朝陽科技大學
資訊與通訊系

碩士論文

# 基於卷積神經網路的聲音分類機制

# CNN-based Sound Classification Mechanism

指導教授：朱鴻棋 博士

研究生：張永霖

中華民國 110 年 1 月

朝陽科技大學資訊與通訊系

Department of Information and Communication Engineering

Chaoyang University of Technology

碩士論文

Thesis for the Degree of Master

# 基於卷積神經網路的聲音分類機制

# CNN-based Sound Classification Mechanism

指導教授：朱鴻棋 (Hung-Chi Chu)

研究生：張永霖 (Young-Lin Zhang)

中華民國 110 年 1 月

January, 2021

# 摘要

日常生活環境中存在各式各樣的聲音，在物聯網日漸的蓬勃發展的現代，利用各種感測器收集到大量的資料(包括聲音訊號資料)。藉由網路通訊技術傳輸至伺服器並儲存，收集到的資料提供系統進行監控與資料分析。

藉由感測器蒐集到的聲音資料相較於其他的感測資料(如溫度、濕度)，收集到的聲音資料需經過訊號前處理轉換後找到相對應的特徵才能夠加以分析，因此處理上較為繁瑣。當我們進行訊號前處理後，假使藉由傳統的聲音訊號處理方式使用單一特徵做為分類依據(如頻率、振幅)，則因資料的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，本研究提出一個利用人工智慧中的卷積神經網路模型來建置聲音的分類機制。藉由梅爾倒頻譜係數原理將聲音訊號轉換成頻譜圖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。在本研究中提出梅爾倒頻譜係數中提取的三角濾波器組數的數量，做為資料量不足時進行資料擴增的依據。

在 ESC-50 公開資料集中，主要探討資料類別繁多且各類別內資料量過少，對於分類結果的影響，並藉由資料擴增的方法來敢善分類模型的準確率。實驗結果顯示所提出的方法，能夠提升分類的準確性 30%。

另外，在 UrbanSound8K 公開資料集中，主要是來探討所提方法在不同資料集的效能差異。實驗模擬是只提取一半的資料集資料進行訓練，並藉由資料擴增的方式來建立訓練模型。實驗結果顯示，聲音分類的準確率可以提升 3%。

**關鍵字:聲音、訊號處理、深度學習、卷積神經網路、分類**

# Abstract

There are all kinds of voices in everyday environments, and in the modern world of the Internet of Things, where the Internet of Things is booming, a large amount of data (including sound signal data) is collected using a variety of sensors. The data collected is transmitted to the server and stored by network communication technology for monitoring and data analysis.

Compared with other sensing data (e.g. temperature and humidity), the sound data collected by sensors need to be converted by signal pre-processing to find corresponding characteristics to analyze, so the processing is more cumbersome. When we do signal pre-processing, if a single feature is used as the basis for classification (e.g. frequency and amplitude) by traditional sound signal processing, the classification accuracy can not be improved effectively because the characteristics of the data cannot be found.

Therefore, this study proposes a classification mechanism that uses the courger neural network model in artificial intelligence to construct sound. The sound signal is converted into a spectrum map into a training model by the Mel Inverted Spectrum Coefficient principle for training. Among them, the courge neural network overcomes the problem that traditional sound recognition can only be classified by a single feature. Mel's inverted spectrum coefficient characterizes digital sound signals by differences in frequency. In this study, the number of triangular filter groups extracted from Mel's inverted spectrum coefficient is proposed as the basis for data amplification when the data quantity is insufficient.

In the ESC-50 open data set, the paper mainly discusses the influence of the large number of data categories and the small amount of information in each

category on the classification results, and dares to improve the accuracy of the classification model by means of data amplification. The experimental results show that the proposed method can improve the accuracy of classification by 30%.

In addition, in the UrbanSound8K public information set, mainly to explore the proposed method in different data sets of performance differences. Experimental simulation is to extract only half of the data set data for training, and by means of data amplification to build a training model. The experimental results show that the accuracy of sound classification can be improved by 3%.

**Keywords: Sound, Signal Processing, Deep Learning, CNN, Classification**

# 致謝

時光匆匆在研究所的生涯即將告一段落，在這段期間透過許多人的幫助才能夠順利的完成研究。首先最感謝的是指導教授朱鴻棋教授，不管在研究或任何方面上都給予最大的支持，並且在研究的過程中不斷的教導與勉勵，幫助我能順利解決許多難題，最終完成這本論文的撰寫。

在口試期間內，承蒙張林煌教授、李宗翰教授、廖俊鑑教授能夠在百忙之中撥冗指導，並且在於論文口試上給予了許多珍貴的意見和教導，使我學會從不同的角度去觀看研究和思考，讓此碩士論文的內容更加得完善，在此致上最深的謝意。並且也同時感謝線上公開的聲音資料分享，使得此研究能夠順利完成。

另外也感謝實驗室的學長、同學及學弟們，在碩士期間的大力相挺，也在你們身上學到了許多，並且從中學習彌補自己不足的部分。也感謝系上的幫忙，在碩士期間幫助我解決了許多問題。

最後要感謝我的家人，能夠在我研究所的期間全力支持，並在我需要時給予關懷與鼓勵，讓我能在求學期間無憂無慮的完成學業。由衷的感謝一路上陪伴我成長的每一個人，謝謝你們。

# 目錄

# 表目錄

# 圖目錄

# 第一章、緒論

## 1.1. 前言

聲音是大自然的傳遞訊息資料的其中一種方式，產生的方式源自於振動所發出的聲波，透過不同的物質狀態(如固態、液態、氣態…等)，進而傳播到人或者動物的聽覺器官上。聲音不但可以提供人與人之間基本的溝通外，大自然的事物也都有屬於自己獨特的聲音並讓人容易分辨此為何物。

所謂「聲音訊號」(Audio Signals)簡稱「音訊」，泛指由人耳聽到的各種聲音的訊號。一般來說，發出聲音的物體會產生振動，振動會對空氣造成壓縮與伸張的狀態，形成聲波，以每秒約 340 公尺的速度在空氣中傳播。當聲波傳遞到人的耳朵中時，耳膜會感覺到壓力，內耳神經再將此訊號傳遞到大腦，由大腦理解與判斷，並理解此訊號的意義。

當我們所收集到的聲音訊號進行應用時，藉由訊號前處理的方式，將訊號提取有效的特徵(如音色、音頻、音量)，提取到的相關特徵進行類別分類。常見的應用如藉由音量來源定位、歌唱分類、樂器分類…等。而在傳統聲音訊號應用，使用單一的聲音特徵，並藉由此特徵的數值差異設定門檻值做為聲音類別分類的依據，因為資料類別的多樣，

無法藉由門檻值仔細分類出各類別之間的差異並做為分類依據，導致無法分類出有效的結果甚至有誤判的情形。因此需對於聲音訊號的前處理進行更深入的特徵提取，如找到能夠更準確分類的門檻值設定，或提取符合各類別的特徵以提供分類模型進行類別分類。

本文提出基於人工智慧的聲音分類機制，藉由聲音訊號的前處理，並將聲音訊號轉換為頻譜圖。並以頻譜圖做為輸入資料，經由訓練模型的特徵學習後，以發展出可分類聲音的分類器，由於資料量的龐大，利用人工智慧的訓練模型可更快更有效率的訓練結果。

## 1.2. 研究動機與目的

在物聯網(Internet of Things, IoT)日漸的發展，對於聲音資料的收集也越來越方便快速，利用感測器將聲音資料收集完成後，藉由網路通訊技術傳輸到伺服器進行儲存，對於儲存的資料進行分析探討。

利用聲音特徵進行分類，假使以傳統的聲音訊號處理方式進行分類，使用單一聲音特徵進行分類，常因為特徵不足導致分類的準確率不佳，因此我們需要更多樣的特徵做為分類的依據。由於聲音的性質，環境聲音分類比起語音、音樂…等，處理上更為複雜，因為沒有像是音色、旋律與節奏較有意義的特性做為分類依據，由於龐大的資料量特徵，在處理上既費時又不易找到相對應的特徵，導致的分類結果準確率不高，甚至聲音資料過於相似而造成誤判。因此，聲音分類是研

究人員一直以來極力想解決的問題。故在整合資通訊技術、人工智慧(Artificial Intelligence, AI)技術，並藉由聲音特徵的提取，根據這些特徵運用聲音分類技術研發出特定聲音的分類機制，以提供一個可變的特徵提取方式提供在聲音分類應用上。

在進行聲音資料分類時，資料集的數量是對於分類有著很大的影響。當資料集內的資料數量過少時，可能會造成資料在分類時找不到特徵，導致分類的效果不好，甚至發生學習模型過擬合(Overfitting)的問題。

因此本研究將提出一個利用 CNN 模型[1]進行環境聲音的分類機制，將聲音訊號進行梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCCs)方法後所產生的頻譜圖，匯入到 CNN 模型裡進行模型訓練。不但保有整張完整時間區間的頻譜圖做為訓練模型之輸入，且克服傳統聲音分類上選用單一特徵的門檻值問題，讓特徵資料更多樣。將輸出的結果進行準確性的驗證，達到分類的能力。

## 1.3. 章節介紹

本論文分為五個章節，各章節簡介如下：

第二章文獻探討：描述物聯網趨勢與聲音訊號的相關應用、使用機器學習與深度學習的方法的比較，並說明本論文所使用的卷積神經網路學習架構。

第三章研究方法：說明本研究所使用的相關資料、實驗架構和流程。

第四章實驗結果：證明本研究方法和各個資料集的分類與分析，並提供聲音分類有更好的分類結果。

第五章結論：講述本研究之結論與未來可行方向。

# 第二章、文獻探討

## 2.1. 物聯網

物聯網[2][3][4]的概念最早在 1995 年比爾蓋茲的「未來之路」一書中被提到。直到 2005 年國際電信聯盟(International Telecommunication Union，ITU)發佈了「ITU 網際網路報告 2005：物聯網」,「物聯網」的概念才被廣為運用與發展。後來在 2009 年 IBM 首席執行官彭明盛提出「智慧地球」的概念，並建議美國政府投資新一代智慧型基礎設施，從此之後物聯網成為產業、軍事機關、學界…等重要的研究與發展重點，積極佈局在物聯網的各項技術與應用。在物聯網的架構中所分為的三層架構中(感知層、網路層、應用層)，如圖 1 所示。其中感知層是做為物聯網的實體層，主要是利用各種感測器(Sensor)、讀取器(Reader)將擷取到的資料。其目的是為了各種資料能夠自動感知、處理與收集的重要技術，現今常代表 Reader 的技術有無線射頻辨識(Radio Frequency IDentification，RFID)。網路層為在物聯網的負責傳輸的架構，主要負責通訊與資訊交換的功能，主要的任務是處理感知層所傳來的資料，其技術除包含現今網路(Internet)架構外，也包含無線感測網路與電信網路透過有線、無線的傳輸協定、TCP/IP 網際網路、VPN、無線通訊技術…等相關技術進行傳送，判斷

是要直接送往雲端或者直接採取適當的動作，是物聯網中不可或缺的技術。應用層則包括了各式與物聯網相關之應用系統，舉凡居家照護系統、智慧校園、智慧城市應用在交通系統、智慧生活應用在遠距離醫療醫療…等。世界各國的產官學界均重點投入發展物聯網應用系統。

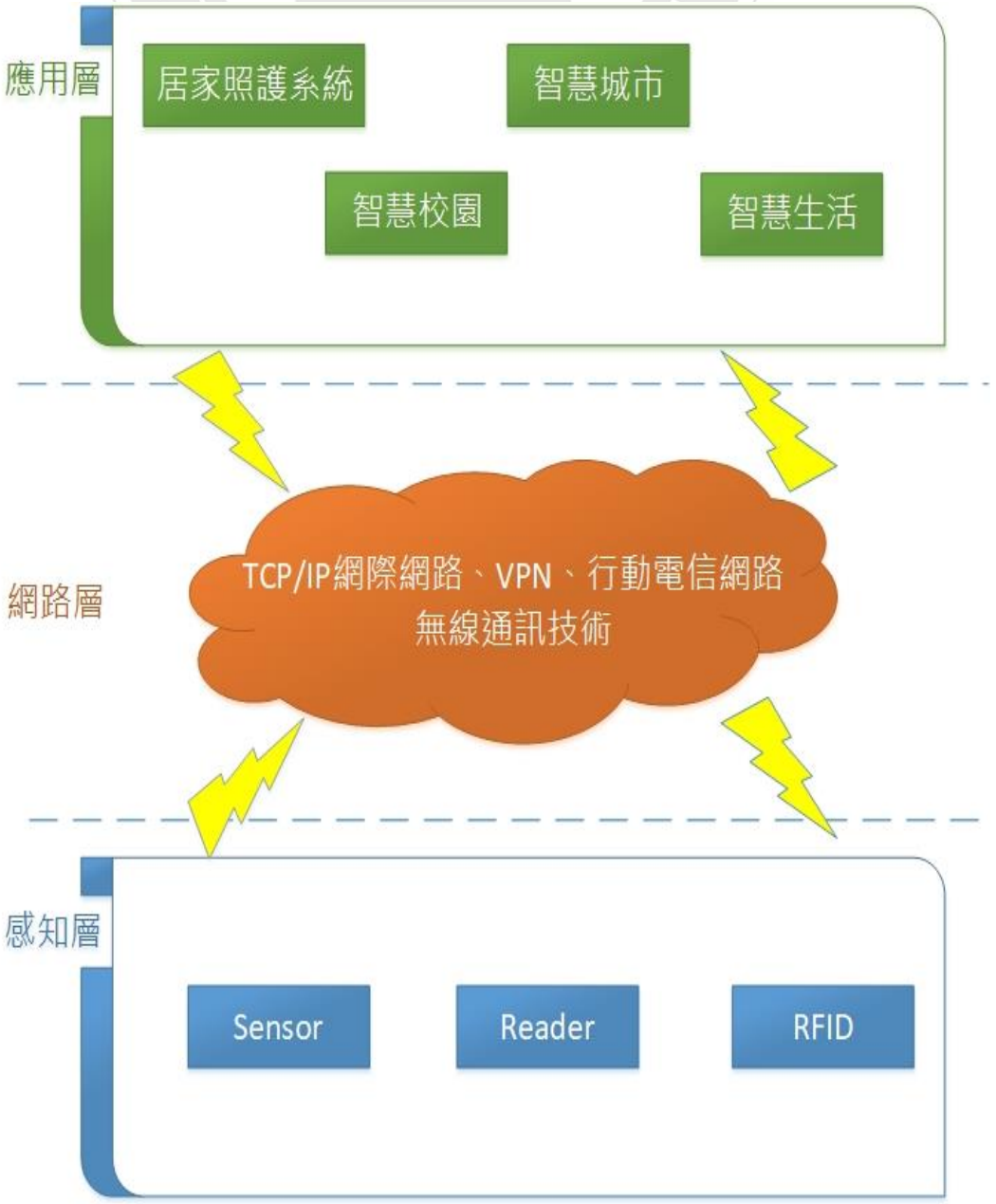



圖 1、物聯網架構

而物聯網應用系統能發展成功，除了大量資訊的收集與網路通訊的傳輸速率提升，提供更加快速的資料傳輸，進而將儲存的資料進行分析與應用。其中，聲音資料也是物聯網中可以收集到的資料，但相較於其他所收集到的資料，假使要藉由聲音資料進行應用，就必須經過資料前處理(Preprocessing)，才有辦法從聲音進行特徵提取做為分析與分類的依據。

假設要將此訊號分析或者儲存起來的話，就必須先將此訊號數位化，簡單來說就是類比訊號轉為數位訊號，將轉換完成的訊號，利用其特徵分析並應用。在醫療方面[5][6]，利用訊號分析的方法，對於人體的器官所發出的特殊聲音(如氣管的呼吸聲、吸氣與吐氣的聲音…等)，進而分析肺部構造，準確判斷病人症狀與防治，在日常生活[7]，老年人口的成長，很多時候都是老年人家自己在家裡，假使我們能夠利用環境聲音的聲音訊號進行識別判斷，根據環境的聲音，進行一些分類分析，假使當某項環境聲音資料與平時不一樣時，就可以即時判斷出來，進而確認老年人家在家裡是否有危險性。

## 2.2. 音頻訊號處理

音頻訊號處理，又稱音訊處理、音樂訊號處理…等，主要用來調整音樂訊號的振幅、頻率、波形…等相關資訊，主要利用初步的數學計算，將聲音訊號升降頻以及窗函數(window function)，就可以提取

出各式的聲音訊號間的特徵，甚至透過一些訊號處理的技巧，可以從聲音訊號取得背後所代表的頻率高低，在進行本文的環境聲音分類時，需要理解聲音訊號的特性，才能夠更進一步的分析與應用，以下將逐一介紹在做音頻訊號前須完成的各項步驟[8]。

### 2.2.1. 聲音訊號的儲存檔案

首先第一步，就是將感測設備所蒐集到的資料，將其儲存起來，可以分成壓縮檔案和無壓縮檔案兩種方法，常見的無壓縮檔案格式如(.wav)檔，而常見壓縮檔案格式為(.mp3)檔。本實驗使用的資料集使用無壓縮的(.wav)檔案，而存取聲音檔時，主要有三個需要考慮因素：取樣頻率(Sample Rate)、取樣解析度(Bit Resolution)與聲道(單聲道或雙聲道)。

(1) 取樣頻率：每秒鐘所取得的聲音資料點數，以 Hertz (Hz)為單位。數值越高，聲音品質越好，但是資料量越大。一般而言，8 kHz 的取樣頻率屬於電話的音質、16 KHz 的取樣頻率屬於一般語音辨識所採用的音質、而 44.1 KHz 的取樣頻率屬於 CD 的音質，本實驗使用的是 CD 的高音質取樣頻率，目的是提取到更多的聲音特徵。

(2) 取樣解析度：聲音資料點所用的位元數，通常使用 8-bit 來表示其數值範圍為 0 ~ 255 或 -128 ~ 127。

(3) 聲道：一般只分單聲道(Mono)或雙聲道(Stereo)。以本實驗為例，以單聲道的聲音，取樣頻率是 44100Hz(44.1 kHz)，解析度是 2 Bytes 的參數來進行錄音一分鐘，所得到的檔案大小大約就是$60\ 秒 \times 44.1kHz \times 2Bytes = 5292KB$，約為 5 MB。

### 2.2.2. 聲音訊號的資訊

在這些聲音訊號處理中進行特徵處理時，以下三個特徵是最常考被使用的：

(1) 音量：從訊號的角度，音量代表的就是訊號的振幅，當振幅越大，所發出的聲音也就越大。一般藉由振幅分類都是依據振幅大小，以分貝的大小進行分類。如藉由振幅大小對於聲音訊號進行聲音位置來源的評估[9]。

(2) 音色：訊號的波形，在各個物體之間所產生的音色也都不相同，因此可以利用波形分析來判斷聲音的類別與來源。如將聲音歌聲進行分類[10]，根據高音、中音、低音進行，可將唱歌的聲音在有限的範圍內進行分類，利用音色特徵，流程包含將聲音進行轉換。利用進行色度歸一化統計(Chroma energy normalized statistics, CENS)，主要功能是將在視窗上進行統計，主要目的可以讓聲音節奏更加平滑、修正發音和音樂的局部偏差，適合用於聲音的匹配和聲音相似性的比對[11]。

(3) 音頻：訊號的頻率，我們稱之為音頻，也代表著音調的高低，當頻率越高音調也就越高，依據音頻轉換的方式對於環境聲音的分類已有不錯的效果[12]。

在本研究中，由於環境聲音的特徵的多樣，若評估聲音的振幅大小做為分類的依據，會發生資料振幅過於相近，且會因未振問題而導致無法單純藉由振幅大小對於環境聲音進行分類。而若評估音色做為分類的依據，則會因為環境聲音沒有音色、旋律與節奏的特徵，而無法有效進行聲音分類，因此最後本研究在特徵提取所選用的是音頻做為聲音訊號前處理的方式。

### 2.2.3. 聲音特徵提取

聲音代表了物體在空氣中隨著時間所變化差異的振動密度，基本上是一個連續的函數，假設要將此訊號儲分析或者儲存起來的話，就必須先將此訊號數位化[13]。

主要考慮訊號的特徵方面，主要分為時域（Time domain, TD）和頻域（Frequency domain, FD）的類型，並以此分析聲音訊號特徵的變化。一般是以傅立葉轉換(Fourier transform)做為轉換的方法，將聲音訊號以頻域的方式呈現。轉換後聲音的本質特性還在，只是將聲音訊號以「頻率」的觀點來呈現分析。傅立葉轉換是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，在數位

訊號處理領域上，透過傅立葉轉換可將資料從時域波形轉換到頻譜上，傅立葉轉換的如公式(1)：

$$F(\omega) = \mathcal{F}|f(t)| = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\ dt \qquad (1)$$

其中$f(t)$為輸入的聲音訊號，將$f(t)$進行積分，其中$\omega$為頻率，$t$代表時間，$e^{-i\omega t}$代表複變函式。

傅立葉轉換根據輸入的訊號類型，可分為如非週期性連續訊號傅立葉變換(Fourier Transform)、週期性連續訊號傅立葉級數(Fourier Series)、非週期性離散訊號離散時域傅立葉變換(Discrete Time Fourier Transform)、週期性離散訊號離散傅立葉變換(Discrete Fourier Transform)，如圖 2 所示。

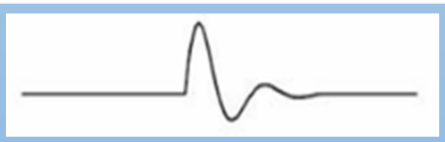
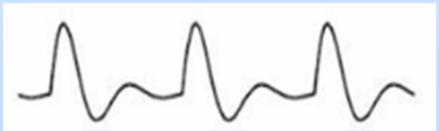
| 轉換方法 | 範例圖型 |
| --- | --- |
| 非週期性連續訊號傅立葉變換 (Fourier Transform) |  |
| 週期性連續訊號傅立葉級數 (Fourier Series) |  |
| 非週期性離散訊號離散時域傅立葉變換(Discrete Time Fourier Transform) |  |
| 週期性離散訊號離散傅立葉變換 (Discrete Fourier Transform) |  |

圖 2、傅立葉轉換變化型

當在分析聲音時，通常以「短時距分析」（Short-term Analysis）為主，因為音訊在短時間內是相對穩定的。因此聲音先切成音框（Frame），每個音框長度大約在 20 ms 左右，再根據音框內的訊號來進行分析。在一個特定音框內，可以發現主要是以音量、音高、音色做為聲音分類的特徵，在選定某特徵做為聲音分類的依據之前，我們必須進行特徵提取。

在聲音特徵提取時，主要的基本提取步驟如下[14]：

(1) 音框：將聲音切成數個音框，其中音框長度約為 20 ms ~ 30ms 之間，若音框太大，較無法找到隨時間變化的聲音特性。反之，若太小的話就無法找到聲音本身的特性，因此音框的用途主要是能夠切割成數個聲音的基本週期，各音框相鄰之間，可以保有重疊，藉此保有資料的連續性，一般而言重疊的部分約為音框長度的 1/2 到 2/3 不等。

(2) 特徵提取：音框化後，將對各個音框進行特徵提取，除了常見藉由聲音基本資訊，如音量、音高的方法，也有如過零率 (zero-crossing rate, ZCR)、MFCCs 進行特徵提取參數…等方法做為特徵提取的方式。

(3) 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。

## 2.3. 梅爾倒頻譜係數

聲音訊號的特徵代表著被轉換為不同的特徵向量係數。這些功能僅包括識別接收到的聲音所需的特徵。從聲音訊號中提取的特徵必須滿足某些標準，其中最重要的標準是：提取的特徵應易於測量，提取的特徵應與時間保持一致，最後，提取的特徵必須將有辨識性的特徵提取出來，將其他不重要的特徵(如背景的噪音，情緒…等)進行過濾，才能夠利用有效的特徵進行分類。

根據前小節提到聲音特徵提取的步驟，在特徵提取步驟時利用MFCCs[15]，其主要是依據人的聽覺做為基礎來進行實驗分析的方法。人耳能夠根據不同聲音的特徵來達到更準確的辨識能力。MFCCs 在1980 年由 Davis 和 Mermelstein 共同所提出的[16]，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[17]。

聲音訊號在提取特徵參數之前，為了使訊號能夠得到正確聲音訊號的特性，因此會先將訊號做前處理(Preprocess)，其流程包含預強調(Pre-Emphasis)、音框化(Frame Blocking)、漢明窗(Hamming Window)、接著是提取特徵參數(Feature Extraction)，最後才進行離散餘弦轉換(Discrete cosine transform, DCT)。針對各項程序的說明如下：

(1) 預強調(Pre-Emphasis)：當聲音收集完成之後，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)：

$$y(t) = x(t) - \alpha \times x(t-1)\,, 0.9 \leq \alpha \leq 1.0 \quad (2)$$

公式(1)為時域進行高通濾波器處理的方式，其中$x(t)$為原先的波型訊號$\alpha x(t-1)$為經過高通濾波器訊號，通常$\alpha$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，而本研究中，為了符合人耳對於聲音的敏感度，忽略了高頻的強調，保留低頻的聲音更多的特徵。因此後續實驗中並沒有使用此步驟。

(2) 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，以本研究為例，取樣頻率 44.1kHz 的聲音資料，因此取樣點為 1103 個，如公式(3)：

$$N = 0.0025 \times 44100 = 1102.5 \cong 1103 \quad (3)$$

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，本研究重疊部分設置為 15ms，因重疊部分一共有 662 個取樣點。

(3) 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，如公式(4)：

$$W(n,\alpha) = (1-\alpha) - \alpha \times cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \quad (4)$$

$W(n,\alpha)$為漢明窗的運算式，$n$ 為第 n 個音框，不同的α 值所產生的漢明窗大小皆為不同，在漢明窗中一般$\alpha$值我們設定為 0.46，如圖 3 所示。

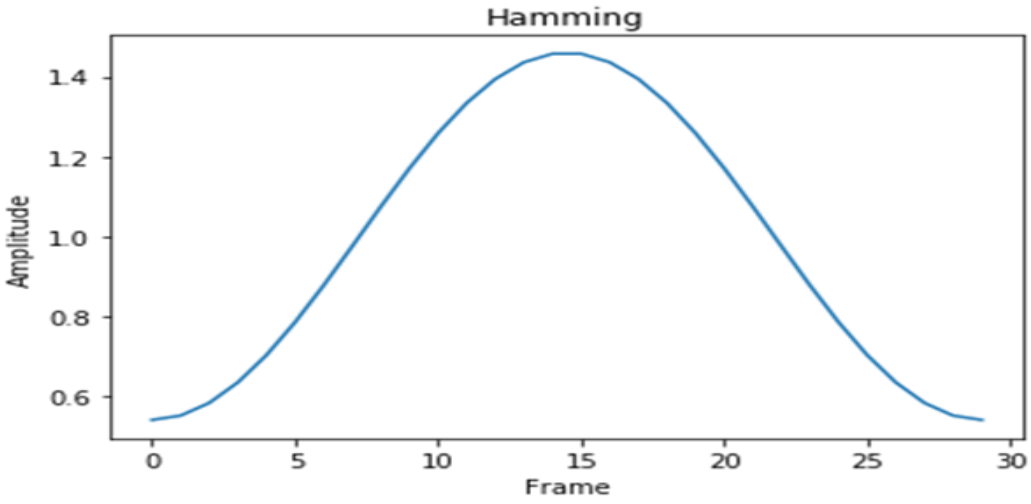



圖 3、漢明曲線$(\alpha = 0.64)$

(4) 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，因此處理較為繁瑣，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，轉換方法如公式(5)：

$$S_i(k) = \sum_{n-1}^{N} S_i(n)\, e^{-j2\pi kn/N} \quad k = 0, \ldots, N-1 \tag{5}$$

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，而梅爾頻率和一般頻率 $f$ 之間的轉換關係式如公式(6)：

$$\text{Mel}(f) = 2595 \times \log\left(1 + \frac{f}{700}\right) = 1125 \times \ln\left(1 + \frac{f}{700}\right) \tag{6}$$

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 4 所示。

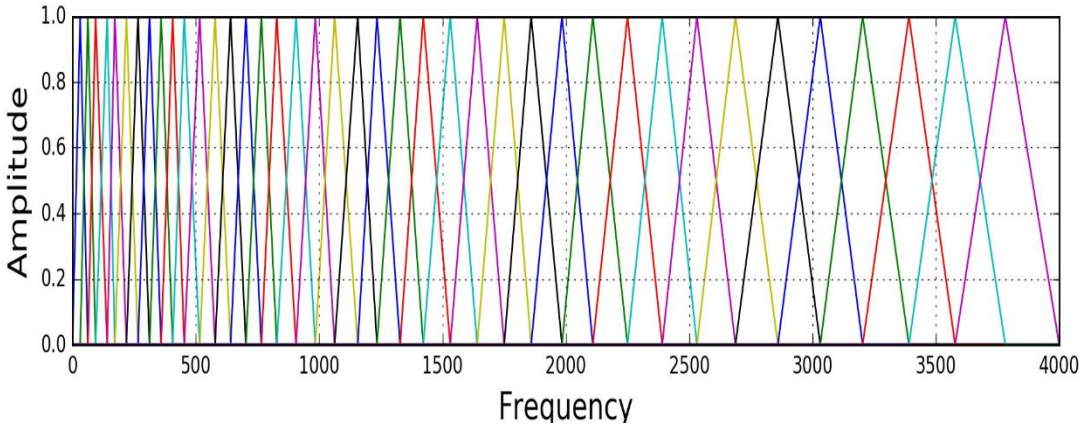



圖 4、三角濾波器

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。

(5) 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，如公式(7)：

$$c(n) = \sum_{m=0}^{m-1} logD(m) \times cos\left(\frac{\pi n(m-0.5)}{m}\right); n = 0,1, \dots, L \quad (7)$$

其中，*c(n)*表示的 MFCC 係數個數，*D(m)*是梅爾頻譜圖，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。

## 2.4. 深度學習(Deep Learning)

深度學習的框架是由 20 世紀 90 年代時類神經網路[18]的概念演進而來，然而因為當時的電腦運算效能低落、數據量不足，使其所顯現的效果不彰而導致類神經網路不被看好，甚至一度被認為是不可能被實現的方法。而在 2006 年有了不一樣的變化，主要是因為深度學習之父 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[19]。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…等產生十分深遠的影響[20]。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。現有的深度學習架構應用的領域十分廣泛有：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[21]。

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 5 所示。

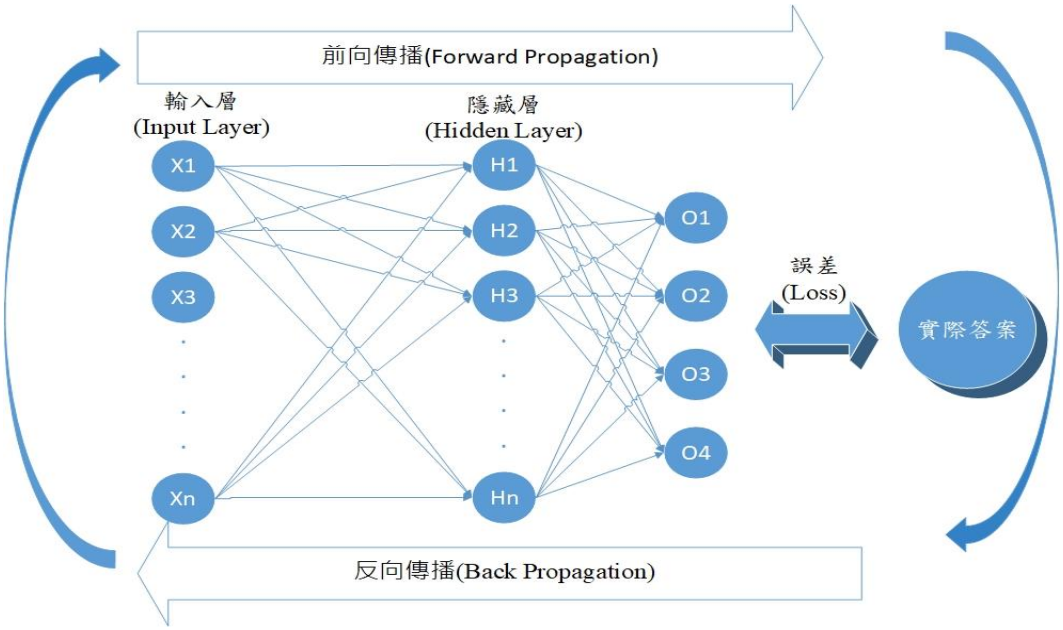



圖 5、神經網路收斂流程

(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 5 中以 $X_1$、$X_2$…$X_n$ 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…等。

(2) 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 $H_1$、$H_2$…$H_n$ 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。

(3) 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 $O_1$、$O_2$、$O_3$ 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

(Back Propagation)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[22]，以決策樹進行聲音分類[23]，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[25]，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[26]。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。

## 2.5. 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[27]，如圖 6 所示。

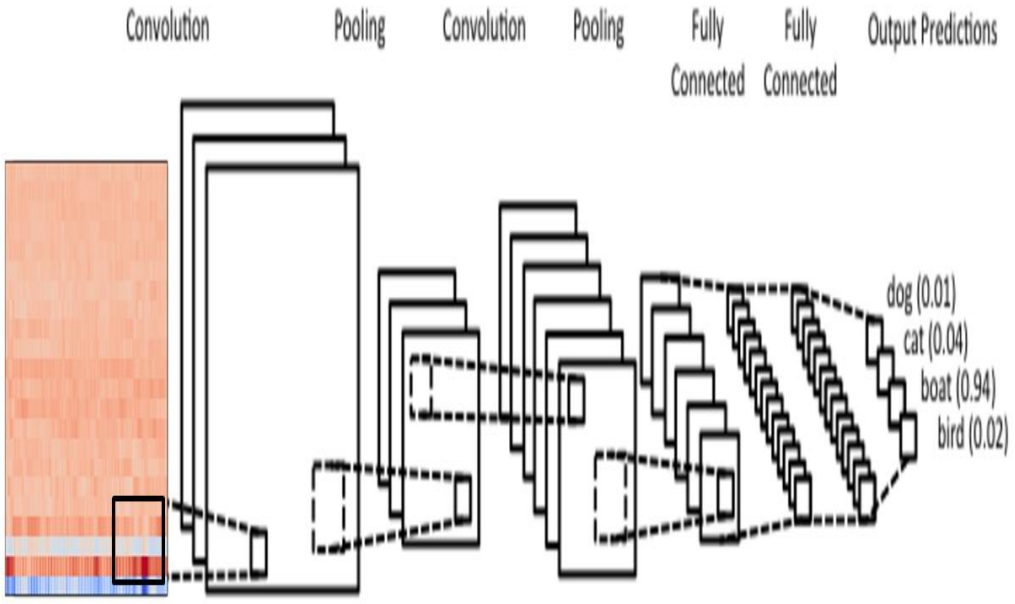



圖 6、CNN 模型架構

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。

### 2.5.1. 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 7 所示。

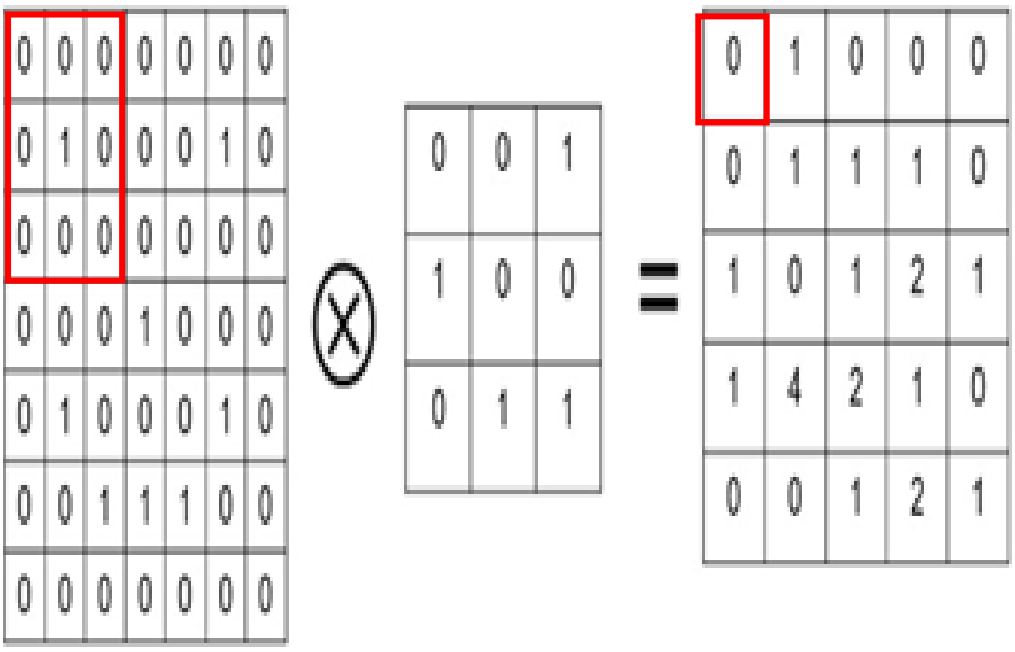



圖 7、卷積層運算

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，如圖 8 所示。

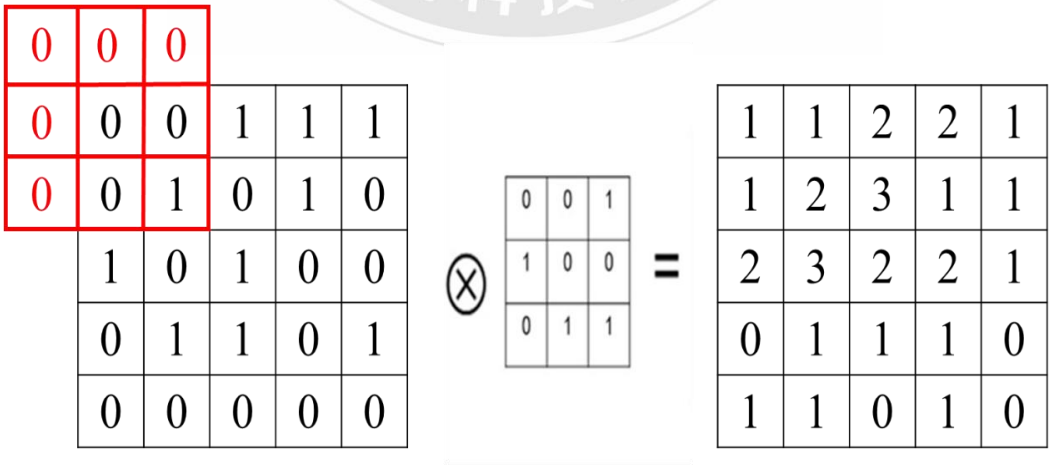



圖 8、卷積層的改善運算

提取出來的特徵有相當多樣，但因特徵的不同，分析出來的類別也就不同。因此選擇使用激活函數來提升特徵提取的能力，其作用是能夠給神經網路加入一些非線性因素，使得神經網路可以更好地解決較為複雜的問題，近幾年在深度學習領域中，激活函數通常都選用 ReLu (Rectified Linear Unit)函數，如 ReLu 在解決梯度消失的問題優於其它的激活函數，這是因為在反向傳播求導時不會因為範圍過大而全部歸零，造成梯度消失。

ReLu 只需判斷輸入值是否大於 0，大大減少了運算的成本，又因為部分神經元輸入小於 0，所以神經網路能變得較為稀疏，能緩解過擬合(Overfitting)的問題。因此本文選用 ReLu 做為聲音分類機制模型的激活函數，如圖 9 所示。

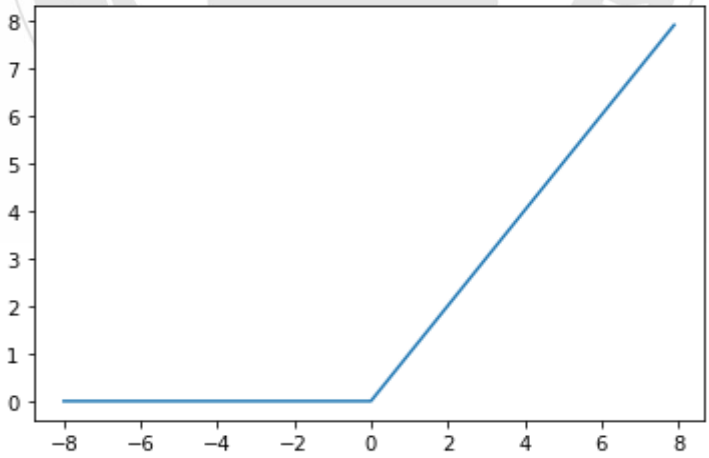



圖 9、ReLu 函式

### 2.5.2. 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。池化層之計算，如圖 10 所示。

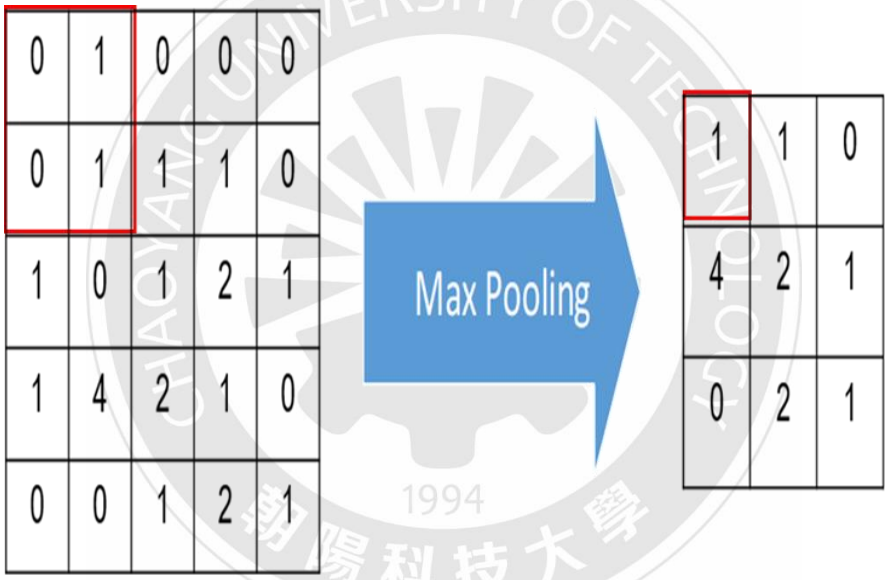



圖 10、池化層計算

### 2.5.3. 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 11 所示。

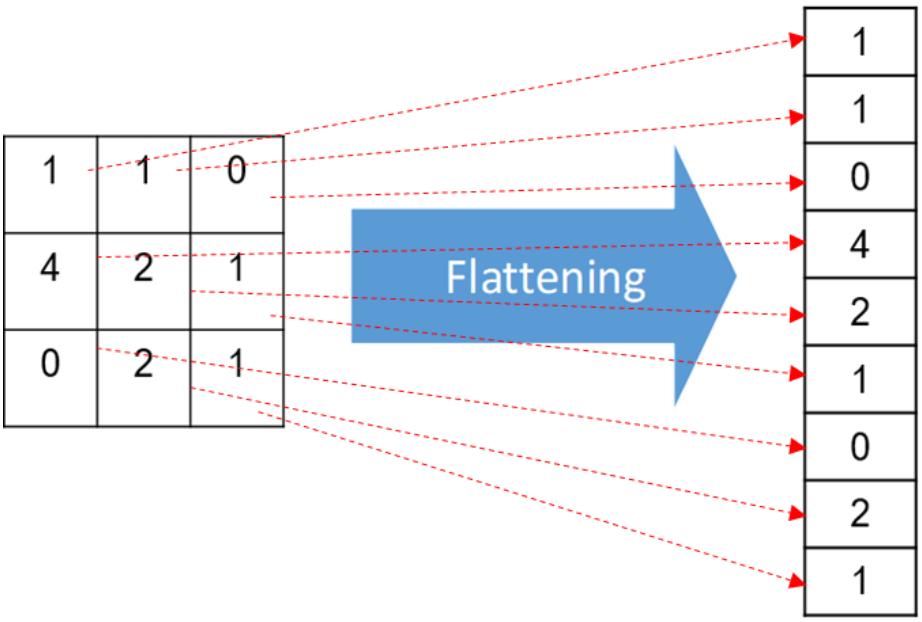



圖 11、全連接層

# 第三章、聲音分類機制流程

## 3.1. 研究流程

本研究的流程可分為六大步驟。分別是使用公開資料集做為模型輸入、並將聲音資料進行前處理(MFCCs)並以比例的方式切割資料集、建立分類模型、將聲音資料匯入模型進行訓練、在訓練過程中探討是否發生過擬合問題、最後評估分類效果，如圖 12 所示。以下章節將針對六大步驟進行介紹。

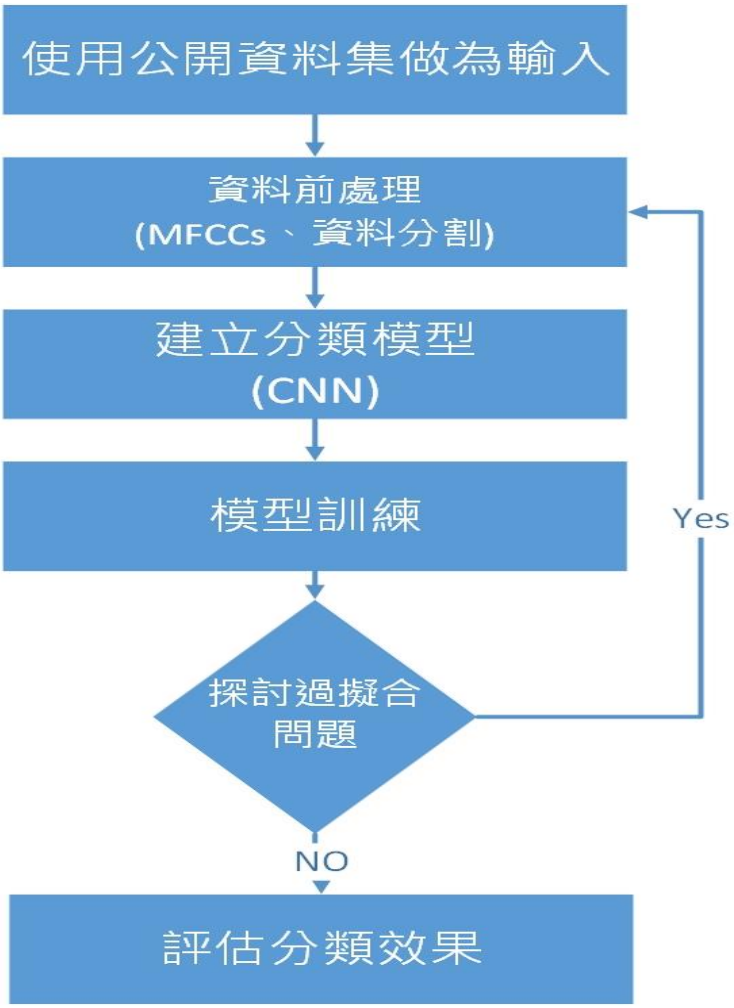



圖 12、聲音分類機制流程

## 3.2. 聲音資料來源

本研究基於考量實際聲音會因為取樣頻率的不同，以及環境的雜訊導致分類的效果不好，因此使用兩組不同的公開資料。將取樣頻率固定、經過整理較無雜訊的聲音資料做為分類資料，目的在於本研究主要探討聲音資料集數量對於分類的效果，不會因取樣頻率的不同、雜訊影響到分類的效果，對聲音資料的數量差異能夠更明確的比較。

### 3.2.1. ESC-50 公開資料集

本研究第一個數據集為 ESC-50 資料集[29]，該數據集由 2000 筆環境聲音所組成，取樣頻率為 44.1kHz，資料集一共分為五個類別，分別為：動物聲音、自然聲音、人類聲音（非語言）、室內聲音以及室外噪音，各類別皆為 400 筆資料，每筆聲音資料都被剪輯為一個 5 秒的聲音資料，此資料集的類別比例較為平均。

### 3.2.2. UrbanSound8K 公開資料集

本研究第二個數據集為 UrbanSound8 公開資料集[30]，此數據集為 8732 筆環境聲音資料所組成，取樣頻率為 44.1kHz，資料集一共分為十個類別，分別為：空調聲音 1000 筆、汽車喇叭 429 筆、兒童嬉鬧聲 1000 筆、狗叫聲 1000 筆、鑽孔聲 1000 筆、發動機運轉聲 1000 筆、槍擊聲 374 筆、手提電鑽 1000 筆、警鈴聲 929 筆、城市街道聲

1000 筆，總計 8732 筆資料，每筆聲音資料皆為 5 秒，相較於 ESC-50 資料集，各類別資料比較不平均。

## 3.3. 資料前處理

將公開資料進行進行前處理，目的是將聲音資料轉換成符合訓練模型所使用的輸入，藉由 MFCCs 進行轉換，產生出來的頻譜圖再以等比例的方式進行切割，分成訓練集以及驗證集以利於後續模型的訓練。

### 3.3.1. MFCCs

藉由 MFCCs 處理流程，逐一將聲音資料進行轉換。將聲音資料以 44.1kHz 的取樣頻率進行提取 5 秒單聲道的聲音，在本實驗我們沒有將聲音訊號進行預強調，因為使用的公開資料已經過篩選與處理，因此不需要再將聲音訊號預強調，保留更好的低頻訊號特徵。接著選定 1103 個音框，並乘上$\alpha = 0.46$ 的漢明窗，接著選擇三角濾波器的量，本文的重點主要就是探討不同的三角濾波器數量於聲音分類的影響(本文使用 40 個三角濾波器為預設)，最後將選擇的三角濾波器組進行離散餘弦轉換，將 MFCC 處理的矩陣(40*173)將轉換成頻譜圖。而此步驟將使用 Python 軟體內建的 LibROSA[31]來完成。轉換的頻譜圖如圖 13 所示。

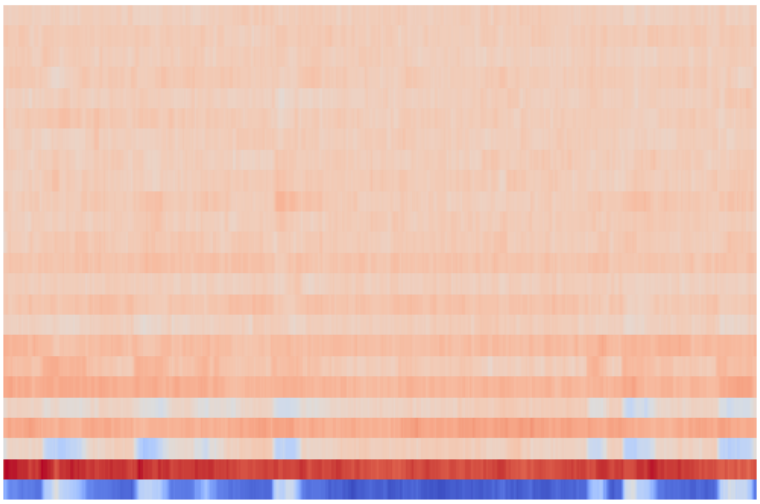


圖 13、頻譜圖範例

### 3.3.2. 資料集切割

將所有輸入資料轉換完的頻譜圖，以 9:1 的比例將頻譜圖分為訓練集與測試集以供模型的訓練及測試。

## 3.4. 建立分類模型

建立 CNN 模型：本篇論文所使用的模型，是基於[32]的動物聲音分類模型而設計改良的，如圖 14 所示。其中輸入層以(40*173)的圖片大小做為輸入層，一共使用了三層的隱藏層(包括卷積層、池化層)，其中各層卷積層的卷積核數量分別為 64、128、256。池化層使用的是最大池化層，使用的激活函數為 ReLu，為防止在模型訓練收斂時產生過擬合問題。我們將各隱藏層中都加入了 Dropout 來延緩過擬合，最後藉由輸出層將資料進行分類。

CNN 模型一共分為九個步驟：

Step(1)：第一步驟輸入圖片設定為 40* 173 pixels。

Step(2)：第一層卷積層的卷積核設定為 2 * 2，特徵圖為 64，以及激活函數為 ReLu。

Step(3)：此步驟為最大池化層，池化層大小設定為 2 * 2。

Step(4)：第二層卷積層的卷積核設定為 2 * 2，特徵圖為 128，以及激活函數為 ReLu。

Step(5)：此步驟為最大池化層，池化層大小設定為 2 * 2。

Step(6)：第三層卷積層的卷積核設定為 2 * 2，特徵圖為 256，以及激活函數為 ReLu。

Step(7)：此步驟為最大池化層，池化層大小設定為 2 * 2。

Step(8)：此步驟包含兩個平坦層，目的是將數據轉換為相應的輸出分類結果。

Step(9)：輸出聲音分類類別(類別)。

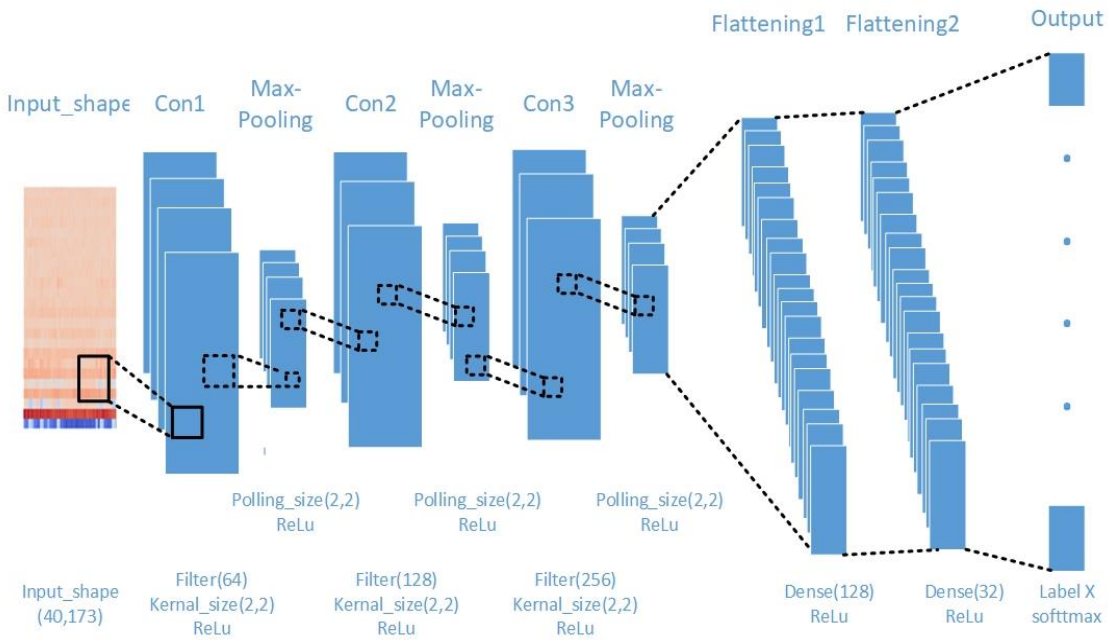



圖 14、本文使用的 CNN 模型架構

## 3.5. 訓練模型

以下將基於圖 14 的 CNN 模型進行聲音分類，說明模型的參數設定，以及探討聲音資料數量對於訓連模型的分類效果的影響導致過擬合問題，最後將提出如何解決此問題的方法。

### 3.5.1. 訓練參數

CNN 模型的超參數設定如表 1 所示。每一個 Epoch 都代表著整個模型訓練”一次”代表的結果，此模型訓練次數我們設定為 500 次 Epoch，但在訓練的過程中，為了防止資料過擬合[33]，我們加入了 Early Stopping 技術[34]，主要是探討當訓練集過於訓練時，訓練集分類結果會逐漸提升，分類結果會逐漸提升，模型也會進行收斂(Loss 值

下降)。但是在驗證集上卻未必與訓練集相同，相差甚遠時便會產生過擬合，因此藉由 Early Stopping 技術，以一個監督式學習的方式，本論文主要是監督訓練驗證集所收斂的狀況，藉由驗證集的 Loss 值，當訓練次數超過 過 50 次以上沒有收斂時，我們就會將訓練模型停止，達到防止過擬合的問題訓練集並藉由 categorical_crossentropy 來做為最後分類結果的交叉驗證方式，以及使用的優化器為 Adam，模型的學習率設定為 0.001。

表 1、模型超參數

| **Parameter** | |
|---|---|
| Epoch | 500 |
| Batch_Size | 4000 |
| Loss_Function | categorical_crossentropy |
| Optimizer | Adam |
| Learning Rate | 0.001 |

### 3.5.2. 探討過擬合問題

過擬合是指一種現象，它與特定數據集過於緊密或精確地匹配，從而無法很好地適應其他數據或預測未來的觀察結果。它通常在以下情況下發生：

(1) 數據集太少：如果重要特徵數據的數量太少，則在訓練模型時將無法正確分類此特徵。

(2) 模型過於複雜：模型過於複雜時，通常採用訓練模型來降低

維數，以簡化系統複雜度。此時，數據集的特徵可能會因過度降維而丟失。

(3) Dropout：當模型反覆提取特徵後，由於模型的複雜度會產生出個多的特徵，因此利用 Dropout 在許多的特徵中，丟棄不需要的特徵，將需要的特徵保留起來。

本研究將運用增加資料量以及 Dropout 來解決過擬合的問題，其中利用兩種梅爾係數轉換頻譜圖來擴增資料量，方法主要是在 Librosa 的 MFCCs 轉換中，對於特徵提取時，運用 MFCCs 中三角濾波器的個數，以一個 20 ~ 40 區間的範圍進行提取(一般預設提取為 40)[35]。而當資料量不足時，只提取單一的組數是無法準確地將聲音資料進行分類，因此我們將會以不同的提取組數做為資料擴增的方法，以解決資料量不足的問題，期望能藉此提升分類準確性的目的。

本研究主要提取整體區間中以相差 5 個三角濾波器的差異(40、35、30、25、20)為間距的數量做為資料前處理方式，擴增方式將以遞減的方式進行提取，第一組：40，第二組：40、30，第三組：40、30、20，第四組：40、35、30、20，第五組：40、35、30、25、20。

## 3.6. 評估模型效果

混淆矩陣是一種用於驗證分類效果的方法[36]，如表 2 所表示。

表 2、混淆矩陣

| True / Prediction | Event | No-event |
|---|---|---|
| **Event** | TP | FP |
| **No-event** | FN | TN |

表 2 的橫軸為實際的結果，縱軸為預測的結果，其中 TP(True Positive)是預測資料的結果與實際資料相同之情形，TN(True Negative)是預測資料的結果為非實際資料，且結果也非實際資料，FP(False Positive) 是預測資料的結果為實際資料，但結果卻非實際資料，FN(False Negative) 是預測資料的結果為非實際資料，但結果卻是實際資料。

依據混淆矩陣的結果，將分別探討準確率、精確率、召回率、F1-Score，以評估所提出的分類性能。

聲音分類的準確性(Accuracy)定義為公式(8)所示，其中分類模型中正確的分類情形與所有分類情形的比例。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

聲音分類的精確度(Precision)定義為公式(9)，在所有預測的結果中，預測正確結果的比例。

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

聲音分類的召回率(Recall)定義為公式(10)，在所有實際結果中，對於預測的比例。

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

聲音分類的 F1Score 定義為公式(11)，其中 P 代表精確度，R 代表 Recall，為兩種評估方式的綜合指標，其中數值範圍為 0~1 之間，越接近 1 表示分類結果越好。

$$\text{F1 Score} = \frac{2PR}{P + R} \tag{11}$$

藉由上述的聲音分類流程機制，我們能夠有完整的步驟來完成聲音分類的研究，並且能夠探討聲音資料的數量，對於練模型的準確性的影響，並藉由整體資料集的擴增以及部分提取進行擴增做為訓練，來評估分類結果。

# 第四章、實驗結果與分析

實驗結果主要透過探討兩個公開資料集藉由不同的三角濾波器組數之間的差異進行資料擴增分類，並由上述所提出的資料擴增方式，提供給不同的資料集擴增並提升分類的準確性，以證明此方式能夠因應不同資料集的分類準確性提升。

## 4.1. 基於 ESC-50 資料集

ESC-50 資料集，一共有 2000 筆資料，初步我們將聲音資料分為五大類別，但各類別中還分別有 10 個小類，我們主要是探討資料以主類別分類時的分類情形，以及藉由擴增後的分類情形，最後並以次類別的分類方式談討各類別的分類情形，並且夠藉由擴增方式，將次類別的資料，更準確的分類，來證明資料擴增的方法是否可行。

### 4.1.1. ESC-50 原始資料分類結果

將 2000 筆資料藉由 MFCCs 特徵提取，且三角濾波器數量固定為 40，並轉換為 CNN 所需的頻譜圖。將資料集分割為 9:1 的比例(訓練集：測試集)進行訓練。而當我們對於訓練集進行訓練時，為了確保能夠即時驗證分類的結果，我們將訓練集再次以 9:1(訓練集：驗證集)的方式進行分割，讓模型每完整訓練一次的時候能夠即時的驗證模型分類的效果，模型訓練結果如圖 15 所示：

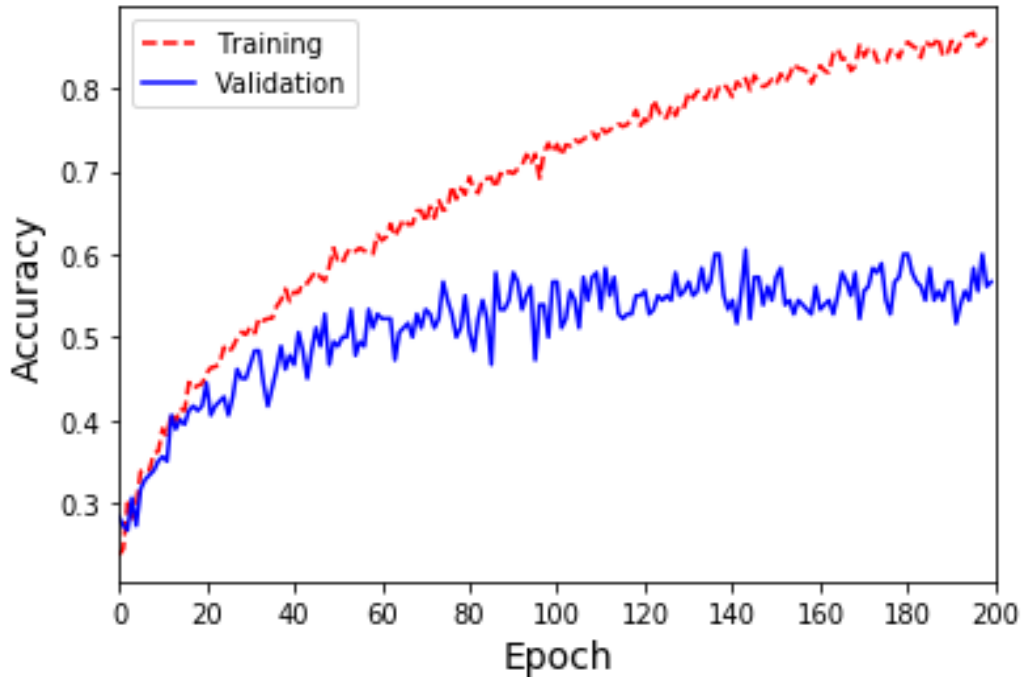



圖 15、ESC-50 原始資料訓練結果

在 500 次的訓練週期中，我們藉由 Early Stopping 方式預防過擬合問題，但我們發現到 ESC-50 原始資料集在訓練第 200 次週期的時候就已經停止，而訓練集與驗證集的結果相差非常的大，此結果可以發現此訓練模型已產生過擬合的問題，而驗證集的準確度也只有 60%。

表 3、ESC-50 各類別混淆矩陣結果

| 實際 / 驗證 | 動物 | 城市 | 人 | 室內 | 大自然 |
|---|---|---|---|---|---|
| 動物 | 23 | 5 | 2 | 3 | 2 |
| 城市 | 7 | 26 | 3 | 1 | 5 |
| 人 | 3 | 5 | 26 | 3 | 3 |
| 室內 | 2 | 4 | 8 | 23 | 6 |
| 大自然 | 0 | 6 | 6 | 2 | 25 |

從表 3 混淆矩陣中可以發現，測試集的分類結果，雖然能夠正確的將大部分資料分類，但有一定的比例產生了誤判的問題。由於測試集數量過少，因此整體分類的結果是不準確的。

表 4、ESC-50 原始資料各類別分類結果

| 評估指標 / 類別 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 動物 | 0.66 | 0.66 | 0.67 | 0.68 |
| 城市 | 0.62 | 0.56 | 0.62 | 0.59 |
| 人 | 0.65 | 0.58 | 0.63 | 0.60 |
| 室內 | 0.53 | 0.72 | 0.53 | 0.61 |
| 大自然 | 0.64 | 0.61 | 0.63 | 0.60 |
| 平均 | 0.62 | 0.63 | 0.62 | 0.62 |

表 4 可以觀察出，ESC-50 原始資料，整體的分類準確度約 60%，表示資料集的數量對於模型的訓練有著很大的影響，資料集過少可能導致模型找不到類別所對應的特徵，導致分類模型效果不好，進而影響分類結果。

### 4.1.2. ESC-50 原始資料擴增分類結果

在原始資料，資料過少導致的分類效果不佳的問題，本小節將基於 ESC-50 原始資料進行擴增，三角濾波器提取的組數我們選擇 40 以及 30 做為 MFCCs 特徵提取轉換的依據，選擇這兩個組數的原因是因為以單一組數分類的結果，這兩組結果為最好以及次好的，因此選擇這兩組三角濾波器組數做為資料擴增的依據，訓練的結果如圖 16 所示。

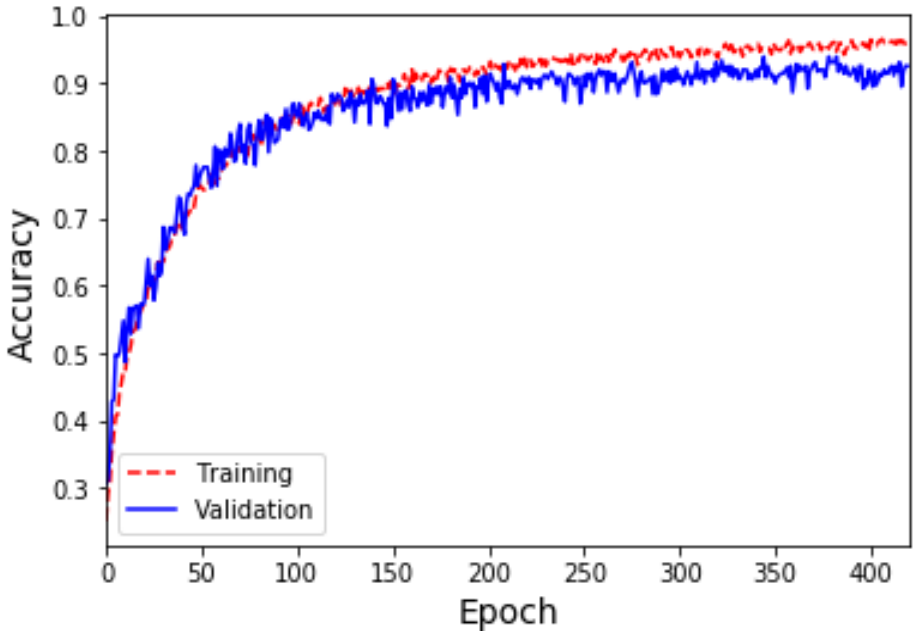



圖 16、ESC-50 資料擴增的訓練結果

從訓練結果可以發現，當資料擴增時，訓練的數量以擴增成長，整體模型提升了一倍的訓練量，在 500 次的訓練週期中，在 Early Stopping 的方式上，整理訓練次數在 420 次週期時停止，但整體驗證集與訓練集的結果是相當的接近，準確度也達到了接近 90%。

表 5、ESC-50 原始資料擴增後各類別混淆矩陣

| 實際 / 驗證 | 動物 | 城市 | 人 | 室內 | 大自然 |
|---|---|---|---|---|---|
| 動物 | 75 | 6 | 3 | 0 | 5 |
| 城市 | 0 | 65 | 3 | 2 | 4 |
| 人 | 3 | 2 | 81 | 1 | 1 |
| 室內 | 0 | 2 | 4 | 69 | 7 |
| 大自然 | 0 | 0 | 0 | 0 | 67 |

在表 5 中可以發現當資料集進行擴增時，各類別分類相較於原始資料的分類明顯提高，以及對於其他類別的誤判也相對的減低，能夠更準確的分類。

表 6、ESC-50 原始資料擴增後各類別分類結果

| 評估指標 / 類別 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 動物 | 0.84 | 0.96 | 0.84 | 0.90 |
| 城市 | 0.88 | 0.87 | 0.88 | 0.87 |
| 人 | 0.92 | 0.89 | 0.92 | 0.91 |
| 室內 | 0.84 | 0.96 | 0.84 | 0.90 |
| 大自然 | 1.00 | 0.80 | 1.00 | 0.89 |
| 平均 | 0.90 | 0.90 | 0.90 | 0.89 |

在表 6 中，整體分類平均可達到 90%，相較於原資料集，明顯提高了 30%，因此可以證明資料擴增時，模型的訓練量也跟著提高，對於聲音資料分類有著顯著的效果。

在本小節一共探討了基於 ESC-50 原始資料以及將原始資料擴增的情形，整體平均一共提升了 30%，證明隨著資料數量的擴增，對於資料量不足時是可行的，如表 7 所示。

表 7、基於 ESC-50 不同的分類結果

| 評估指標 / 方法 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 原始資料 | 0.68 | 0.63 | 0.62 | 0.62 |
| 擴增資料 | 0.90 | 0.90 | 0.90 | 0.89 |

### 4.1.3. ESC-50 原始資料各類別細分之情形

上述在探討 5 大類別時，由於主類別中的資料過於多樣，因此分類時，無法找到相對應的特徵導致分類不準確，而當資料量過少時也會影響分類的準確性。因此藉由資料擴增的方法可以發現，資料量的增加對於模型的訓練確實能夠達到顯著的效果，但是主類別的多樣性，藉由此方式雖然可以將主類別的分類準確度提升，但我們無法探討出次類別中獨有的特徵，無法證明此擴增方式是否有更好的優勢，因此在本章節，我們將 5 大主類別中，10 個次類別進行探討，並藉由擴增的方法，觀察出各類別的分類結果，以下我們將逐一探討各類別的分類結果：

(1) 動物：

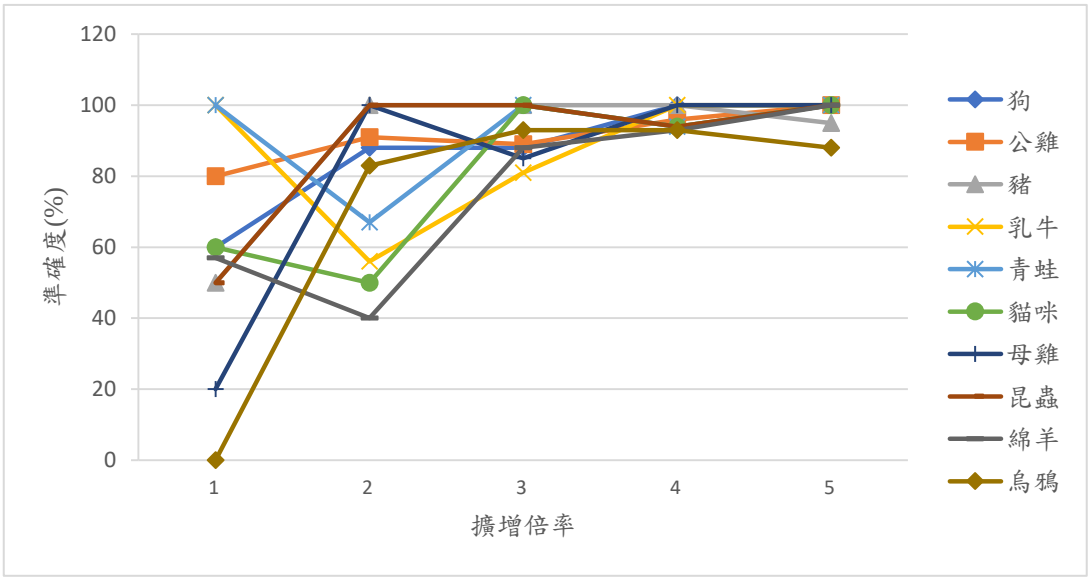



圖 17、動物類別 Accuracy

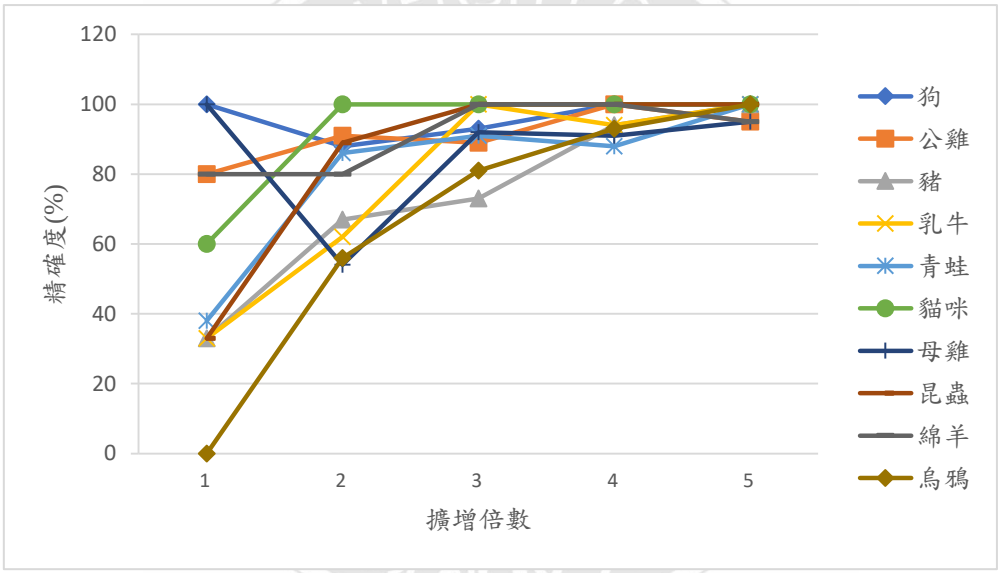



圖 18、動物類別 Precision

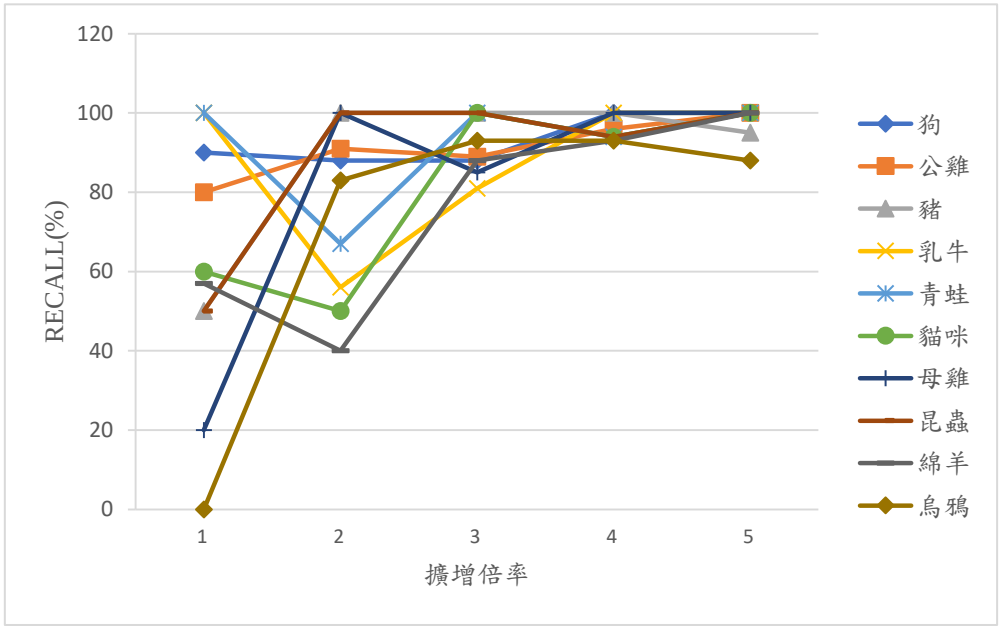



圖 19、動物類別 Recall

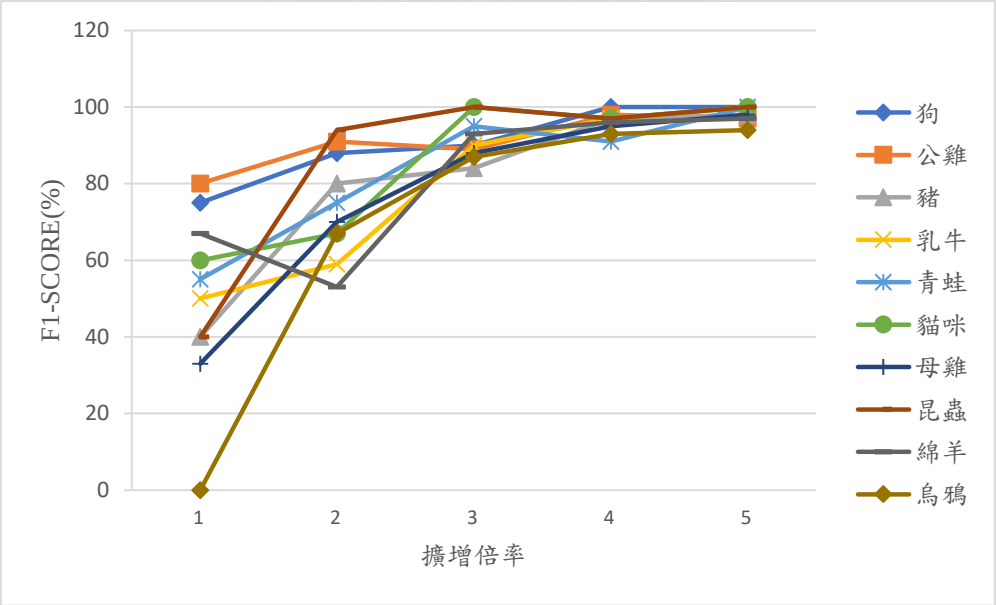



圖 20、動物類別 F1-Score

以上四個評估指標的圖中，橫軸為擴增的倍率，1 為只提取 40 個三角濾波器；2 為提取 40、30 個的三角濾波器做為擴增；3 為 40、30、20；4 為 40、 35、30、20；5 為 40、35、30、25、20。根據四個評估指標，我們可以發現，在動物類別時，在原始資料進行分類時，平均值都偏低，而準確率十分的分散，甚至有的在資料量少的時候已能夠準確的分類。而在藉由我們擴增機制進行資料量的提升後，有著明顯的提升，資料分類的準確率也逐漸的趨於平穩。而當在資料擴增達到 5 倍時分類效果已達到了不錯的水準。

(2) 人類：

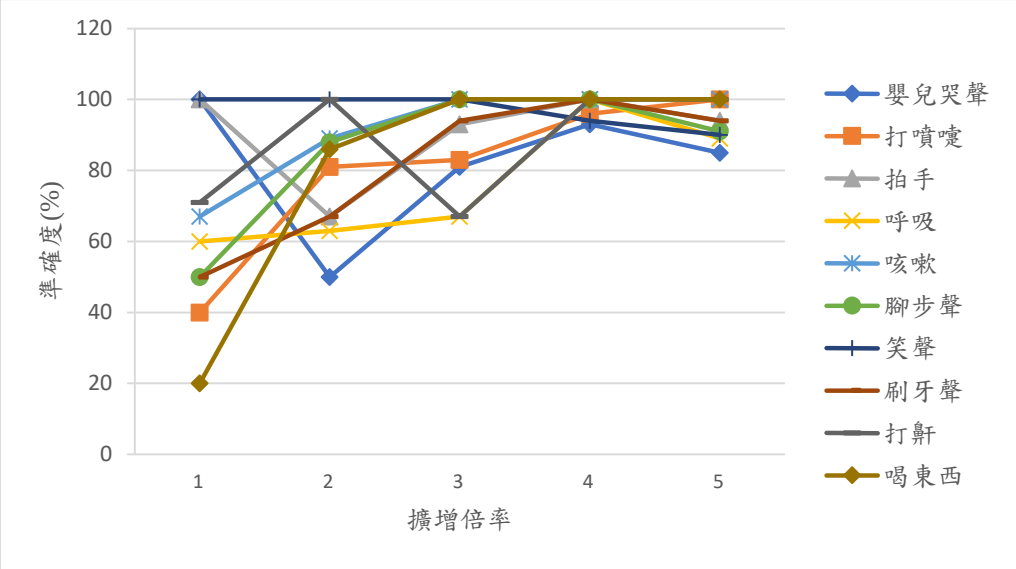



圖 21、人類類別 Accuracy

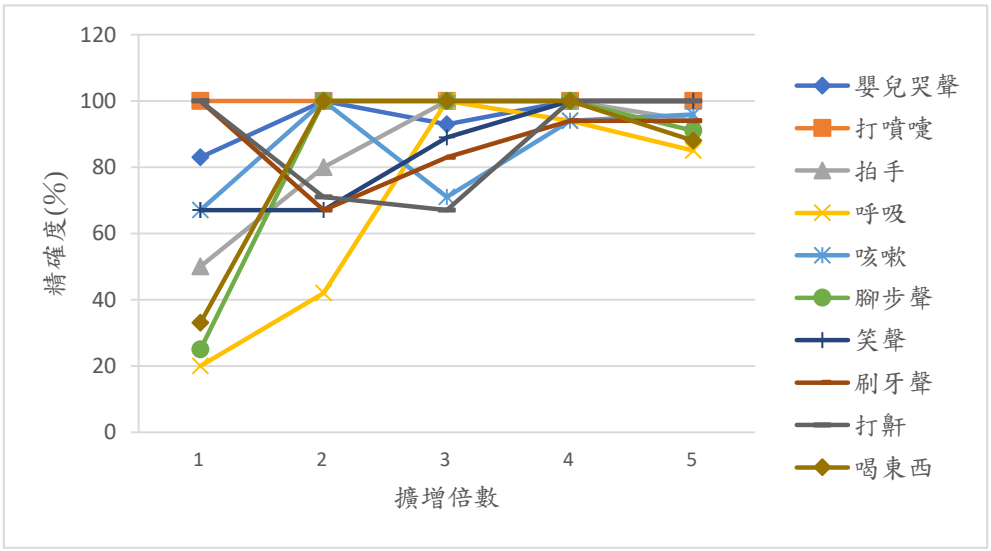



圖 22、人類類別 Precision

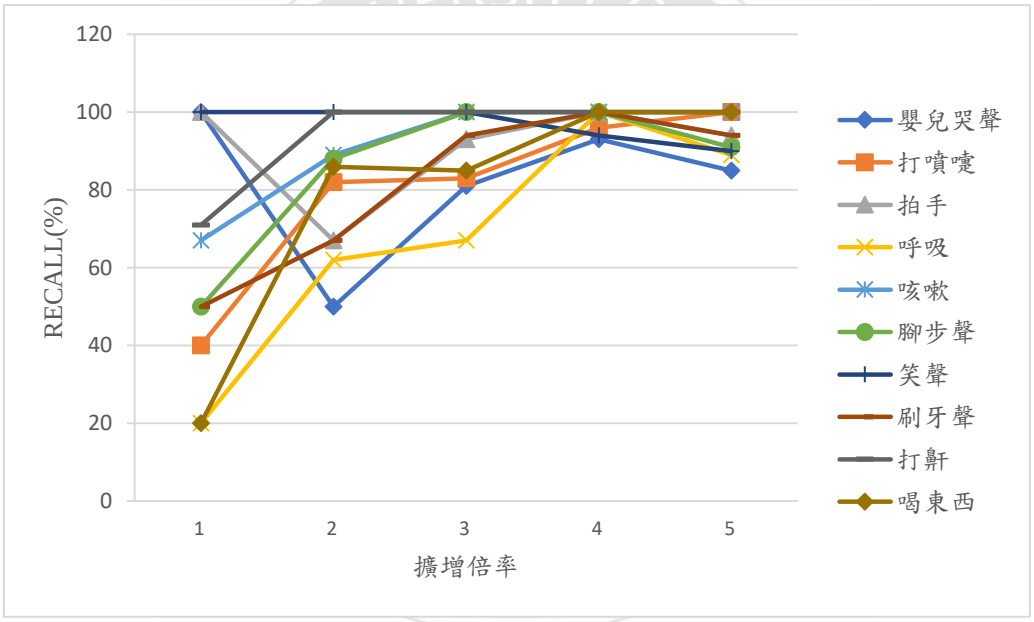



圖 23、人類類別 Recall

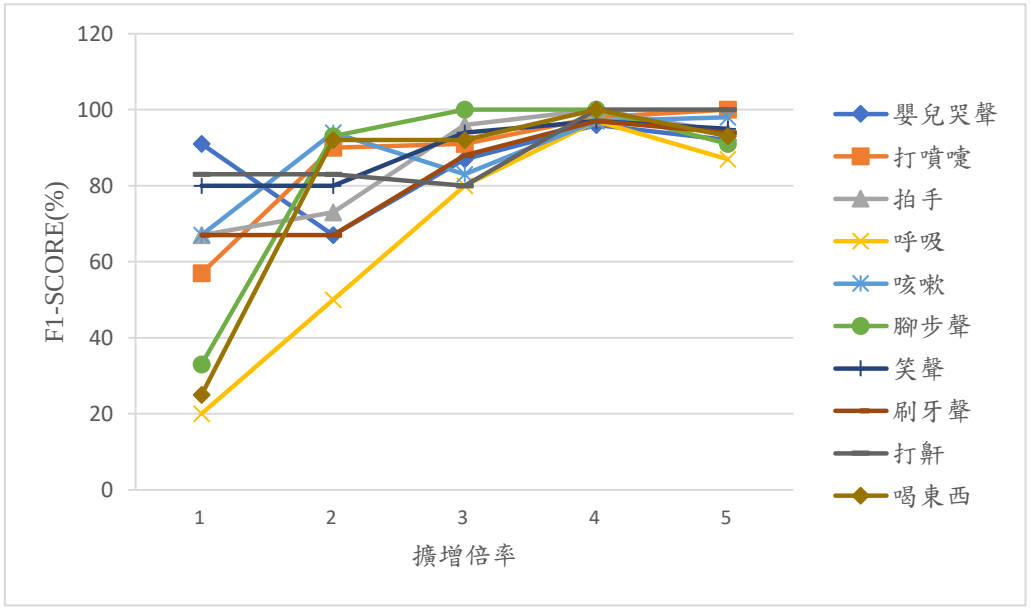



圖 24、人類類別 F1-Score

以上四個評估指標的圖中，橫軸為擴增的倍率，1 為只提取 40 個三角濾波器；2 為提取 40、30 個的三角濾波器做為擴增；3 為 40、30、20；4 為 40、 35、30、20；5 為 40、35、30、25、20。根據四個評估指標，我們可以發現，在人類類別時，在原始資料進行分類時，

平均值也是偏低，而準確率的結果十分的分散，甚至有的在資料量少的時候已能夠準確的分類。而在藉由我們擴增機制進行資料量的提升後，有著明顯的提升，資料分類的準確率也逐漸的趨於平穩。而當在資料擴增達到 5 倍時分類效果已達到了不錯的水準。

(3) 大自然：

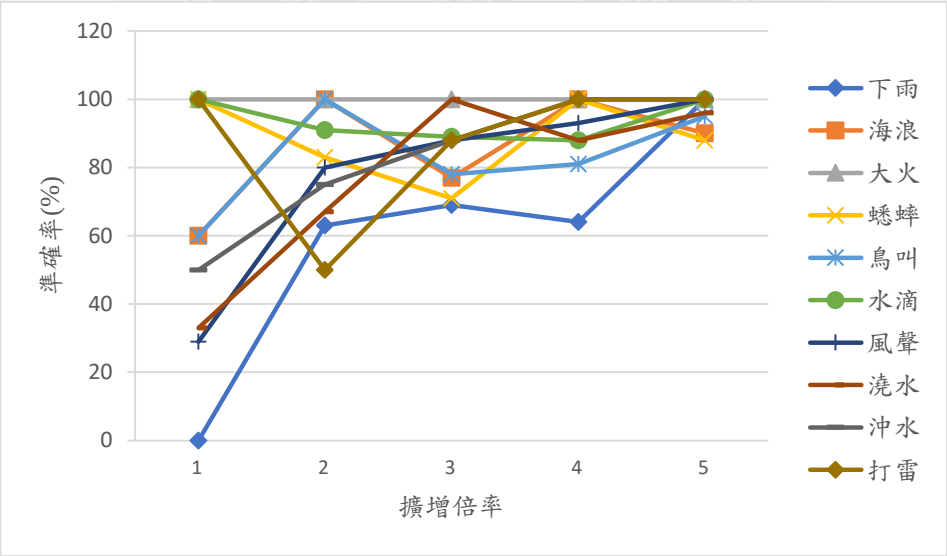



圖 25、大自然類別 Accuracy

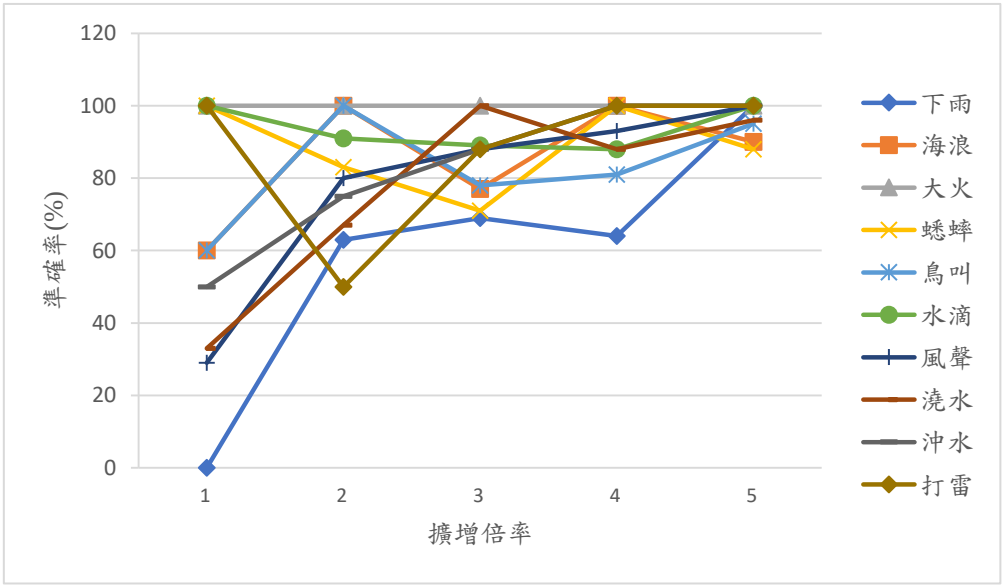



圖 26、大自然類別 Precision

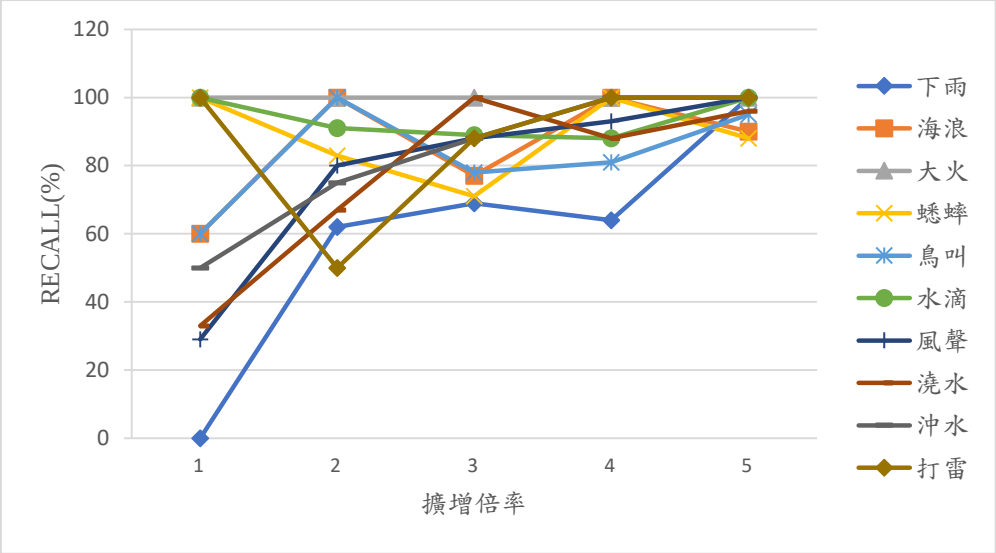



圖 27、大自然類別 Recall

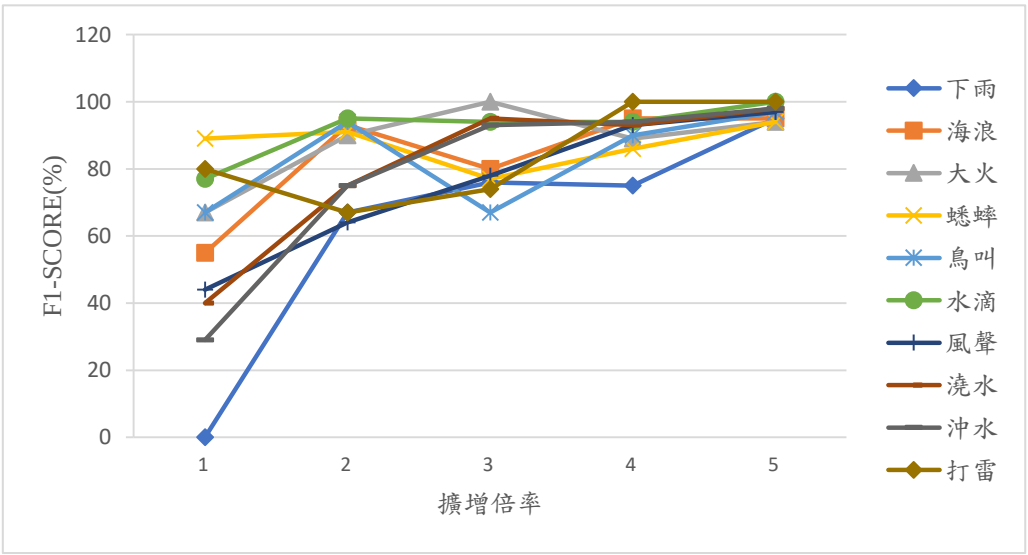



圖 28、大自然類別 F1-Score

以上四個評估指標的圖中，橫軸為擴增的倍率，1 為只提取 40 個三角濾波器；2 為提取 40、30 個的三角濾波器做為擴增；3 為 40、30、20；4 為 40、35、30、20；5 為 40、35、30、25、20。根據四個評估指標，我們可以發現，在大自然類別時，在原始資料進行分類時，相較於其他類別，有著比較好的準確率，大部分都能夠達到 50%以上，

雖然在資料量少的時候已能夠正確的分類。但在藉由我們擴增機制進行資料量的提升後，還能夠明顯的提升，資料分類的準確率也逐漸的趨於平穩。而當在資料擴增達到 5 倍時分類效果已達到了不錯的水準。

(4) 室內聲音：

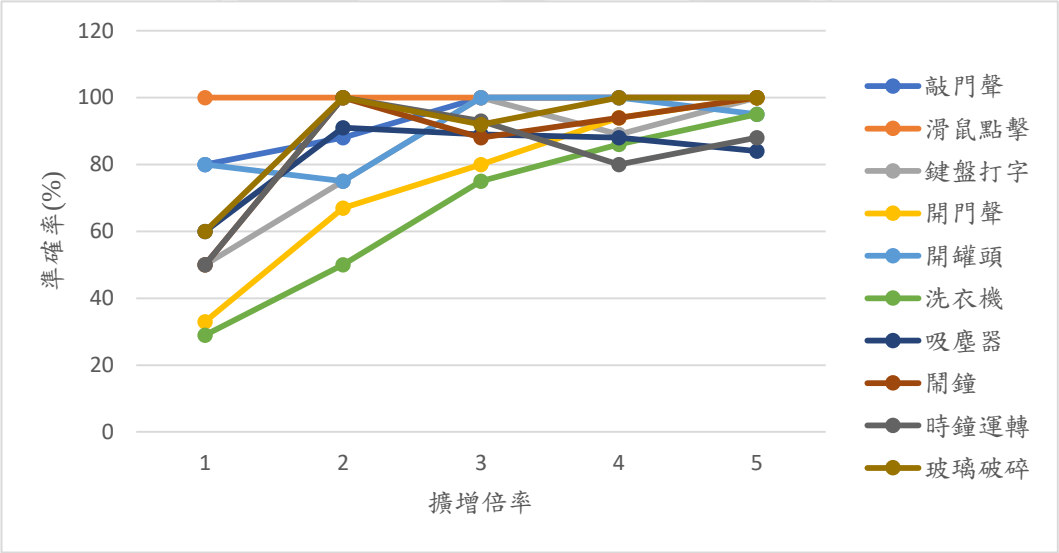



圖 29、室內聲音類別 Accuracy

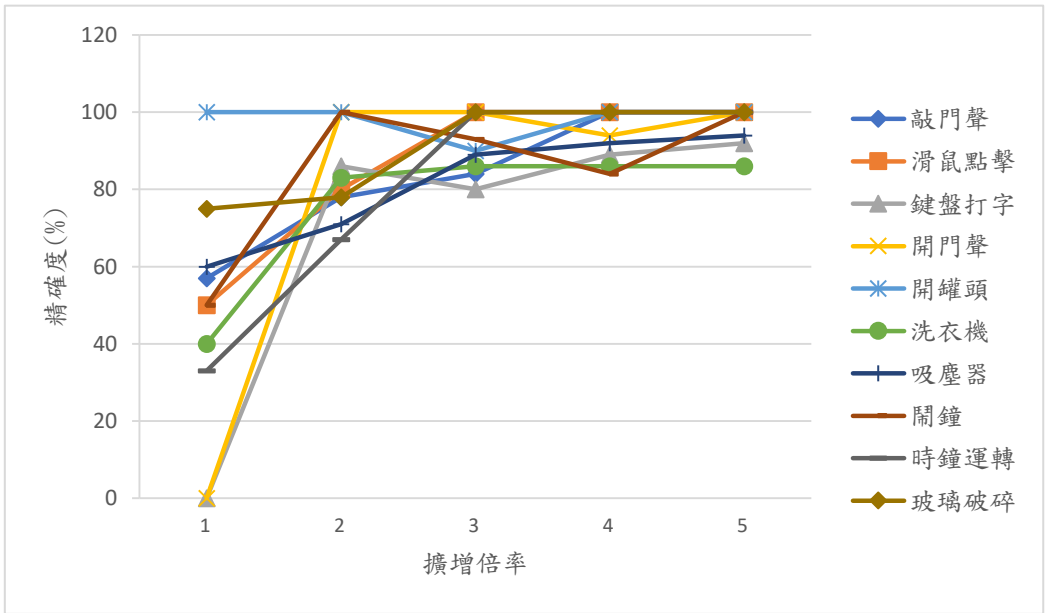



圖 30、室內聲音類別 Precision

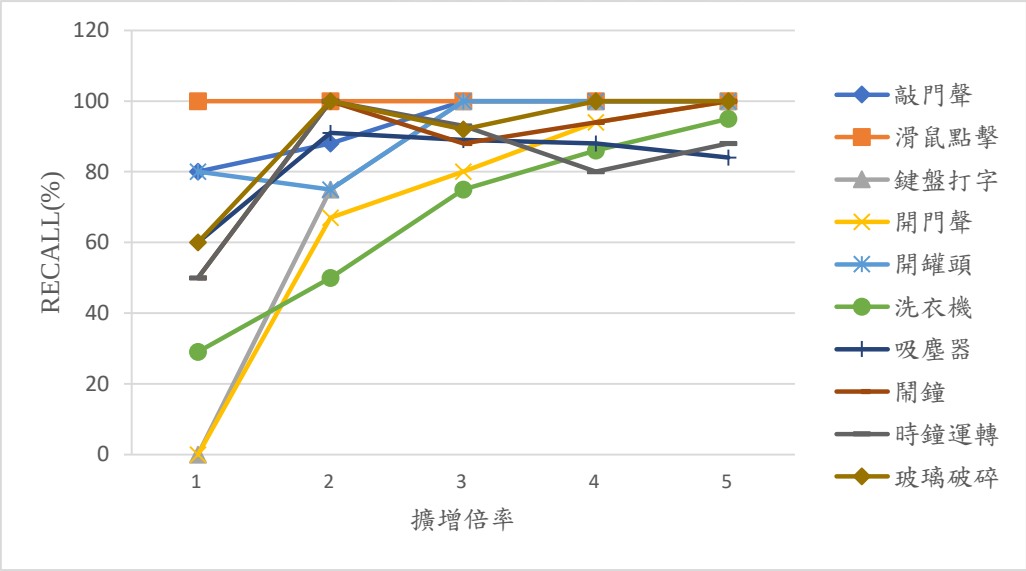



圖 31、室內聲音類別 Recall

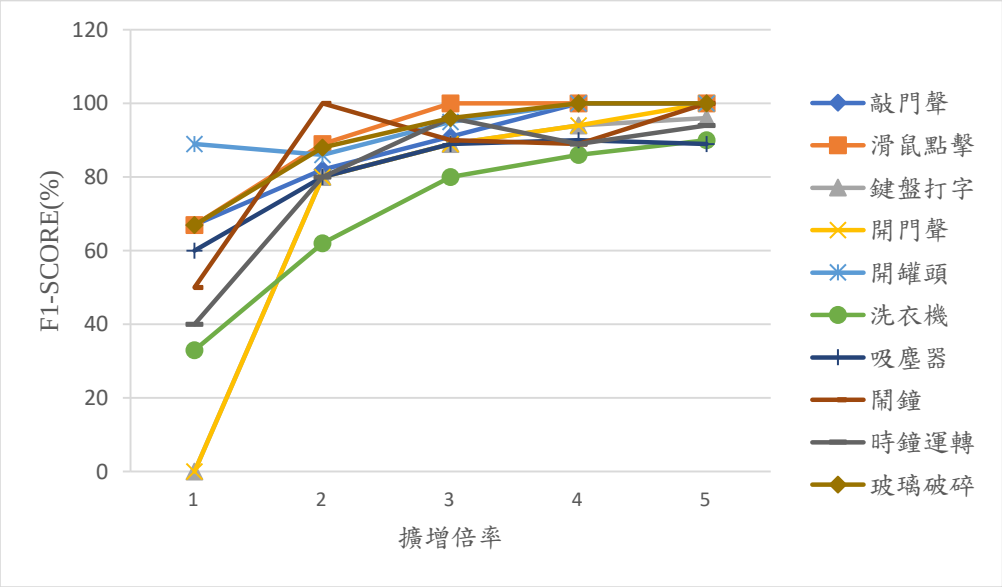



圖 32、室內聲音類別 F1-Score

以上四個評估指標的圖中，橫軸為擴增的倍率，1 為只提取 40 個三角濾波器；2 為提取 40、30 個的三角濾波器做為擴增；3 為 40、30、20；4 為 40、35、30、20；5 為 40、35、30、25、20。根據四個評估指標，我們可以發現，在室內聲音類別時，在原始資料進行分類時，可以發現甚至其中有某類別完全無法分類出來，雖然在資料量少

的時候已能夠正確的分類。但在藉由我們擴增機制進行資料量的提升後，還能夠明顯的提升資料分類的準確率也逐漸的趨於平穩。而當在資料擴增達到 5 倍時分類效果已達到了不錯的水準。

(5) 城市聲音：

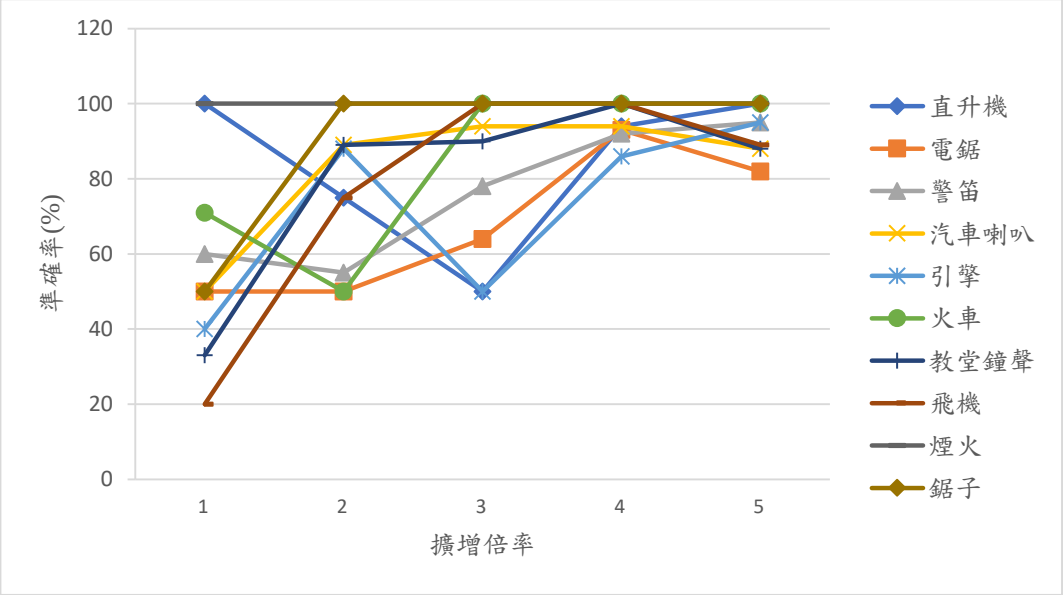



圖 33、城市聲音類別 Accuracy

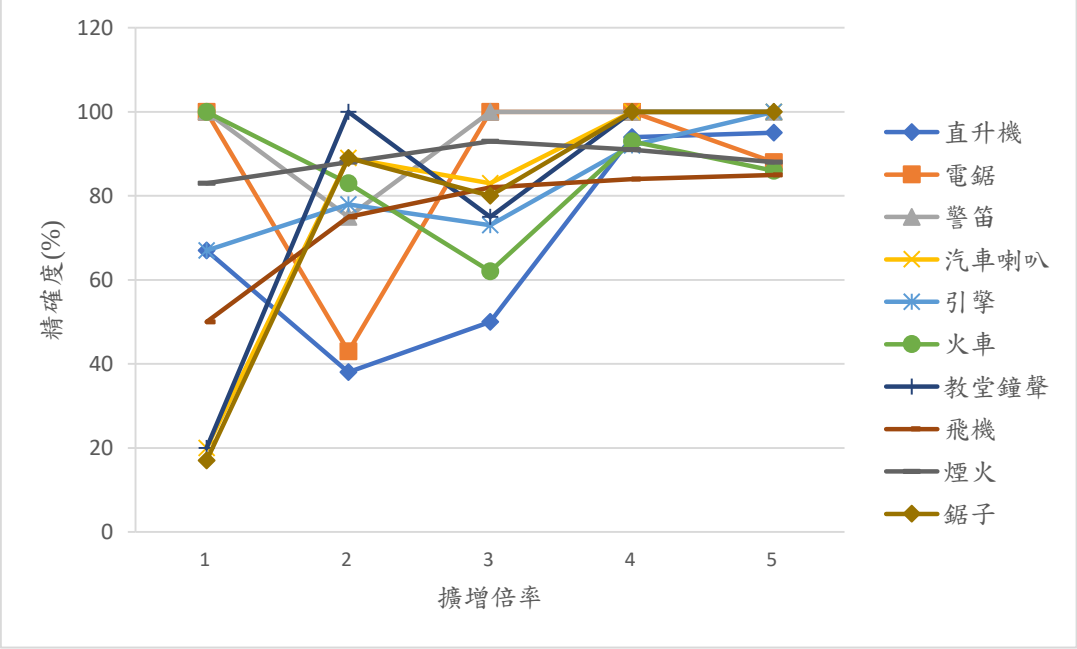



圖 34、城市聲音類別 Precision

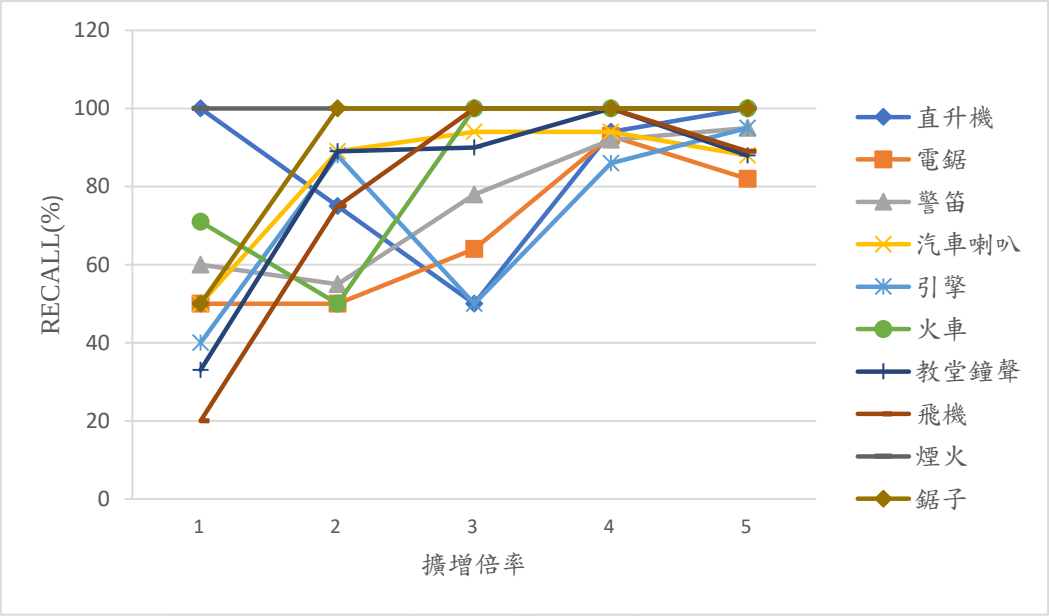



圖 35、城市聲音類別 Recall

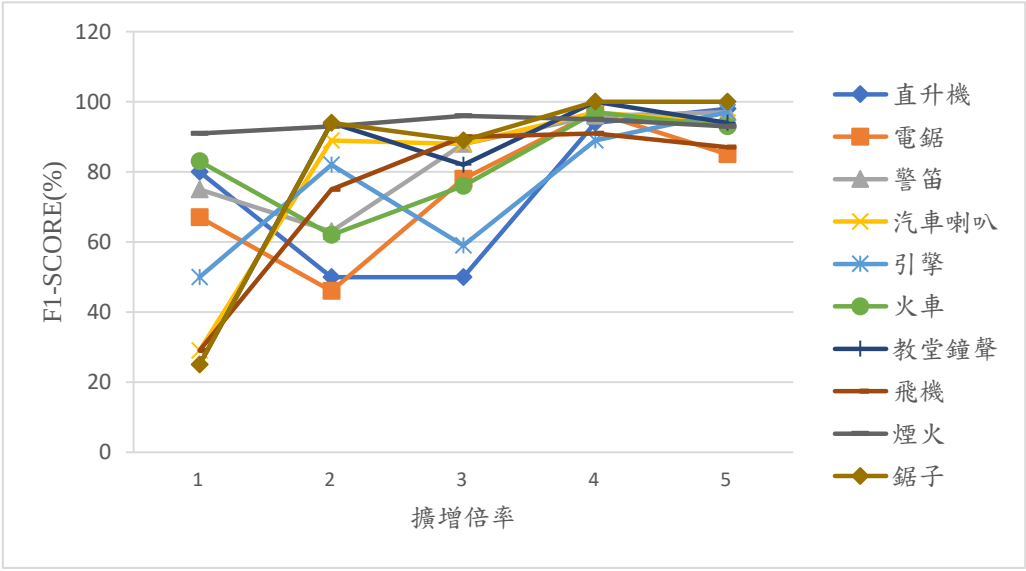



圖 36、城市聲音類別 F1-Score

以上四個評估指標的圖中，橫軸為擴增的倍率，1 為只提取 40 個三角濾波器；2 為提取 40、30 個的三角濾波器做為擴增；3 為 40、30、20；4 為 40、35、30、20；5 為 40、35、30、25、20。根據四個評估指標，我們可以發現，在城市聲音類別時，在原始資料進行分類時，平均值都偏低，而準確率十分的分散，甚至有的在資料量少的時

候已能夠準確的分類。而在藉由我們擴增機制進行資料量的提升後，有著明顯的提升，資料分類的準確率也逐漸的趨於平穩。而當在資料擴增達到 5 倍時分類效果已達到了不錯的水準。

本實驗主要探討，將 ESC-50 聲音資料 5 大主類別進行更細分的次類別分類時，我們可以發現，在次類別的聲音資料時，能夠藉由擴增的方式做為資料量提升的方法，增加訓練資料，並從三角濾波器數量之間的差異做為分類特徵，進而提升分類的效果。

## 4.2. 基於 UrbanSound8K 資料集

當資料集不足時，導致分類的不準確，從上述的資料即可以探討出，可藉由資料擴增的方式來提高分類的準確率，而在這節要介紹的資料集，探討一個和 ESC-50 資料集數量不相同，且各類數量不平均的資料集，已探討擴增的方法是否影響分類的結果。並對於資料集進行分類、資料擴增後的分類，以及提取部分資料進行擴增訓練的情形。

### 4.2.1. UrbanSound8K 原始資料分類結果

原始資料一共 8732 筆資料，藉由 MFCCs 特徵提取，三角濾波器組數一樣固定為 40，藉由上述的步驟進行分類分析，將資料集分割為 9:1 的比例(訓練集：驗證集)進行訓練，再將訓練集以 9:1(訓練集：驗證集)的方式進行分割，讓模型每完整訓練一次的時候能夠即時的驗證模型分類的效果，模型訓練結果如圖 37 所示。

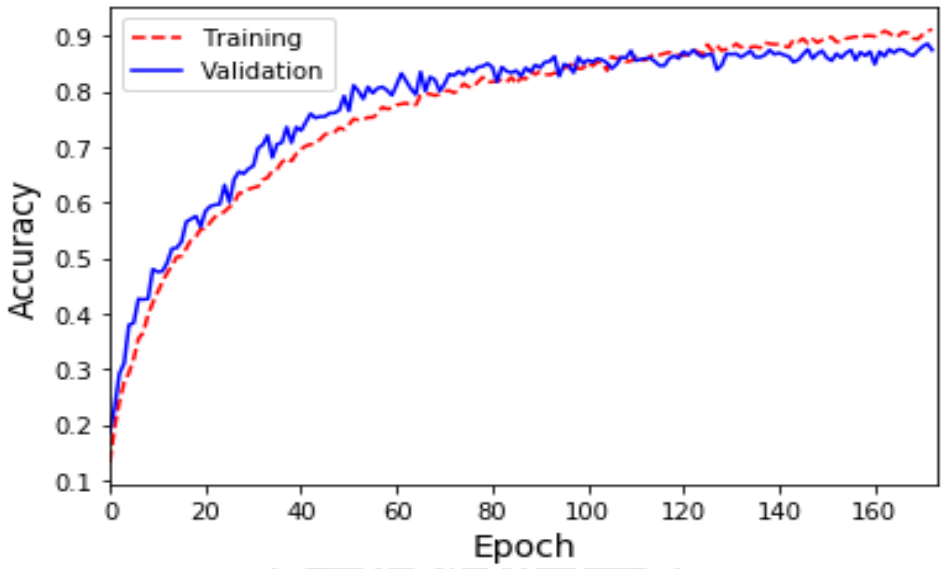



圖 37、UrbanSound8K 原始資料訓練結果

在 500 次的訓練週期中，使用 Early Stopping，可以發現原始 UrbanSound8K 資料集在訓練週期第 173 次週期的時候就已經停止，訓練集與驗證集的結果非常的接近，此結果可以發現，當資料集數量足夠時，訓練的準確度就會很好，而驗證集的準確度也有所提升。

表 8、UrbanSound8K 原始資料各類別混淆矩陣

| 實際 / 驗證 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 86 | 0 | 3 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| 2 | 0 | 33 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 4 |
| 3 | 4 | 0 | 85 | 3 | 0 | 1 | 0 | 0 | 0 | 4 |
| 4 | 0 | 1 | 7 | 87 | 1 | 0 | 0 | 1 | 3 | 1 |
| 5 | 1 | 0 | 3 | 0 | 83 | 0 | 2 | 7 | 0 | 2 |
| 6 | 1 | 0 | 0 | 0 | 0 | 107 | 0 | 3 | 2 | 2 |
| 7 | 0 | 0 | 0 | 1 | 1 | 0 | 32 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 90 | 2 | 2 |
| 9 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 97 | 1 |
| 10 | 0 | 0 | 11 | 0 | 2 | 5 | 0 | 0 | 1 | 82 |

表 8 類別分別為，1：空調聲音、2：汽車喇叭、3：兒童嬉鬧聲、4：狗叫聲、5：鑽孔聲、6：發動機運轉聲、7：槍擊聲、8：手提電鑽、9：警鈴聲、10：城市街道聲，可以發現，在原始資料的分類已經達到一定水準的分類。

表 9、UrbanSound8K 原始資料各類別分類結果

| 評估指標 / 類別 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 空調聲音 | 0.93 | 0.93 | 0.83 | 0..93 |
| 汽車喇叭 | 0.87 | 0.97 | 0.87 | 0.92 |
| 兒童嬉鬧 | 0.88 | 0.77 | 0.88 | 0.82 |
| 狗叫聲 | 0.86 | 0.96 | 0.86 | 0.91 |
| 鑽孔聲 | 0.85 | 0.91 | 0.85 | 0.88 |
| 發動機運轉聲 | 0.93 | 0.93 | 0.93 | 0.93 |
| 槍擊聲 | 0.94 | 0.91 | 0.94 | 0.93 |
| 手提電鑽 | 0.92 | 0.88 | 0.92 | 0.90 |
| 警鈴聲 | 0.97 | 0.92 | 0.97 | 0.94 |
| 城市街道聲 | 0.81 | 0.84 | 0.81 | 0.82 |
| 平均 | 0.90 | 0.92 | 0.90 | 0.89 |

在表 9 中，UrbanSound8K 原始資料的平均分類結果達到了 90%，證明資料量足夠的情況下，分類效果已經可以達到一定的水準。

### 4.2.2. UrbanSound8K 原始資料擴增分類結果

在上一小節中，UrbanSound8K 原始資料的分類準確度已達到了 90%，接著我們將 UrbanSound8K 原始資料進行擴增，這裡一樣使用三角濾波器提取的組數 40 以及 30 做為 MFCCs 特徵提取轉換的依據，此組實驗是為了驗證當資料量足夠時，訓練的準確度可以因為此

擴增資料集的方法，再一次的提升，模型訓練結果如圖 38 所示。

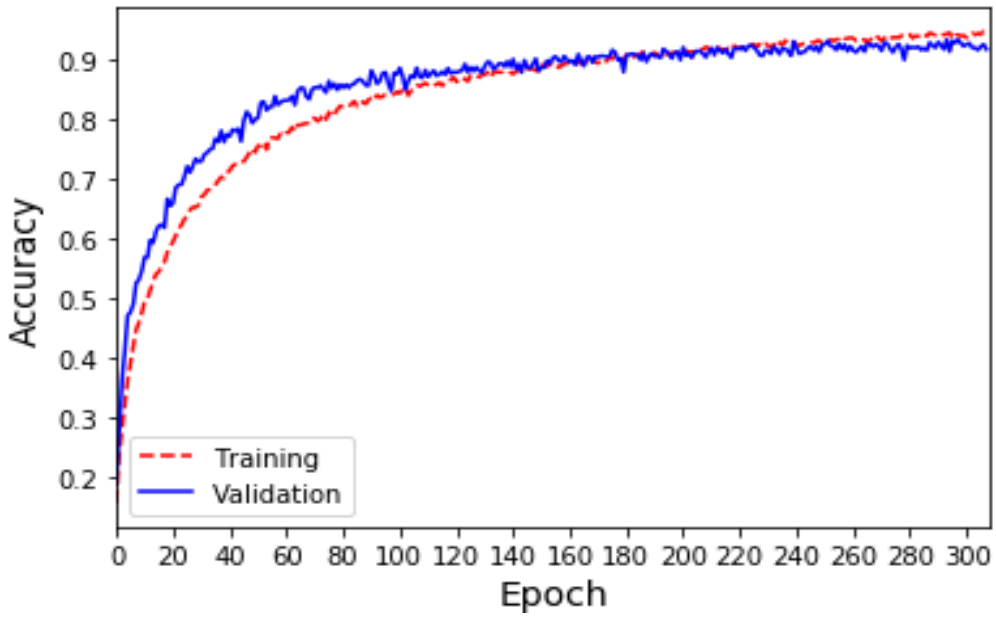



圖 38、UrbanSound8K 原始資料擴增訓練結果

在資料進行擴增後，訓練的準確度與原始資料做比較，發現 500 次的訓練週期中，在 Early Stopping 機制中，資料在訓練第 308 次週期時停止，訓練集與驗證集的結果非常的接近，此結果可以發現，當資料集數量再進行擴增時，訓練的準確度相較於原始資料只提升了 2%，而驗證集的準確度雖有所提升，但提升的效果有限。

表 10、UrbanSound8K 原始資料擴增後各類別混淆矩陣

| 實際 / 驗證 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 189 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 71 | 0 | 0 | 5 | 4 | 0 | 1 | 4 | 2 |
| 3 | 4 | 0 | 183 | 1 | 1 | 4 | 0 | 0 | 5 | 7 |
| 4 | 1 | 1 | 11 | 183 | 3 | 4 | 0 | 0 | 9 | 1 |
| 5 | 2 | 0 | 0 | 0 | 164 | 0 | 0 | 10 | 2 | 2 |
| 6 | 4 | 0 | 0 | 1 | 0 | 194 | 0 | 0 | 0 | 2 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 68 | 0 | 0 | 0 |
| 8 | 4 | 0 | 0 | 0 | 4 | 0 | 0 | 205 | 0 | 1 |
| 9 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 187 | 0 |
| 10 | 3 | 5 | 5 | 2 | 0 | 2 | 0 | 0 | 6 | 156 |

在表 10 中類別分別為：1 表示為空調聲音、2 表示為汽車喇叭、3 表示為兒童嬉鬧聲、4 表示為狗叫聲、5 表示為鑽孔聲、6 表示為發動機運轉聲、7 表示為槍擊聲、8 表示為手提電鑽、9 表示為警鈴聲、10 表示為城市街道聲。

表 11、UrbanSound8K 原始資料擴增後各類別分類結果

| 評估指標 / 類別 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 空調聲音 | 0.98 | 0.91 | 0.98 | 0.94 |
| 汽車喇叭 | 0.81 | 0.91 | 0.81 | 0.86 |
| 兒童嬉鬧 | 0.89 | 0.86 | 0.89 | 0.88 |
| 狗叫聲 | 0.86 | 0.96 | 0.86 | 0.91 |
| 鑽孔聲 | 0.91 | 0.93 | 0.91 | 0.92 |
| 發動機運轉聲 | 0.97 | 0.93 | 0.97 | 0.95 |
| 槍擊聲 | 1.00 | 1.00 | 1.00 | 1.00 |
| 手提電鑽 | 0.96 | 0.94 | 0.96 | 0.95 |
| 警鈴聲 | 0.98 | 0.88 | 0.98 | 0.93 |
| 城市街道聲 | 0.87 | 0.91 | 0.80 | 0.85 |
| 平均 | 0.92 | 0.92 | 0.92 | 0.92 |

在表 11 中，UrbanSound8K 原始資料的平均分類結果達到了 92%，相較於原始資料，但準確度提升有限，證明資料量足夠的情況下，分類效果提升有限，在擴增資料方法中，我們可以發現在第 7 類：槍擊聲中，有 100%的分類結果，準確的將槍擊聲分辨出來，各類也能夠更準確的分辨出來，雖然整體類別的分類都有小幅度的提升，但第 9 類警鈴聲雖然有所降低，但是整體平均還是有提升的趨勢。

### 4.2.3. UrbanSound8K 部分資料提取擴增分類結果

在上述所探討的實驗中，當資料量不足時，藉由資料擴增的方式，可以有效的提升分類的準確度，反觀當資料量足夠時，對於資料擴增的方法，提升的效果有限，甚至是會有所降低。因此對於資料量足夠的 UrbanSound8K 資料集中，我們將模擬資料量不足的情形，對 UrbanSound8K 資料集只提取一半(4366 筆)的資料集，並藉由這一半的資料進行資料擴增的方法，如此一來，我們就能夠以少資料量的情形下，藉由資料擴增的方式達到如同原始資料相同甚至是更好的分類準確度，訓練結果如圖 39 所示。

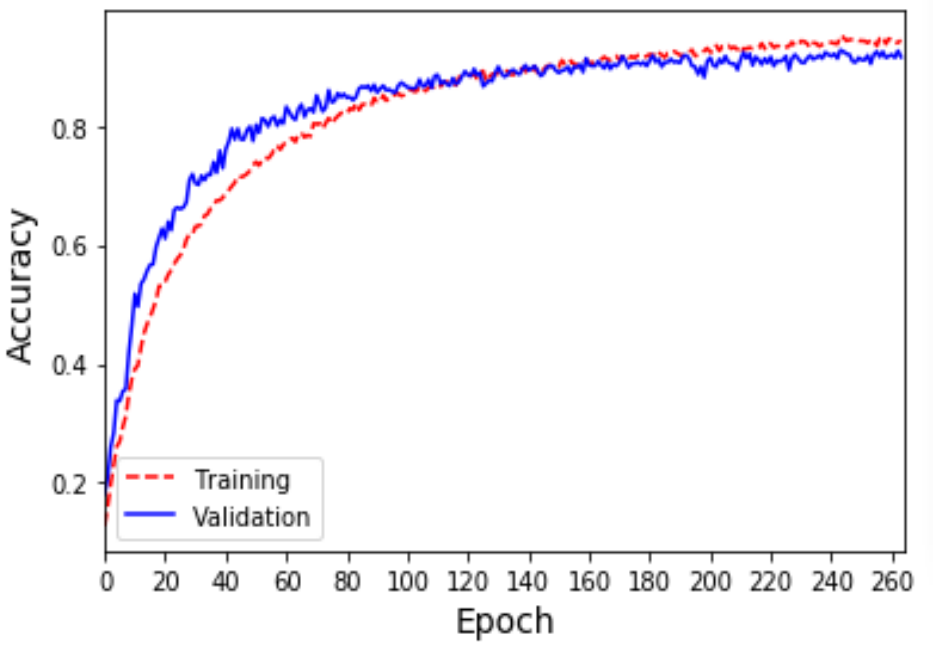



圖 39、UrbanSound8K 部分提取資料擴增訓練結果

在部分提取資料擴增後訓練，發現 500 次的訓練週期中，在 Early Stopping 機制中，資料在訓練第 372 次週期時停止，訓練集與驗證集的結果非常的接近，此結果可以發現，雖然資料量不足，對於找到資料類別相關的特徵，訓練次數相較於前面的實驗多出了約 50 次週期，但少資料量藉由資料擴增的方式，準確度卻相較於原始資料提升了 5%，而驗證集的準確度也有所提升，因此證明資料擴增的方式能夠提供給聲音分類中，以少量的資料達到更好的準確度。

表 12、UrbanSound8K 部分提取資料擴增後各類別混淆矩陣

| 實際 / 驗證 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 89 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 |
| 2 | 2 | 28 | 3 | 0 | 1 | 2 | 0 | 0 | 0 | 2 |
| 3 | 1 | 0 | 85 | 2 | 1 | 0 | 0 | 0 | 1 | 7 |
| 4 | 0 | 1 | 5 | 88 | 2 | 1 | 0 | 0 | 2 | 2 |
| 5 | 1 | 0 | 0 | 3 | 87 | 1 | 0 | 3 | 0 | 3 |
| 6 | 1 | 0 | 0 | 1 | 0 | 107 | 0 | 1 | 3 | 1 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 34 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 90 | 0 | 5 |
| 9 | 0 | 0 | 2 | 1 | 1 | 1 | 0 | 0 | 91 | 4 |
| 10 | 1 | 0 | 4 | 0 | 0 | 1 | 0 | 0 | 1 | 94 |

在表 12 中類別分別為，1：空調聲音、2：汽車喇叭、3：兒童嬉鬧聲、4：狗叫聲、5：鑽孔聲、6：發動機運轉聲、7：槍擊聲、8：手提電鑽、9：警鈴聲、10：城市街道聲，可以發現，相較於原始資料所分類出來的結果，更加的準確。

表 13、UrbanSound8K 部分提取資料擴增後各類別分類結果

| 評估指標 / 類別 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 空調聲音 | 0.97 | 0.93 | 0.97 | 0.95 |
| 汽車喇叭 | 0.74 | 0.97 | 0.74 | 0.84 |
| 兒童嬉鬧 | 0.84 | 0.85 | 0.88 | 0.86 |
| 狗叫聲 | 0.90 | 0.93 | 0.87 | 0.90 |
| 鑽孔聲 | 0.89 | 0.92 | 0.89 | 0.90 |
| 發動機運轉聲 | 0.94 | 0.93 | 0.93 | 0.93 |
| 槍擊聲 | 1.00 | 1.00 | 1.00 | 1.00 |
| 手提電鑽 | 0.95 | 0.96 | 0.92 | 0.94 |
| 警鈴聲 | 0.91 | 0.93 | 0.91 | 0.92 |
| 城市街道聲 | 0.93 | 0.80 | 0.93 | 0.86 |
| 平均 | 0.91 | 0.92 | 0.90 | 0.91 |

在表 13 中，UrbanSound8K 部分資料提取擴增的分類結果達到了 95%，相較於原始資料，提升了 5%。雖然訓練時間增加了 50 次週期，但是藉由增加的訓練次數，使準確度有所提升；我們可以發現當資料不足時，資料擴增能夠提升分類效果，不只是第 7 類別：槍擊聲可以 100%的分類，連同第 3 類類別的兒童嬉鬧聲能夠 100%的準確分類，整體資料都提升到了 90%，此實驗可證明，當我們藉由擴增方式讓資料量提升時，能夠提升分類的準確率甚至更好。

在本小節探討基於 UrbanSound8K 在不同情形時所分類的結果，在原始資料集整體平均已達到了 90%，因此對於足夠數量的資料集時，資料擴增的方式進行分類的效果有限，因此模擬了當資料量不足時(將資料集只提取一半)，進行資料擴增的方法，此結果證明資料的擴增，對於資料量不足時是可行的，甚至比原始資料分類的結果更佳，如表 14 所示。

表 14、基於 UrbanSound8K 不同的分類結果

| 評估指標 / 方法 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| 原始資料 | 0.91 | 0.92 | 0.90 | 0.89 |
| 擴增資料 | 0.92 | 0.92 | 0.92 | 0.92 |
| 部分提取擴增 | 0.95 | 0.95 | 0.94 | 0.95 |

# 第五章、結論與未來展望

聲音和我們的生活息息相關，日常生活中包含了各式各樣的聲音，不僅是提供人與人之間基本的溝通，且讓大自然的事物也都有屬於自己獨特的聲音。但現今聲音應用的各個場域下，因為資料量的不足，進而導致分類的準確率不高。因此本文提出基於卷積神經網路的聲音分類機制，利用 MFCCs 聲音特徵提取方式，將聲音訊號轉換為頻譜圖。並以頻譜圖做為輸入資料，經由 CNN 模型的特徵學習後，分辨聲音的類別，並利用 MFCCs 裡不同的三角濾波器的數量提取，做為資料擴增的方法，運用於當資料量不足的結果下，亦能夠提供良好的分類效果，進而發展出優良的聲音分類器。

本文一共探討了兩個不同的資料集，資料集 ESC-50 時，因為資料主類別中類別的複雜，導致的聲音分類不準確。藉由資料擴增的方法，雖然能夠有所提升，但不足以代表此方法的可行性。因此我們以次類別的方式進行探討，當整體資料集在資料進行擴增到 5 倍時，資料分類的效果可以達到最佳，而在第二個資料集 UrbanSound8K 中，藉由不同的資料集，去探討資料擴增的方式是否有辦法因應不同來源所收集到的聲音資料進行分類，且探討進行資料擴增時分類的準確率是否能夠再提升，實驗結果顯示當資料量足夠時，使用資料擴增的方

式，對於分類的準確度提升效果有限。因此我們將資料集UrbanSound8K提取一半的資料量，藉由這一半的資料進行資料擴增，雖然訓練的週期時間相較於原始資料花費更久，但分類的準確度卻比原始資料提升 5%，而且我們只利用了一半的原始資料，即可建立出高準確率的訓練模型。

在未來工作中，由於實際收音時所產生的環境噪音、雜訊、聲音資料重疊…等相關問題都會影響到聲音分類模型的效能，因此我們未來可以加入不同的特徵提取方法或修正模型架構來加強本文所提出來的擴增方法；也能夠選擇深度更深的神經網路模型，根據上述的改善方式，未來希望能夠以多類別的方式，提供實際收音的類別進行探討與應用。

# 參考文獻


[1] Y. LeCun, Y. Bengio, G.E. Hinton,“Deep Learning,” Nature, Vol. 521, pp. 436-444, 27 May. 2015.

[2] 張志勇, 翁仲銘, 石貴平, 廖文華, ” 物聯網概論,” 碁峰資訊股份有限公司, Jan. 2013.

[3] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, S. Karnouskos and D. Boyle, “From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence,” 1st edition, Elsevier, 16 Apr. 2014.

[4] D. Singh, G. Tripathi, and A.J. Jara, “A survey of Internet-of-Things: Future Vision, Architecture, Challenges and Services,” IEEE World Forum on Internet of Things (WF-IoT), pp. 287-292, Seoul, 6-8 Mar. 2014.

[5] B. Wang, L. Miao, H. Dong and Z. Zheng, "The Research of Lung Sound Signals Based on Cepstrum Analysis," 2012 International Conference on Biomedical Engineering and Biotechnology, pp. 934-938, Macao, 28-30 May 2012.

[6] A. Hamidah, R. Saputra, T. L. R. Mengko, R. Mengko and B. Anggoro, "Effective heart sounds detection method based on signal's characteristics," 2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1-4, Phuket, 24-27 Oct. 2016.

[7] T. Hayashi, M. Nishida, N. Kitaoka and K. Takeda, "Daily activity recognition based on DNN using environmental sound and acceleration signals," 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 2306-2310, Nice, 31 Aug.-4 Sep. 2015.

[8] “音頻訊號處理” [Online], https://www.wikiwand.com/zh-mo/%E9%9F%B3%E9%A2%91%E4%BF%A1%E5%8F%B7%E5%A4%84%E7%90%86, (Accessed on 7 Oct. 2020)

[9] T. Han, K. Kim and H. Park, "Location Estimation of Predominant Sound Source with Embedded Source Separation in Amplitude-Panned Stereo Signal," in IEEE Signal Processing Letters, vol. 22, no. 10, pp. 1685-1688, Oct. 2015.

[10] E. Półrolniczak and M. Kramarczyk, "Estimation of singing voice types based on voice parameters analysis," 2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, pp. 63-68, 20-22 Sept. 2017.

[11] M. Meinard, E. Sebastian, ”Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features,” 12th

International Society for Music Information Retrieval Conference, PP. 215-220, 24-28 Oct. 2011.
[12]K. Z. Thwe and N. War, "Environmental sound classification based on time-frequency representation," 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 251-255, Kanazawa, 26-28 June 2017.
[13]T. Viarbitskaya and A. Dobrucki, "Audio processing with using Python language science libraries," 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 350-354, Poznan, Poland, 19-21 Sep. 2018.
[14]"Basic Acoustic Features (基本聲學特徵)" [Online], http://mirlab.org/jang/books/audioSignalProcessing/audioBasicFeature.asp?title=3-2%20Basic%20Acoustic%20Features%20(%B0%F2%A5%BB%C1n%BE%C7%AFS%BCx)&language=chinese, (Accessed on 7 Oct 2020)
[15]U. Bhattacharjee, S. Gogoi and R. Sharma, "A statistical analysis on the impact of noise on MFCC features for speech recognition," 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-5, Jaipur, India, 23-25 Dec. 2016.
[16]M. Sadeghi and H. Marvi, "Optimal MFCC features extraction by differential evolution algorithm for speaker recognition," The 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), Shahrood, pp. 169-173, Iran, 20-21 Dec. 2017.
[17]M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," The 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, Gold Coast, QLD, Australia, 13-15 Dec. 2010.
[18]J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol 61, pp 85-117, Jan. 2015.
[19]H-C. Shin, H-R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R-M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE Transactions on Medical Imaging, Vol. 35, Iss. 5, May 2016.
[20]L. Deng, D. Yu,"Deep Learning : Methods and Applications," 1stedition, Now Foundations and Trends, 30 Jun. 2014.
[21]"AI 來襲！三分鐘看懂人工智慧" [Online],

https://makerpro.cc/2019/05/introduction-to-ai/ (Accessed on 21 Oct 2020)
[22]R. Palaniappan and K. Sundaraj, "Respiratory sound classification using cepstral features and support vector machine," 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 132-136, Trivandrum, 19-21 Dec. 2013.
[23]C. Y. Yu, H. Liu and Z. M. Qi, "Sound event detection using deep random forest", Detection and Classification of Acoustic Scenes and Events, Munich, Germany, 16 Nov. 2017.
[24]S. Tangkawanit, C. Pinthong and S. Kanprachar, "Development of gunfire sound classification system with a smartphone using ANN," 2018 International Conference on Digital Arts, Media and Technology (ICDAMT), pp. 168-172, Phayao, 25-28 Feb. 2018.
[25]J. K. Das, A. Ghosh, A. K. Pal, S. Dutta and A. Chakrabarty, "Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features," 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), pp. 1-9, Fez, Morocco, 21-23 Oct. 2020.
[26]J. Xie, K. Hu, M. Zhu, J. Yu and Q. Zhu, "Investigation of Different CNN-Based Models for Improved Bird Sound Classification," in IEEE Access, vol. 7, pp. 175353-175361, Dec. 2019.
[27]Z. Chi, Y. Li and C. Chen, "Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), pp. 251-254, Dalian, China, 19-20 Oct. 2019.
[28]
[29]K. J. Piczak, "ESC: Dataset for environmental sound classification," Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018, Oct. 2015.
[30]J. Salamon, C. Jacoby and J. P. Bello, "A dataset and taxonomy for urban sound research", Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041-1044, 2014, [Online], https://doi.org/10.1145/2647868.2655045, (Accessed on 3 Dec 2020)
[31]"librosa — librosa 0.8.0 documentation" [Online], https://librosa.org/doc/latest/index.html# (Accessed on 9 Dec 2020)
[32]E. Şaşmaz and F. B. Tek, "Animal Sound Classification Using A Convolutional Neural Network," 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 625-629, Sarajevo, 20-23 Sep. 2018.
[33]H. Li, J. Li, X. Guan, B. Liang, Y. Lai and X. Luo, "Research on

Overfitting of Deep Learning," 2019 15th International Conference on Computational Intelligence and Security (CIS), pp. 78-81, Macao, 13-16 Dec. 2019.

[34] “ Early stopping”, [Online], https://en.wikipedia.org/wiki/Early_stopping (Accessed on 9 Dec 2020)

[35] “語音信號的梅爾頻率倒譜係數(MFCC)的原理講解及 python 實現” [Online], https://www.cnblogs.com/LXP-Never/p/10918590.html (Accessed on 9 Dec 2020)

[36]J. L. Garcia-Balboa, M. V. Alba-Fernandez, F. J. Ariza-López and J. Rodriguez-Avi, "Homogeneity Test for Confusion Matrices: A Method and an Example," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 1203-1205, Valencia, 22-27 Jul. 2018.