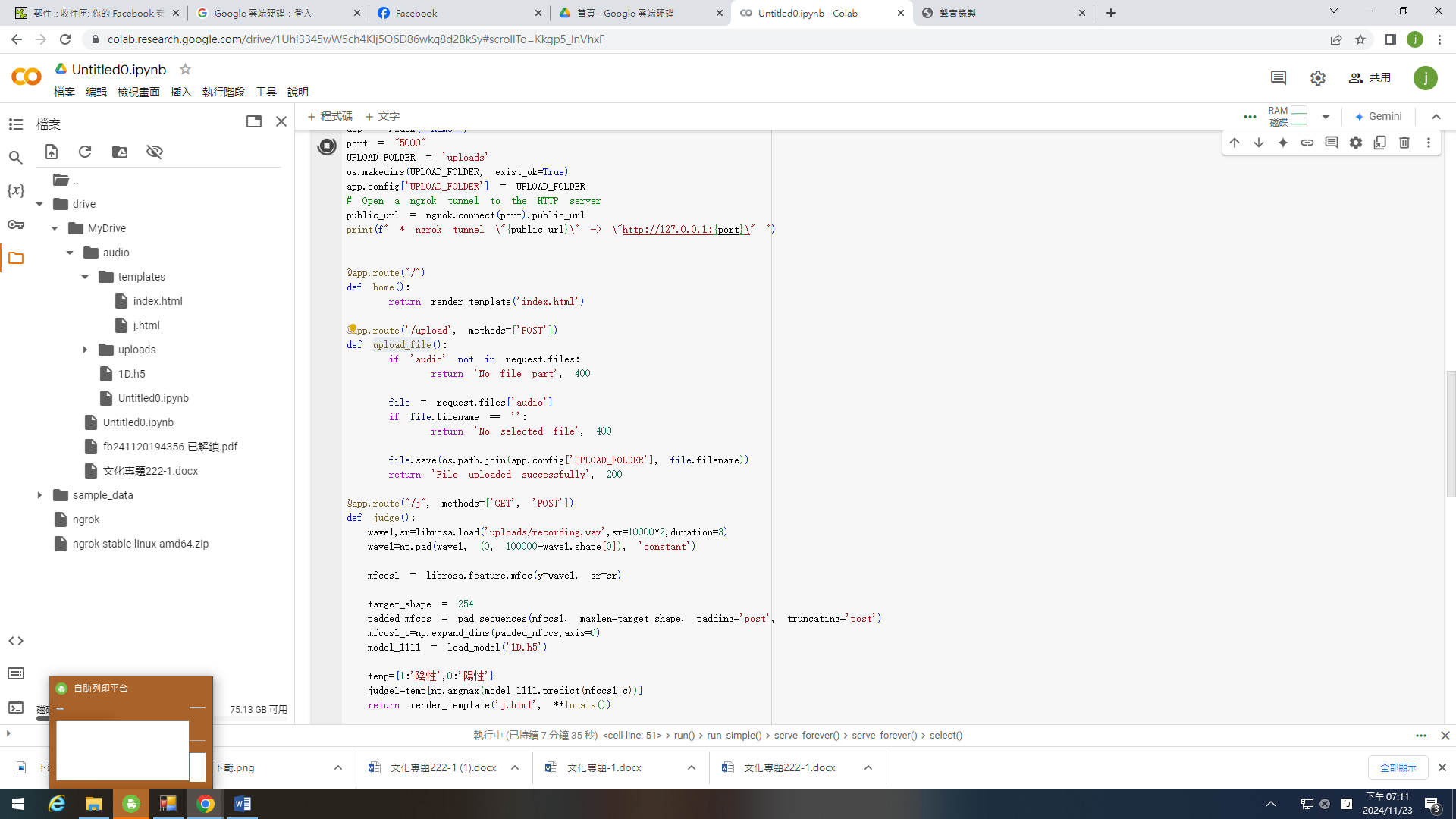


圖4. 16 upload\_file之程式碼

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4. 17前端示意圖

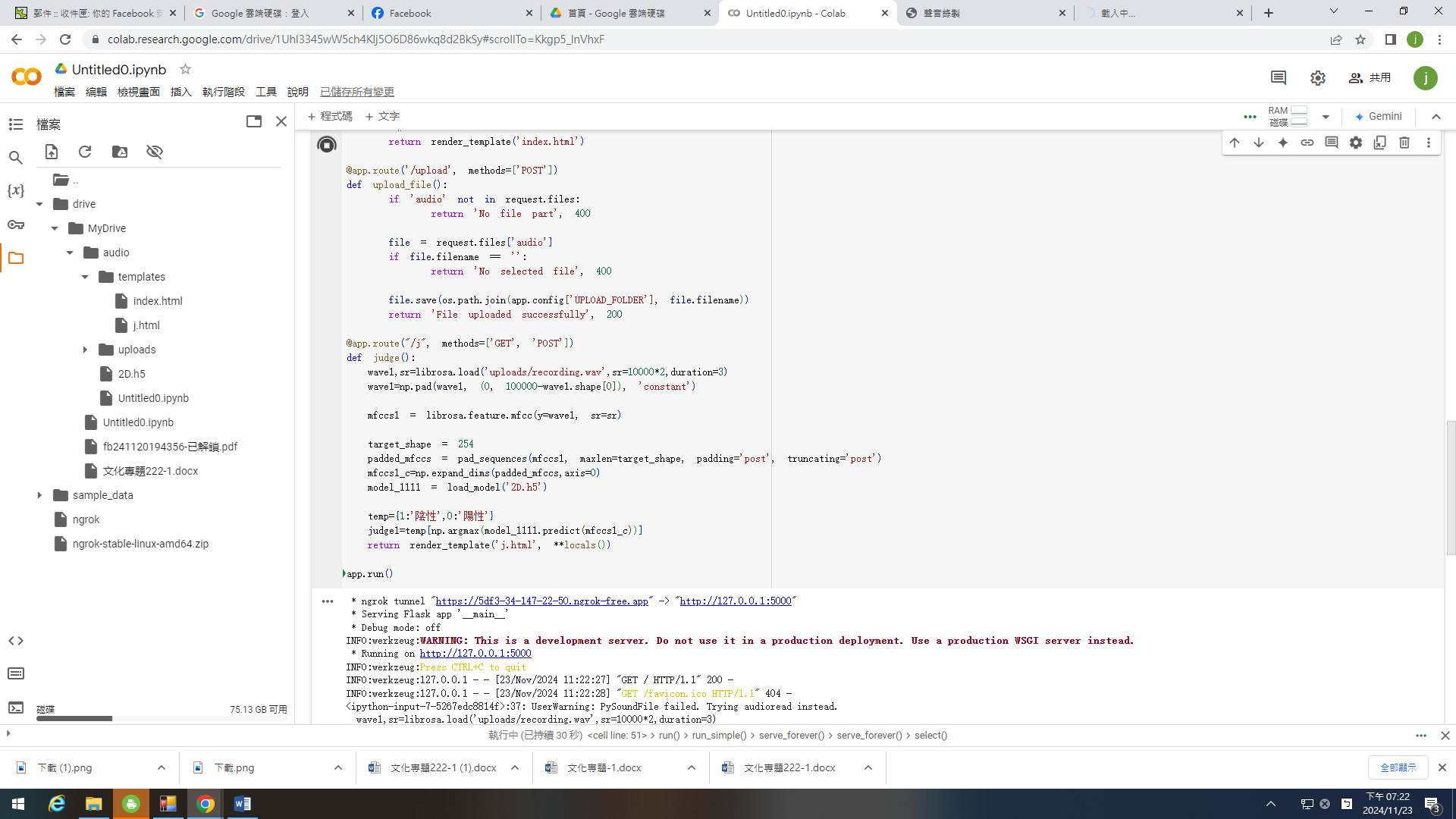


圖4. 18 judge之程式碼

由圖4. 18將預測好的類別存在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4. 19，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。

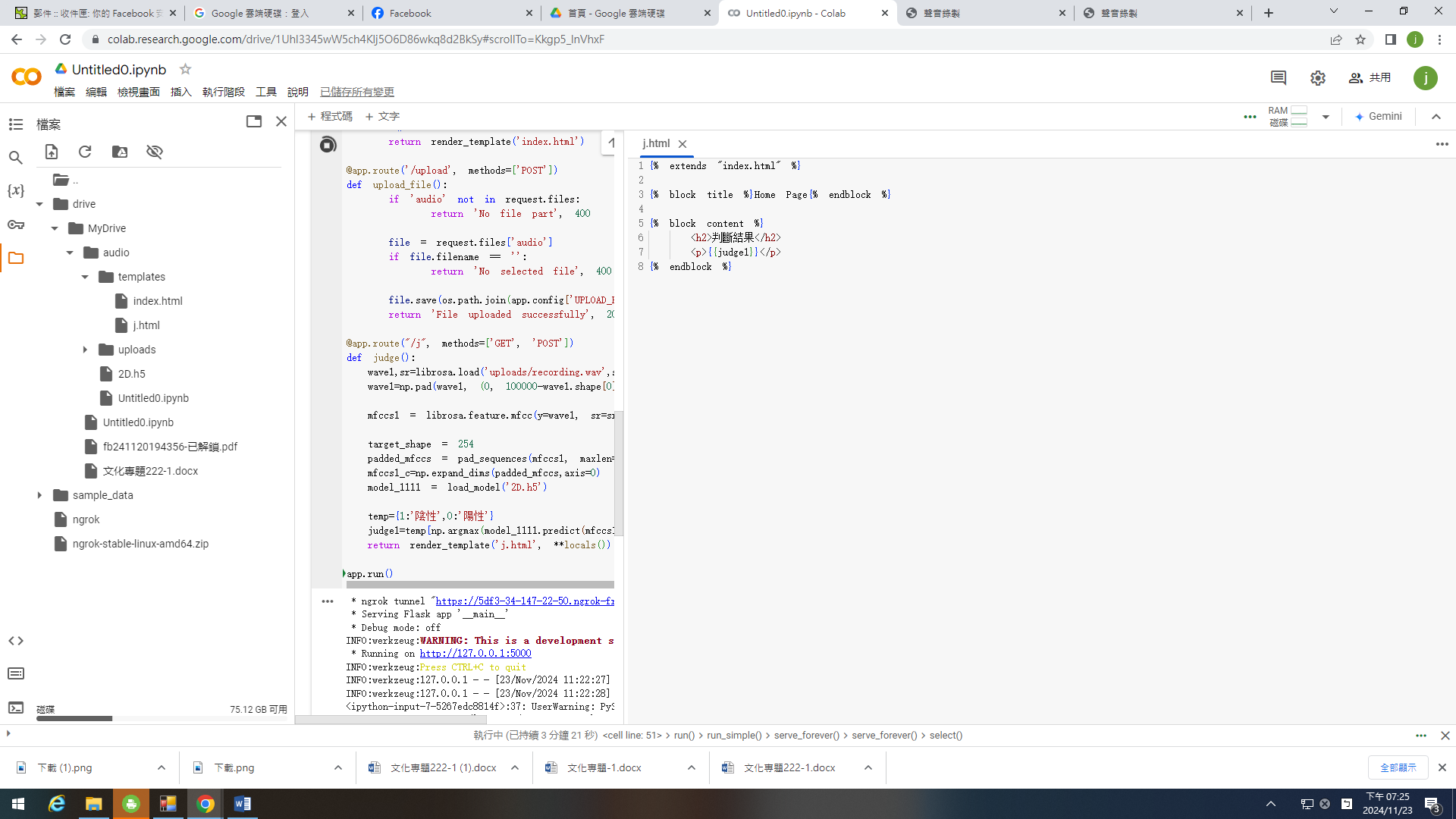


圖4. 19 輸出結果之模板程式碼

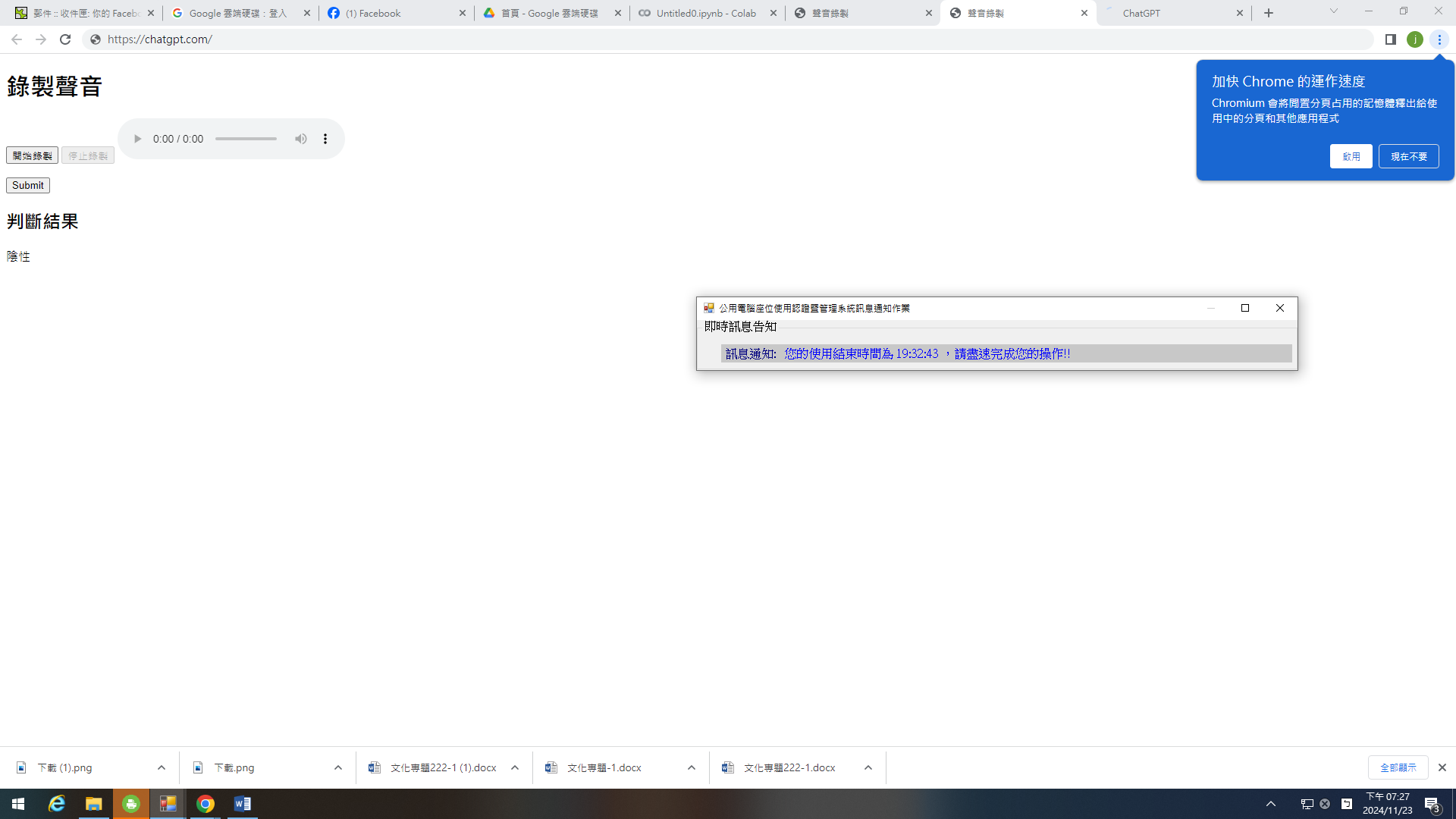


圖4. 20 最終顯示結果介面

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。

# 人力配置

**表 5.1工作人員工作分配圖**

|  |  |
| --- | --- |
| 姓名 | 工作內容 |
| 方建智 | 模型建立2d及優化 評估  模型建立1d及優化 評估  遷移學習 |
| 康宇翔 | 模型建立1d及優化 評估  輸出靜態網頁報表 |
| 曾敬峰 | 測試SVM  類神經網路特徵分析可視化 |
| 全體成員 | 蒐集與分析資料、製作報告、設計系統、功能實現、系統測試 |

# 第6章 未來展望：醫療聲音辨識處理系統

隨著技術的進步和醫療需求的增長，醫療聲音辨識處理系統在未來有著廣闊的發展前景。以下是對其未來展望的詳細分析：

(1) 提高系統精度

隨著更多醫療聲音數據的收集和精細標註，醫療聲音辨識系統的機器學習模型將持續優化，進而顯著提升系統的精度和穩定性。這些數據可以來自不同的醫療環境，如診所、醫院、遠程醫療平台等，涵蓋不同的病症和患者群體。特別是利用深度學習中的先進算法，如卷積神經網絡（CNN）和長短期記憶網絡（LSTM），能夠更有效地從複雜的聲音信號中提取特徵，提高疾病識別的準確率。此外，多模態數據融合技術（如結合聲音、影像和文本數據）將進一步增強系統的綜合判斷能力。

(2) 強化實時處理能力

為了滿足即時診斷和決策的需求，未來將引入高效的硬件加速技術，如使用GPU（圖形處理器）或專用的AI加速芯片，顯著提高系統的實時處理能力。這將使醫療聲音辨識系統能夠在診療過程中快速響應，提供即時的診斷建議和警示，特別是在急診室和手術室等需要快速決策的環境中。此外，邊緣計算技術的應用將使得系統能夠在設備端進行部分處理，減少延遲，提升患者體驗。

(3) 使用更便捷的套件

近期NVIDIA推出NeMo，這是一個開源的深度學習框架，專注於自然語言處理（NLP）和語音領域。NeMo提供了一個靈活、高效的開發環境，旨在幫助開發者快速構建和訓練語音識別、語音合成、語音轉文字、文字轉語音和自然語言理解等領域的模型。

其中有一個模組名為asr，參考<https://github.com/NVIDIA/NeMo/blob/main/tutorials/asr/Voice_Activity_Detection.ipynb>

這個網站，可以發現能夠直接將聲音直接不經過處理，只要輸入聲音檔案之位置，即可以進行模型訓練，未來可以嘗試利用該模組進行聲音預測。

# 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。