以語音情緒辨識增強深度偽造音訊檢測

Enhancing Deepfake Audio Detection with Speech Emotion Recognition

王智弘

中央警察大學資訊管理研究所
桃園市龜山區大崗里樹人路 56 號
im1113058@mail.cpu.edu.tw

張明桑

中央警察大學資訊管理研究所
桃園市龜山區大崗里樹人路 56 號
mschang@mail.cpu.edu.tw

*林曾祥

中央警察大學資訊管理研究所
桃園市龜山區大崗里樹人路 56 號
jslin168@mail.cpu.edu.tw

摘要

隨著資訊科技的發展，運算硬體成本降低的同時，軟體技術也不斷受到優化，促使深度學習技術門檻降低並廣泛使用，然而不肖人士也透過深度偽造（Deepfake）技術來偽造音訊，製作出不實的影音片段並散布，導致深度偽造的影音影響人們的生活甚鉅，所以深度偽造音訊檢測的重要性與日俱增。同樣地，在深度學習模型的助益之下，語音情緒辨識的研究蓬勃發展，學者可以透過神經網路萃取音訊中的情緒特徵，並用於情緒類別的辨識任務。本研究建置基準和增強兩種系統：基準系統是將深度偽造音訊資料集轉換成梅爾頻率倒譜係數（Mel-Frequency Cepstral Coefficients，MFCCs）後輸入 XGBoost（eXtreme Gradient Boosting）分類器進行分類任務並記錄效能，包含英文的 ASVSPOOF（Automatic Speaker Verification Spoofing and Countermeasures Challenge）資料集和中文的 CFAD（A Chinese Dataset for Fake Audio Detection）資料集；增強系統則是先訓練出語音情緒辨識系統，該語音情緒辨識系統是以 EfficientNet 模型為基礎進行遷移式學習，加入瑞爾森情緒語言與歌曲視聽資料集（Ryerson Audio-Visual Database of Emotional Speech and Song，RAVDESS）微調，以產生可達成

語音情緒辨識任務的語音情緒辨識系統，並加入基準系統之中，將梅爾頻率倒譜係數加上所萃取的情緒特徵向量，再輸入 XGBoost 分類器進行分類任務並記錄效能。實驗結果顯示，增強系統在各項評估指標中勝過基準系統，顯示加入語音情緒辨識系統可以增強深度偽造音訊檢測系統。

關鍵字：深度偽造音訊檢測、語音情緒辨識、遷移式學習。

Abstract

With the advancement of information technology and the simultaneous reduction in computational hardware costs, continuous optimization of software technologies has facilitated the widespread use of deep learning techniques by lowering entry barriers. However, unscrupulous individuals exploit deepfake technology to manipulate audio, creating deceptive audiovisual content and disseminating misinformation. The profound impact of deepfake audiovisual content on people's lives underscores the escalating importance of detecting such deceptive audio. Concurrently, leveraging deep learning models, research in the field of speech emotion recognition has flourished. Scholars can extract emotional features from audio using neural networks for emotion categorization tasks. This study establishes two systems: a baseline system and an enhanced system. The baseline system transforms deepfake audio datasets into MFCCs（Mel-Frequency Cepstral Coefficients） and utilizes an XGBoost (eXtreme Gradient Boosting) classifier for classification tasks. Performance is recorded using the ASVSPOOF (Automatic Speaker Verification Spoofing and Countermeasures Challenge) dataset in English and the CFAD (A Chinese Dataset for Fake Audio Detection) dataset in Chinese. The enhanced system, on the other hand, first trains a speech emotion recognition system based on the EfficientNet model through transfer learning. This system is fine-tuned with the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The resulting speech emotion recognition system is integrated into the baseline system by combining Mel Frequency Cepstral Coefficients with extracted emotional feature vectors. The combined input is then classified using the XGBoost classifier, and performance metrics are recorded. Experimental results demonstrate that the enhanced system outperforms the baseline system across various evaluation metrics, indicating that the integration of a speech emotion recognition system enhances the capability of deepfake audio detection systems.

Keywords: Deepfake Audio Detection, Speech Emotion Recognition, Transfer Learning.

壹、 導論

警察機關致力於執行警察任務，包含依法維持公共秩序、保護社會安全、防止一切危害及促進人民福利等，隨著科技的進步，警察執行勤務的方式也與時俱進，而近年來資訊科技普遍受到大眾關注和積極使用，促使資通訊硬體設備取得成本降低，軟體技術也不斷受到優化，學者們在網路上進行學術交流，各項應用軟體的下載連結和使用說明也在網路上廣為流傳，形成資訊應用技術門檻的降低，讓大眾可以自由快速地享受科技化為生活帶來的便利。以正面的角度來看，讓更多使用者可以透過工具程式完成其需求，例如語音情緒辨識（Speech Emotion Recognition，SER）、剪輯影音和流程自動化等，然而以負面的角度來看，則是讓更多犯罪者透過新興技術來達成犯罪的目的，像是散播惡意程式（Malicious Software，Malware）控制殭屍電腦來製造分散式阻斷服務攻擊，或透過深度學習來進行換臉、變聲的深度偽造影音等科技化犯罪。

新興的科技化犯罪態樣繁多，近年來全球興起以假消息挾帶深度偽造影音，企圖誘導民眾點擊惡意連結或是接收錯誤資訊，面對深度偽造影音，警察機關需要隨著時代的演進加強偵辦刑案技術的科技化，現今在檢測深度偽造影音的研究上已有可以對面部輪廓進行檢測，或是對音訊特徵進行檢測，對於中文的深度偽造音訊檢測效能則是我們特別關注的重點。所以我們希望能將跨語言通用的情緒特徵，用以增強深度偽造音訊檢測。本文接著將於第二節進行相關文獻的探討，第三節說明研究方法包含實驗的環境、方法和流程，第四節展示實驗結果並進行討論，最後在第五節提出結論及未來研究方向的建議。

貳、 文獻探討

一、語音情緒辨識

由機器對於人類反應中所代表的情緒種類進行辨識，是近年來因為人機互動興起所浮現的需求之一，有多種方法可以識別出人類的情緒，像是透過視覺訊號、聲音訊號、物理訊號或文字語意等方法[1]。也能透過結合不同方法來加強情緒辨識的效果[2][3]。而在深度偽造語音的檢測任務之中，適合透過語音進行情緒辨識。而近期達成語音情緒辨識效果的主要技術是深度學習，以一系列的流程，將帶有情緒的音訊進行前處理，再輸入深度學習系統，萃取出情緒特徵，並加以辨識情緒種類[4]。作為輸入深度學習的資料庫，需要經由人工標記情緒，近年來，在數個有標記的語音資料庫當中，瑞爾森情緒語言與歌曲視聽資料庫受到廣泛地使用，該資料庫是由經過英語口音篩選的專業演員所錄製，錄音人員的男女比例也有經過平衡，適合用來進行深度學習的訓練[5]。而在取得語音資料庫後，需要對音訊進行前處理，使音訊能夠輸入進深度學習系統，而此一階段也能夠做適當的處理來提高之後情緒辨識的準確率[6]。為了找出音訊的情緒特徵並輸入卷

積神經網路進行訓練，需要將音訊進行前處理，在前處理的方式當中以梅爾頻譜圖最能保留音訊中的情緒特徵[7]。

二、遷移式學習

為了對龐大的神經網路進行訓練，需要耗費龐大的運算資源和訓練時間，但是即使擁有足夠強大的運算環境，訓練資料的不足仍然會對神經網路造成過度擬合的問題，近年來學者發現以非情緒辨識任務為目標訓練的預訓練模型，經過添加神經層微調後，便能夠提升神經網路的泛化性而更好地達成任務[8][9]。在預訓練的模型當中，EfficientNet 曾以特殊的結構在視覺辨識比賽中獲得佳績，打敗了經典的 ResNet、DenseNet 等模型，使用較小的模型取得優異的性能[10]。

三、深度偽造音訊檢測

由於深度偽造音訊的門檻降低，使得檢測深度偽造音訊的需求逐漸受到重視，在深度偽造音訊資料集的建置上有著名的英文的 ASVSPOOF[11]和中文的 CFAD[12]等資料集。在使用深度學習模型執行檢測深度偽造音訊任務時，有多個步驟可以提升檢測效果：使用預訓練的模型進行遷移式學習，能夠有效提高準確率和 F1 分數等效能指標[13]，若選擇使用真實音訊資料所預訓練的模型，經過微調前端，可以減少在檢測深度偽造音訊的錯誤率[14]，而在輸入預訓練模型前，對於前端特徵提取進行微調能提升性能[15]，如果使用語音情緒辨識所萃取的情緒特徵，能有加強深度偽造音訊的檢測效能[16]，將多個預訓練模型進行融合或改進，能夠加強模型對於新資料集的泛化性[17]，在卷積神經網路加入分類器可以進一步提升模型效能[18]。

參、研究方法

一、實驗環境

本文實驗環境在個人電腦上操作，相關設備資訊如下：

(一) 作業系統（Operating System）：

Windows-11。

(二) 中央處理器（Central Processing Unit，CPU）：

Intel-Pentium-Gold-G6405-4.10GHz。

(三) 隨機存取記憶體（Random-access memory, RAM）：

DDR4-24GB。

(四) 圖形處理器（Graphics Processing Unit，GPU）：

GTX-3060-Ti-8GB。

(五) 整合開發環境（Integrated Development Environment, IDE）：

本研究使用 Anaconda 之下的 Spyder 作為整合開發環境。

(六) 程式語言（Programing Language）：

Python 3.9。Python 是一種直譯式的程式語言，以簡潔的語法和大量的

第三方模組為特色而受到廣泛地使用，且由於其開源特性，許多團體開發了實用的套件並公開存取使用。本研究使用了 Numpy、Pandas、Librosa、Torch、Tensorflow、Sklearn、Keras 及 Matplotlib 等多項函式庫及套件進行資料整理、訓練模型及呈現結果等功能。

二、資料集

(一) ASVSPOOF（Automatic Speaker Verification and Spoofing Countermeasures Challenge）

1. 簡介

ASVSPOOF 是一項針對欺騙攻擊（Spoof Attack）語音所定期舉辦的國際挑戰賽，由 2015 年開始第一屆競賽，核心的挑戰任務是從十多種語音轉換、合成或生成技術產生的語音資料集和真實語音資料集當中，分類出人類語音和欺騙攻擊語音，2017 年競賽新增檢測重播攻擊（Replay Attack）的任務，2019 年競賽拓展了實體存取攻擊資料集數據，2021 年競賽新增了語音深度偽造資料集及識別深度偽造演算法挑戰。ASVSPOOF 挑戰賽致力於推動語音處理系統對抗語音偽造的能力，希望透過全球挑戰者的參與，提出具備大數據處理能力、高度通用性、優良準確率及充分運算效率的分類器或模型，用以研究出防禦偽造語音技術的對策。

2. 邏輯存取（Logical Access）資料集

邏輯存取資料集是 ASVSPOOF 挑戰賽中經過語音轉換和文字生成語音技術所產生的欺騙攻擊資料集，因為沿襲多年的測試和擴增，邏輯存取資料集是該挑戰賽中的穩定性較高的資料集，共有 181566 筆音訊資料，如圖一所示。本文以 ASVSPOOF2021-LA 資料集所附錄之標籤說明文件紀錄順序由第一筆資料往下取 6000 筆，因為該資料集所附錄的標籤說明文件中，前段部分以相較平均的數量標記每 1 名錄音者在 6 種錄音演算法與 13 種欺騙攻擊演算法情境下的排列組合約 1200 筆資料不等，同一錄音者資料中欺騙攻擊標籤與真實音訊標籤比例約為 11 比 1，而後段部分標籤規律不同於前半段，故考慮運算資源限制後取至 6000 筆資料，其中包括 5 名錄音者完整的排列組合的 5901 筆資料和其餘錄音者 99 筆不完整的排列組合資料，總共 6000 筆資料，其中欺騙（Spoof）標籤 5482 筆和真實（Bonafide）標籤 518 筆。

(二) CFAD（A Chinese Dataset for Fake Audio Detection）

1. 簡介

CFAD 是一個用於檢測偽造音訊的標準化中文資料集，由中國科學院自動化研究所的學者們所建置。有鑑於主流偽造音訊的資料集都是以英文建置，而使用中文的國家仍然有中文偽造音訊資料集的需求，故其蒐集了多個中文真實語音資料集，並使用十二種

主流語音生成技術來生成偽造音訊，並加入多種干擾檢測的方法，例如進行不同信噪比的雜訊添加、透過不同編解碼器的音訊格式轉換等，希望模擬現實中不可預期的環境干擾，使得 CFAD 資料集更具挑戰性，也能藉此資料集訓練的模型能有更好的泛化能力。

2. Dev_Clean 資料集

在 CFAD 的資料集中，Dev_Clean 資料集是蒐集了四種中文真實語音資料集並以八種偽造語音技術生成偽造語音後加入，共有 14400 筆音訊資料，如圖二所示，而因運算環境資源限制，以及測試語音資料集中真實語音與偽造語音數量是否平衡化的影響，本文使用 CFAD-Dev_Clean 部分 6000 筆資料進行實驗，其中包含欺騙（Spoof）語音資料隨機抽樣 3001 筆和真實（Bonafide）語音資料隨機抽樣 2999 筆。本研究的資料分布如圖三所示。

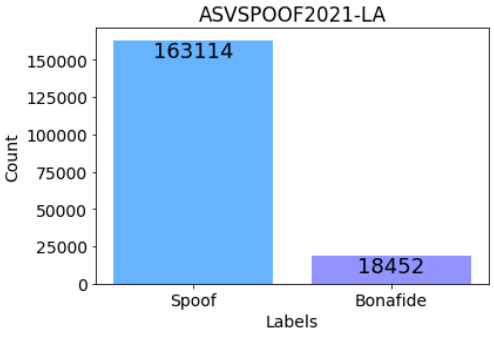
(三) 瑞爾森情緒語言與歌曲視聽資料集（Ryerson Audio-Visual Database of Emotional Speech and Song, RAVDESS）

1. 簡介

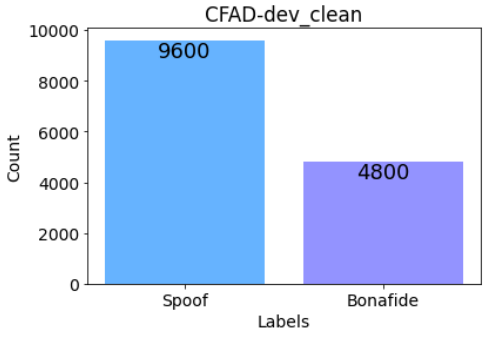
瑞爾森情緒語言與歌曲視聽資料集是一個用於情緒辨識的多模態英文資料集，是由位在加拿大多倫多的瑞爾森大學等校所建置，包含情緒語音和歌曲。瑞爾森情緒語言資料集之中性別分布為經過口音篩選的 12 名男性與 12 名女性，不論男女錄音員均為專業演員，年齡範圍在 21 到 33 歲之間。

2. 資料集

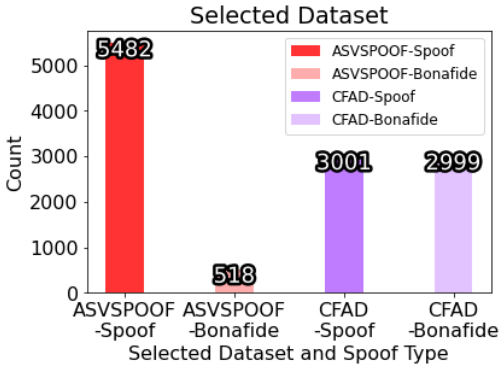
瑞爾森情緒語言與歌曲視聽資料集中共有 8 種情緒類別的音訊檔案，包含冷靜（Calm）音訊檔案 192 筆、開心（Happy）音訊檔案 192 筆、傷心（Sad）音訊檔案 192 筆、生氣（Angry）音訊檔案 192 筆、害怕（Fearful）音訊檔案 192 筆、厭惡（Disgust）音訊檔案 192 筆、驚訝（Surprised）音訊檔案 192 筆及中性（Neutral）音訊檔案 96 筆等共 1440 個情緒音訊檔案，如圖四所示。優點為情緒類別分布平均，但缺點為中性類別檔案數量較少。本文使用全部 1440 個情緒音訊檔案。



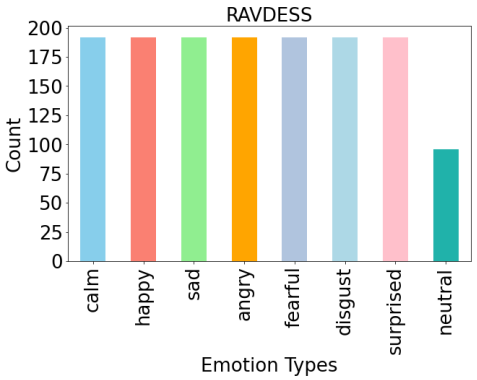
圖一：ASVSPOOF 邏輯存取資料集示意圖



圖二：CFAD 之 Dev_Clean 資料示意圖



圖三：本研究使用之資料集示意圖



圖四：RAVDESS 情緒音訊資料集示意圖

三、語音情緒辨識

(一) 梅爾頻譜圖（Mel-spectrogram）

梅爾頻譜圖，是一種基於聲音特徵的表徵方式，經常使用在語音的辨識與分析，為了模擬人類耳朵的感知，將連續變化的聲音訊號（如圖五所示）透過數學原理轉換特性，並萃取初階的音訊特徵作為特徵向量，方便後續進行觀察分析與機器處理，如圖六所示，其過程中包含下列主要步驟：

1. 預強化（Pre-emphasis）：
   在轉換成梅爾頻譜的初期階段，通常會將音頻訊號進行預強化，透過將音頻訊號通過高通濾波器來達成，用以消除聲帶和嘴唇對音訊帶來的影響，改善信噪比來使後續步驟順利進行。
2. 訊框化（Frame Blocking）：
   使用漢明窗（Hamming Window）或其他函數，將音頻訊號依照時間總長度轉換成數個固定時間長度的訊框，藉此產生出方便機器處理的音訊單位，通常為了穩定訊框之間的連續性，採樣時會將訊框重疊一部份，避免因為訊框交界而失去音頻訊號的重要特徵。
3. 快速傅立葉轉換（Fast Fourier Transform，FFT）：
   將音頻訊號在時域（Time Domain）的變數轉換成頻域（Frequency Domain）上的變數，使得觀察或處理音訊特徵的時候能夠將複雜的音訊轉換成方便計算的頻域特徵，將能量振幅作為每個訊框的特徵指標來進行後續的機器處理步驟。
4. 梅爾濾波器（Mel Filter Bank）：
   旨在消除音訊諧波，並將頻譜進行平滑化的處理。透過輸入頻譜進入一組由 20 個三角帶通濾波器（Triangular Bandpass Filters）組成的梅爾濾波器組，將每個三角帶通濾波器平均對應梅爾頻率（Mel Frequency）刻度上的一段區域，並計算其對數能量，藉以

呈現音頻訊號在不同梅爾頻率上的分布。

5. 結果呈現：

梅爾頻譜圖的 X 軸代表時間，Y 軸代表頻率，顏色則代表在不同時間和頻率上的訊號強度，通常以分貝（dB）為表示單位。

(二) 深度學習（Deep Learning）

深度學習是一種人工的神經網路架構，致力於模擬人類大腦的學習過程，使得人工智慧在對大量輸入的資料進行表徵學習之後，能夠擁有自主學習和特徵提取的能力。深度學習架構包含多層由神經元組成的神經層，包括輸入層、多個隱藏層及輸出層，並透過權重、偏差值和激活函數等操作進行運算。

1. 卷積神經網路

卷積神經網路（Convolutional Neural Network，CNN）是一種深度學習模型，通常用於電腦視覺領域如圖像與影片，起源於模擬生物接收視覺訊號的過程，對視野中刺激進行相互重疊來感知，而卷積神經網路由包含一層或多層卷積層的神經網路架構所組成，每層神經網路層都包含多個神經元。通常具有下列神經網路層：

(1) 卷積層（Convolutional Layer）：

卷積層的目的是提取圖像的特徵，是卷積神經網路的核心部分，使用一組卷積核在所輸入的圖像上滑動，透過對局部圖像執行加權運算，進而捕捉圖像的邊緣、紋理等特徵，將複雜的細節簡化為主要的特徵，並據以產生特徵圖。

(2) 池化層（Pooling Layer）：

池化層的目的是減少特徵圖的參數並且保留重要特徵，包含最大池化層（Max Pooling）、平均池化層（Average Pooling）。最常見為使用最大池化層緊接於卷積層之後，將輸入的特徵圖劃分為數個區域，並從每個子區域中取最大值保留作為特徵，減少了參數數量也避免過度擬合。

(3) 線性整流層（Rectified Linear Units Layer，ReLU Layer）：

線性整流層用途為判定並整理神經元的輸出，將所有負數輸入都映射為零，而正數輸入則保持原狀，能夠減緩梯度消失的問題，也有利了後續的反向傳播和學習計算。

(4) 完全連接層（Fully Connected Layer）：

完全連接層也稱作密集層（Dense Layer），用途為將輸入的特徵映射到最終的輸出結果之上，每個神經元上都有權重（Weights）和偏差值（Bias），權重用來調整輸入特徵的重要程度，偏差值用於增強特徵的活躍性及靈活性。

2. EfficientNet

EfficientNet 的設計目的是提高模型效能同時兼顧模型的預測準確

率，是 Google Brain 團隊所提出的一系列卷積神經網路模型，以其獨特的複合縮放方法而聞名，在 2019 年的 ImageNet 大規模視覺辨識挑戰賽（ImageNet Large Scale Visual Recognition Challenge，ILSVRC）中表現優異因而受到矚目。這個方法產生了從 B0 到 B7 共八種不同大小的模型，讓使用者能根據自身資源限制和效能需求來做選擇，其特色如下：

(1) 複合縮放（Compounding Scaling）：
通過同時調整神經網路的深度（神經層數）、寬度（神經元通道數）和解析度（輸入圖像大小）來實現整體模型的平衡，確保在提高準確率的過程中，兼顧運算參數的數量以避免模型的肥大對模型效能的負面影響。

(2) ImageNet 大規模視覺辨識挑戰賽表現優異：
ImageNet 大規模視覺辨識挑戰賽使用 ImageNet 數據集，其中包含 1000 個不同分類，總共約 1400 萬張已標記的圖像，參賽者任務為訓練模型以識別圖像中的目標並進行正確的分類。而 EfficientNet 在保持高度準確性的同時，減少模型所需的參數，相較於同期的其他模型消耗更少的運算資源。

(3) 一系列的模型規模可供選擇：
EfficientNet 提供了 B0、B1、B2、B3、B4、B5、B6 及 B7 等八種不同規模的模型可供套用，每種模型都在 ImageNet 大規模視覺辨識挑戰賽上受測具有良好的效能，使用者可以根據運算資源和運算任務自行選擇，從移動設備所需的輕量級運算模型到高精度圖像分類的大規模運算都能符合需求。

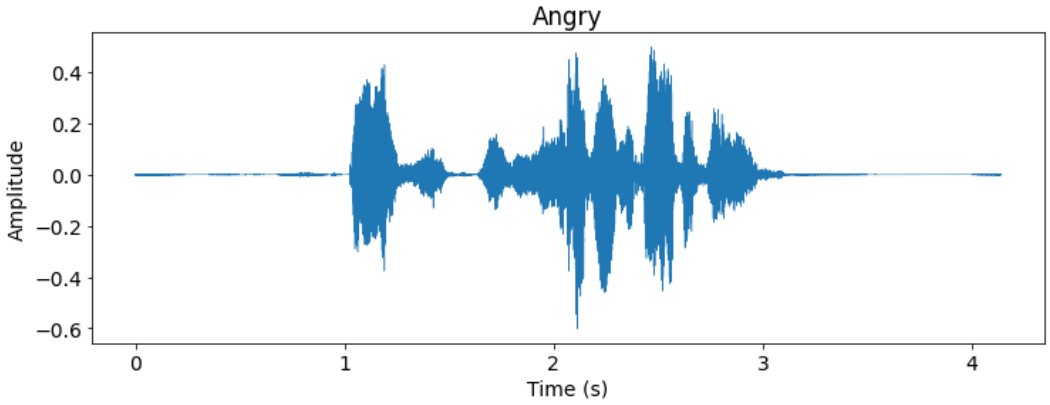
(三) 遷移式學習（Transfer Learning）

遷移式學習是一種儲存並利用模型的機器學習方法，通常應用在與儲存模型原任務目標相關的領域上，因而能提高模型的問題解決能力。遷移式學習的核心概念是將來源領域學習到的知識，透過模型的權重及參數方式進行儲存，並應用到目標領域進行適應學習，通常在來源領域和目標領域越相似的情形之下，模型的適應性和準確率會越高。遷移式學習的主要特色如下：

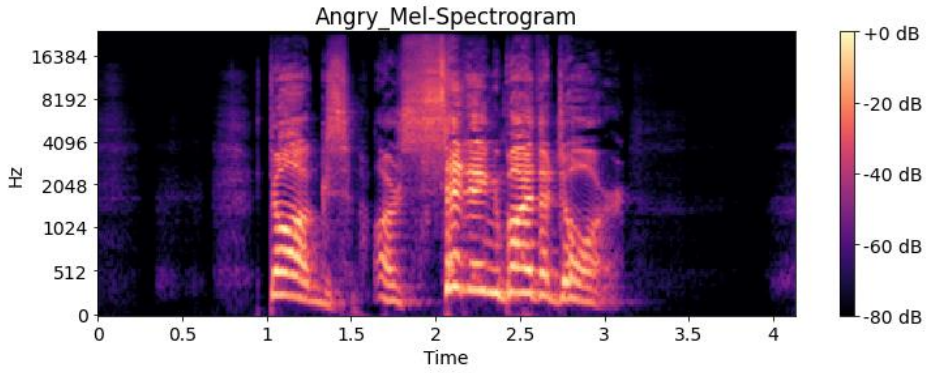
1. 預訓練模型（Pre-trained Model）：
已經在大規模數據集上進行訓練的模型，因為在來源領域上學習到了豐富的知識及特徵表示，具有良好的泛化能力，能夠在目標領域上提供已訓練的基礎權重，所以能夠節省運算資源同時追求更高的準確率，例如在自然語言處理（Natural Language Processing，NLP）領域裡的 BERT（Bidirectional Encoder Representations from Transformers）預訓練模型是 Google 公司使用大量文字語料庫進行預訓練並公開，受到學者們高度使用及良好評價。

2. 微調（Fine-tuning）：

微調的目的是將預訓練模型適應在目標領域的任務之上，做法是載入預訓練模型的權重與參數，視目標任務需求刪改或新增預訓練模型的神經層，然後將目標領域的資料集輸入到模型中進行額外的訓練，微調神經元參數，通常越相似的領域則調整速度越快，直到模型能夠萃取目標領域的任務特徵，解決新的任務目標。



圖五：聲音訊號示意圖



圖六：梅爾頻譜圖

四、深度偽造音訊檢測

(一) 梅爾頻率倒譜係數

梅爾頻率倒譜係數（Mel-Frequency Cepstral Coefficients，MFCCs）通常用於音訊處理及語音辨識的領域之中，是一種基於音訊頻譜特徵的特徵表示方式，通過一系列對於音訊頻率的線性變換，將音訊原始檔案轉換成機器學習模型能夠理解並處理的特徵向量。音訊原始檔案轉換成特徵向量的過程中包含下列主要步驟：

1. 預強化（Pre-emphasis）：

在轉換成梅爾頻率倒譜係數的初期階段，通常會將音頻訊號進行預強化，透過將音頻訊號通過高通濾波器來達成，用以消除聲帶

和嘴唇對音訊帶來的影響，改善信噪比來使後續步驟順利進行。

2. 訊框化（Frame Blocking）：
   使用漢明窗（Hamming Window）或其他函數，將音頻訊號依照時間總長度轉換成數段固定時間長度的訊框，藉此產生出方便機器處理的音訊單位，通常為了穩定訊框之間的連續性，採樣時會將訊框重疊一部份，避免因為訊框交界而失去音頻訊號的重要特徵。
3. 快速傅立葉轉換（Fast Fourier Transform，FFT）：
   將音頻訊號在時域（Time Domain）的變數轉換成頻域（Frequency Domain）上的變數，使得觀察或處理音訊特徵的時候能夠將複雜的音訊轉換成方便計算的頻域特徵，將能量振幅作為每段訊框的特徵指標來進行後續的機器處理步驟。
4. 梅爾濾波器（Mel Filter Bank）：
   旨在消除音訊諧波，並將頻譜進行平滑化的處理。透過輸入頻譜進入一組由 20 個三角帶通濾波器（Triangular Bandpass Filters）組成的梅爾濾波器組，將每個三角帶通濾波器平均對應梅爾頻率（Mel Frequency）刻度上的一段區域，並計算其對數能量，藉以呈現音頻訊號在不同梅爾頻率上的分布。
5. 離散餘弦轉換（Discrete Cosine Transform， DCT）：
   由於倒頻譜（Cepstrum）在展現人聲和音樂訊號的特徵上有良好的效果，所以為了將目前取得的梅爾頻率上分布的對數能量頻譜轉換成倒頻譜，將梅爾濾波器所得到的一組對數能量，進行離散於弦轉換，求出梅爾頻率倒譜係數，而每一段訊框通常只會保留 13 個梅爾頻率倒譜係數，故時間長度越長的音訊會轉換出更多訊框，並最終轉換出更多梅爾頻率倒譜系係數。

(二) XGBoost（eXtreme Gradient Boosting）

XGBoost 是一種用於預測結果的梯度提升演算法，核心概念是以集成學習的方法，結合多個決策樹來建構出效能更強大的預測模型。最初版本在 2014 年被開發，後於 2019 年在 Higgs 機器學習挑戰獲勝後受到矚目，當前版本 2. 0. 2 於 2023 年更新，作為一個開源框架，XGBoost 為 C++、Java、Python、R 和 Julia 等程式語言提供開源軟體包實現，其主要特色說明如下：

1. 梯度提升（Gradient Boosting）演算法
   是一種集成學習演算法，透過組合並迭代地訓練一系列弱學習器，進而建構出一個總體強大的預測模型，該預測模型能夠透過弱學習器的逐步改進，由底層往上修正錯誤，逐漸最小化損失函數，追求最高的模型效能，同時具有良好的泛化能力。
2. 正則化（Regularization）
   正則化是一種降低機器學習過度擬合發生的訓練方法，核心概念

是通過對損失函數進行添加項目，幫助模型限制參數的數值範圍，防止模型過度膨脹並降低複雜度，進而提高泛化能力，防止過度擬合而在未見過的資料集上無法發揮預期的效能。常見的正則化為 L1（Lasso）正則化和 L2（Ridge）正則化。

五、實驗流程

(一) 深度偽造音訊資料集

1. ASVSPOOF 英文資料集

本文使用 ASVSPOOF2021-LA 部分 6000 筆資料進行實驗，其中包含欺騙語音資料 5482 筆和真實語音資料 518 筆。先讀取資料集之 FLAC 編碼音訊檔案，並整理、匹配音訊檔案之欺騙與真實標籤，將音訊檔案分別轉換成梅爾頻率倒譜係數，但因為各個音訊檔案錄製時間長度不同導致梅爾頻率倒譜係數結果不一致，故將特徵係數總量較少者採取零填充（Zero-padding）的方式將係數總量補齊。本研究因為運算資源限制，將時間長度最長的資料取 1300 個梅爾頻率倒譜係數作為上限，並將其餘因時間長度較短造成特徵係數較少者採取零填充方式補齊至 1300 個。

2. CFAD 中文資料集

本文使用 CFAD-Dev_Clean 部分 6000 筆資料進行實驗，其中包含欺騙語音資料 3001 筆和真實語音資料 2999 筆。先讀取資料集之 WAV 編碼音訊檔案，並整理、匹配音訊檔案之欺騙與真實標籤，將音訊檔案分別轉換成梅爾頻率倒譜係數，並比照 ASVSPOOF 英文資料集將時間長度最長者取 1300 個梅爾頻率倒譜係數，其餘採取零填充方式補齊至 1300 個。

(二) 基準系統

1. 深度偽造音訊檢測

(1) XGBoost 分類器

我們將深度偽造音訊資料集所轉換成的梅爾頻率倒譜係數特徵和其所匹配的欺騙、真實標籤作為本階段用於 XGBoost 分類器的輸入資料，並將輸入資料依照比例切分八成作為訓練資料集、兩成作為驗證資料集。在 XGBoost 分類器的相關參數的調整如下：決策樹數量（n_estimators）設為 200，以較龐大的模型規模來追求更高的準確率和泛化能力；隨機種子編號（ransom_state）設為 11，以固定訓練過程來實現可複製性；決策樹建構演算法（tree_method）設為直方圖算法（hist），用於大量數據時可加速決策樹的建構過程；運算設備（device）設為圖形處理器（cuda），將主要運算設備由中央處理器轉移到圖形處理器以加速訓練過程，且由於圖形處理器的資源不足，所以造成採取樣本時的限制。

(2) 效能評估

對於 XGBoost 分類器的最終結果進行整理，我們獲得了真陽性、偽陰性、偽陽性、真陰性的實際數量作為混淆矩陣，並據以計算出準確率、精確度、召回率、F1 分數等評估指標，最後繪製接收者操作特徵曲線，並計算曲線下面積。

(三) 增強系統

1. 語音情緒辨識

(1) RAVDESS 英文資料集

本文使用瑞爾森情緒語言與歌曲視聽資料集中情緒語言部分 1440 筆資料進行實驗，共有 8 種情緒類別的情緒語音，其中包含冷靜（Calm）音訊檔案 192 筆、開心（Happy）音訊檔案 192 筆、傷心（Sad）音訊檔案 192 筆、生氣（Angry）音訊檔案 192 筆、害怕（Fearful）音訊檔案 192 筆、厭惡（Disgust）音訊檔案 192 筆、驚訝（Surprised）音訊檔案 192 筆及中性（Neutral）音訊檔案 96 筆等。先讀取資料集之 WAV 編碼音訊檔案，並整理、匹配音訊檔案之情緒類別標籤，將音訊檔案分別轉換成梅爾頻譜圖以利後續步驟。

(2) EfficientNet 遷移式學習

取得 RAVDESS 英文資料集所生成的梅爾頻譜圖和相對應的情緒類別標籤之後，我們接著使用辨識圖像能力優異的 EfficientNet 模型來進行遷移式學習，基於訓練速度考量，選擇了模型尺寸最小的 EffcientNet-B0 模型，該模型所需之輸入格式為寬度 224 像素乘以高度 224 像素乘以顏色 3 原色的圖像，但是上一步驟產出的梅爾頻譜圖的寬度和高度均大於輸入格式且並寬高比並非 1 比 1，故需將每張梅爾頻譜圖的尺寸縮小至輸入格式後方能輸入 EffcientNet-B0 模型進行訓練，於模型訓練之前，載入預訓練的 ImageNet 權重後鎖定神經層的參數訓練，以加速進行遷移式學習，並於微調過程中在模型輸出之前加入兩層神經層：全局平均池化層（GlobalAveragePooling2D Layer）降低模型的維度並保留主要特徵；密集層（Dense Layer）及 Softmax 激勵函數來將特徵轉換並對應最終的八種情緒類別，儘管 ImageNet 的原型是用於進行 1000 個不同類別圖像的分類任務，有別於本研究此處的 8 種情緒類別之圖像分類任務，但是其經過大量資料的訓練後的優異圖像分類能力，在遷移式學習之後，仍有助於進行 8 種情緒類別的圖像分類任務。編譯模型時相關設定如下：優化器（Optimizer）設為 Adam 自適應隨機梯度下降演算法，用動態方式調整模型的學習率來獲得更好的學習成效；

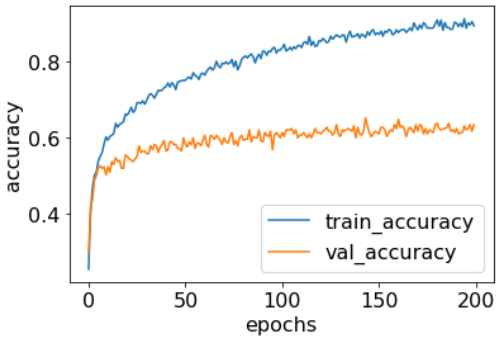
損失函數（Loss Function）設定選擇為分類標籤交叉熵（Categorical_crossentropy）函數，適合衡量多類別分類問題；監測指標（Metrics）設為準確率（Accuracy），常用於分類問題模型的性能指標。

(3) 效能評估

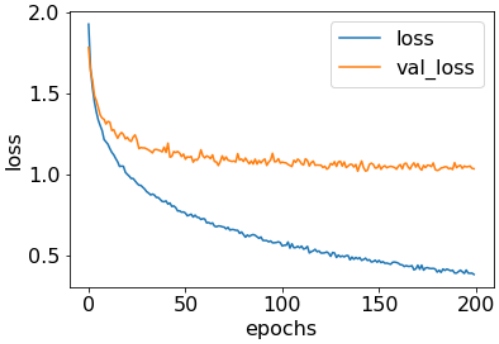
將輸入之梅爾頻譜圖和相對應的情緒類別標籤，依照比例切分八成作為訓練資料集、兩成作為驗證資料集，訓練批次大小（Batch_size）設定為 32 以加速訓練過程，經過 200 輪（Epochs）訓練後：訓練準確率為 0.894、驗證準確率為 0.635，如圖七所示；訓練損失值為 0.380、驗證損失值為 1.033，如圖八所示。確認結果後將模型儲存備用。

(4) 情緒特徵萃取

經過訓練後的語音情緒辨識模型，可以將輸入的音訊檔案進行情緒特徵萃取，再將萃取出的情緒特徵於最後一層進行情緒分類，表示最後一層會接收到最終的情緒特徵，而為了繼續使用該情緒特徵並減少資料量，我們把輸出層由情緒分類任務的 Softmax 密集層改成能夠保留 256 個情緒特徵的 ReLU 密集層。



圖七：語音情緒辨識模型準確率示意圖



圖八：語音情緒辨識模型損失函數示意圖

2. 深度偽造音訊檢測

(1) XGBoost 分類器

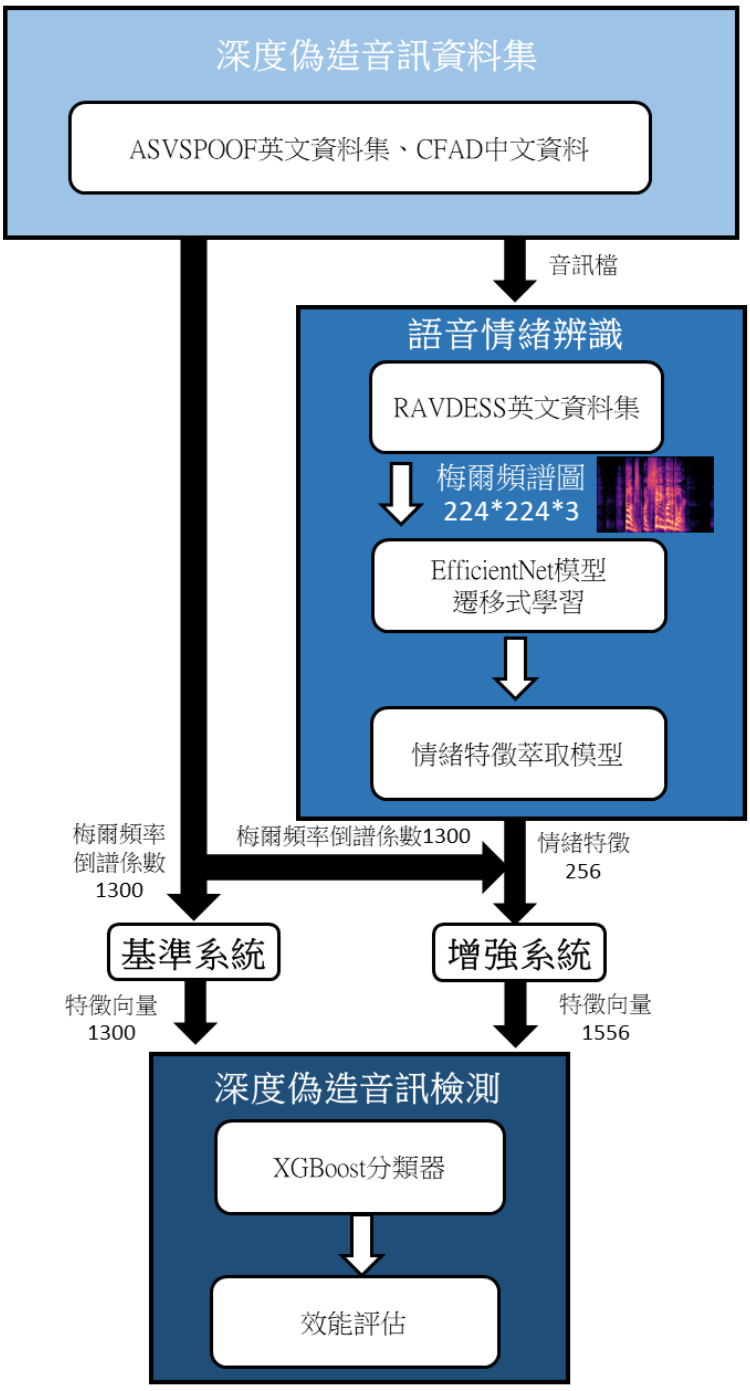
我們將深度偽造音訊資料集中每一筆音訊分別進行梅爾頻率倒譜係數特徵的轉換，及進行語音情緒辨識萃取出 256 個情緒特徵，再將梅爾頻率倒譜係數特徵直接和 256 個情緒特徵合併在一起，使得每一筆音訊都成為一組特徵向量。每一組特徵向量再和其所匹配的欺騙、真實標籤作為本階段用於 XGBoost 分類器的輸入資料，其餘設定值同基準系統。

(2) 效能評估

對於 XGBoost 分類器的最終結果進行整理，我們獲得了真陽性、偽陰性、偽陽性、真陰性的數值作為混淆矩陣，並據以計算準確率、精確度、召回率、F1 分數等評估指標，最後繪製接收者操作特徵曲線，並計算接收者操作特徵曲線下面積。

(四) 實驗流程圖

實驗流程分為兩個系統，如圖九所示：



圖九：實驗流程圖

肆、結果與討論

一、實驗結果評估指標

本實驗使用準確率、精確率、召回率、F1 分數及接收者操作特徵曲線下面積等數據作為評估模型效能的指標，混淆矩陣中 Spoof 代表為欺騙、bonafide 代表為真實，各項數據之計算方式如下：

(一)資料標籤分類為真且模型分類預測為真：真陽性（True Positive，TP）。

(二)資料標籤分類為真且模型分類預測為假：偽陰性（False Negative，FN）。

(三)資料標籤分類為假且模型分類預測為真：偽陽性（False Positive，FP）。

(四)資料標籤分類為假且模型分類預測為假：真陰性（True Negative，TN）。

(五)真陽性率(True Positive Rate, TPR)：

$$TPR = TP/(TP + FN) \tag{1}$$

(六)偽陽性率(False Positive Rate, FPR)：

$$FPR = FP/(FP + TN) \tag{2}$$

(七)準確率（Accuracy）：

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{3}$$

(八)精確率(Precision)：

$$Precision = TP/(TP + FP) \tag{4}$$

(九)召回率(Recall)：

$$Recall = TP/(TP + FN) \tag{5}$$

(十)F1 分數(F1_score)：

$$F1\_score = 2/(1/Precision + 1/Recall) \tag{6}$$

(十一)接收者操作特徵曲線（Receiver Operating Characteristic Curve, ROC Curve）：由實驗結果的真陽性、偽陰性、偽陽性和真陰性數值來計算真陽性率與偽陽性率後，以偽陽性率為 X 軸、真陽性率為 Y 軸來繪製接收者操作特徵曲線。

(十二)接收者操作特徵曲線下面積（Area Under the Curve of ROC, AUC）：AUC 是計算接收者操作特徵曲線下方的面積，其數值必在 0~1 之間，通常 AUC 數值越高之模型效能越好。

二、實驗結果

依照實驗流程操作之後，深度偽造語音檢測系統實驗結果彙整表如表一所示，各項實驗數據臚列如下：

(一) ASVSPOOF

1. 基準系統

使用英文深度偽造音訊資料集 ASVSPOOF 在基準系統 XGBoost 分類器的檢測效能如下：準確率為 0.9741、精確度為 0.8863、召回率為 0.8111、F1 分數為 0.8470、接收者操作特徵曲線下面積為 0.9006。混淆矩陣中，真陽性為百分之 7.17、偽陰性為百分之 1.67、偽陽性為百分之 0.92、真陰性為百分之 90.25。接收者操作特徵曲線圖如圖十所示，混淆矩陣如圖十三所示。

2. 增強系統

使用英文深度偽造音訊資料集 ASVSPOOF 在增強系統即語音情緒辨識系統加上 XGBoost 分類器的檢測效能如下：準確率為 0.9800、精確度為 0.9275、召回率為 0.8394、F1 分數為 0.8812、接收者操作特徵曲線下面積為 0.9166。真陽性為百分之 7.42、偽陰性為百分之 1.42、偽陽性為百分之 0.58、真陰性為百分之 90.58。接收者操作特徵曲線圖如圖十所示，混淆矩陣如圖十四所示。

(二) CFAD

1. 基準系統

使用中文深度偽造音訊資料集 CFAD 在基準系統 XGBoost 分類器的檢測效能如下：準確率為 0.9800、精確度為 0.9915、召回率為 0.9686、F1 分數為 0.9799、接收者操作特徵曲線下面積為 0.9801。真陽性為百分之 48.75、偽陰性為百分之 1.58、偽陽性為百分之 0.42、真陰性為百分之 49.25。接收者操作特徵曲線圖如圖十一所示，混淆矩陣如圖十五所示。

2. 增強系統

使用中文深度偽造音訊資料集 CFAD 在增強系統即語音情緒辨識系統加上 XGBoost 分類器的檢測效能如下：準確率為 0.9858、精確度為 0.9900、召回率為 0.9817、F1 分數為 0.9858、接收者操作特徵曲線下面積為 0.9859。真陽性為百分之 49.42、偽陰性為百分之 0.92、偽陽性為百分之 0.50、真陰性為百分之 49.17。接收者操作特徵曲線圖如圖十一所示，混淆矩陣如圖十六所示。

(三) ASVSPOOF 混合 CFAD 後

1. 基準系統

使用英文深度偽造音訊資料集 ASVSPOOF 加上中文深度偽造音訊資料集 CFAD 在基準系統 XGBoost 分類器的檢測效能如下：準確率為 0.9750、精確度為 0.9684、召回率為 0.9433、F1 分數為

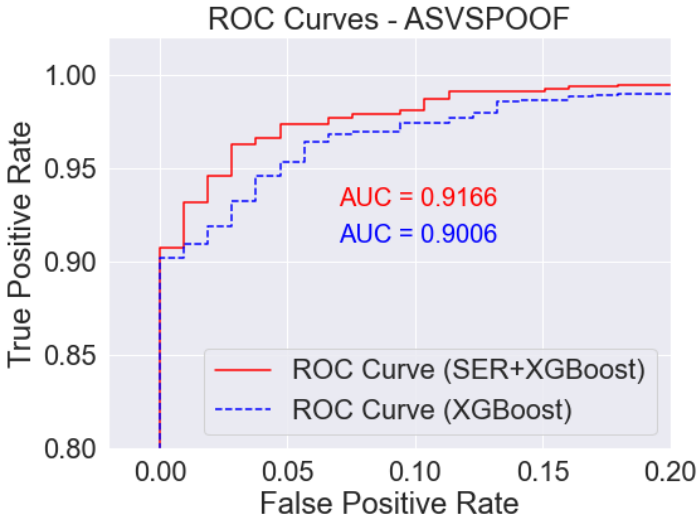
0.9557、接收者操作特徵曲線下面積為 0.9654。真陽性為百分之 26.96、偽陰性為百分之 1.62、偽陽性為百分之 0.88、真陰性為百分之 70.54。接收者操作特徵曲線圖如圖十二所示，混淆矩陣如圖十七所示。

2. 增強系統

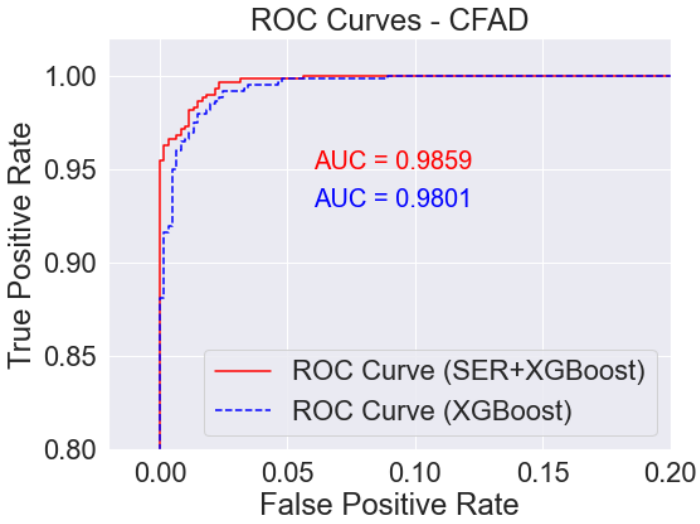
使用英文深度偽造音訊資料集 ASVSPOOF 加上中文深度偽造音訊資料集 CFAD 在增強系統即語音情緒辨識系統加上 XGBoost 分類器的檢測效能如下：準確率為 0.9804、精確度為 0.9747、召回率為 0.9563、F1 分數為 0.9654、接收者操作特徵曲線下面積為 0.9732。真陽性為百分之 27.33、偽陰性為百分之 1.25、偽陽性為百分之 0.71、真陰性為百分之 70.71。接收者操作特徵曲線圖如圖十二所示，混淆矩陣如圖十八所示。

表一：深度偽造語音檢測系統實驗結果彙整表

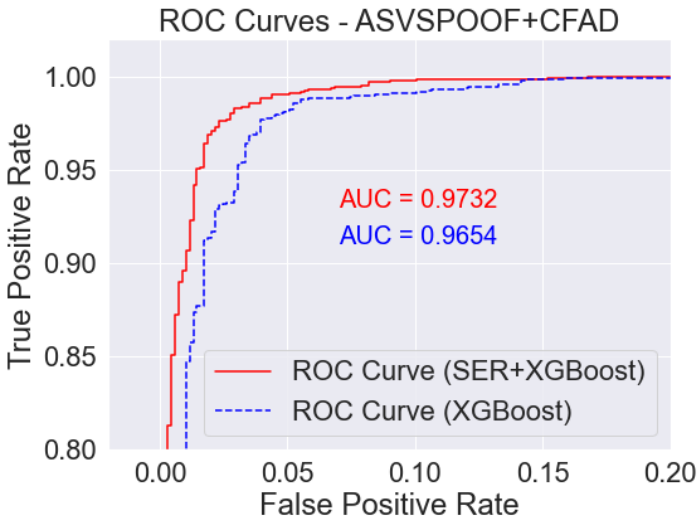
深度偽造音訊資料集	深度偽造音訊檢測系統	準確率（Accuracy）	精確度（Precision）	召回率（Recall）	F1分數（F1_score）	ROC下面積（AUC）
ASVSPOOF	基準系統（XGBoost）	0.9741	0.8863	0.9111	0.8470	0.9006
	增強系統（SER+XGBoost）	0.9800	0.9275	0.8394	0.8812	0.9166
CFAD	基準系統（XGBoost）	0.9800	0.9915	0.9686	0.9799	0.9801
	增強系統（SER+XGBoost）	0.9858	0.9900	0.9817	0.9858	0.9859
ASVSPOOF 混合CFAD後	基準系統（XGBoost）	0.9750	0.9684	0.9433	0.9557	0.9654
	增強系統（SER+XGBoost）	0.9804	0.9747	0.9563	0.9654	0.9732



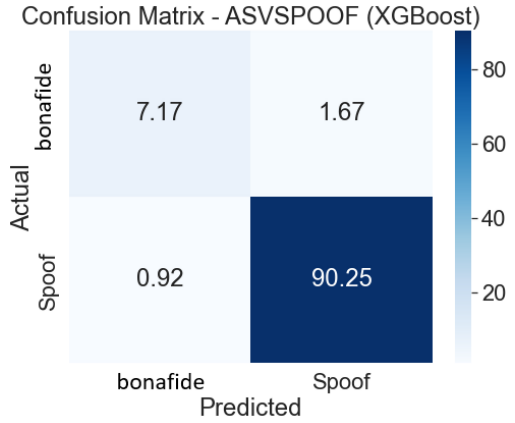
圖十：AVSPOOF 接收者操作特徵曲線圖



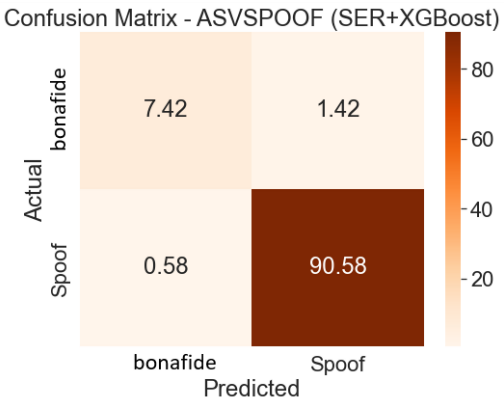
圖十一：CFAD 接收者操作特徵曲線圖



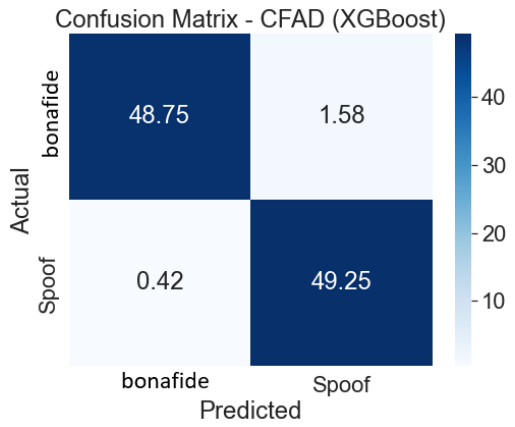
圖十二：AVSPOOF 混合 CFAD 後接收者操作特徵曲線圖



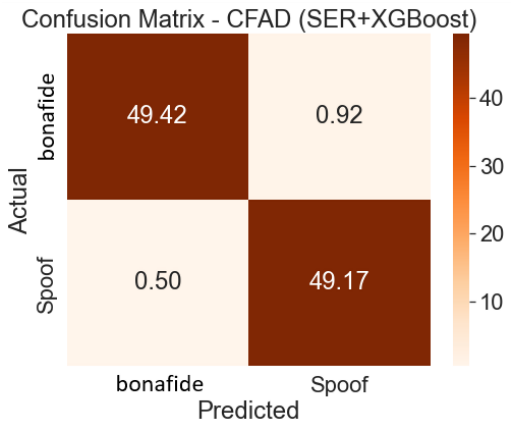
圖十三：AVSPOOF 基準系統混淆矩陣圖



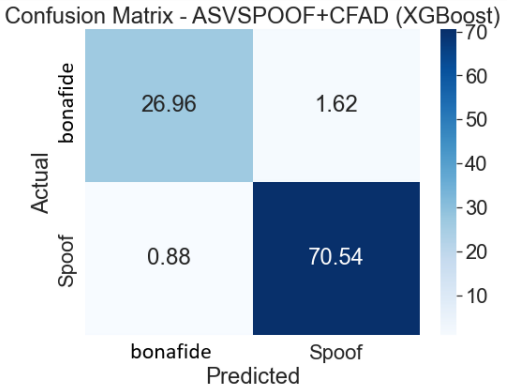
圖十四：AVSPOOF 增強系統混淆矩陣圖



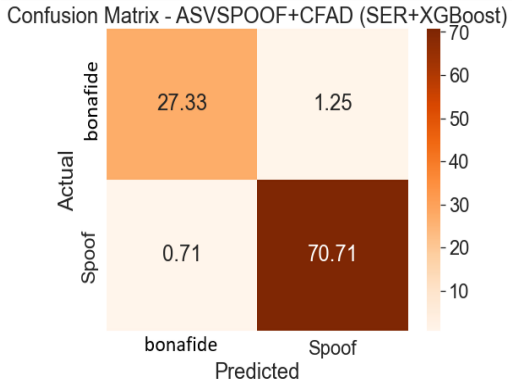
圖十五：CFAD 基準系統混淆矩陣圖



圖十六：CFAD 增強系統混淆矩陣圖



圖十七：ASVSPOOF 混合 CFAD 後基準系統混淆矩陣圖



圖十八：ASVSPOOF 混合 CFAD 後增強系統混淆矩陣圖

三、結果討論

(一) ASVSPOOF

1. 基準系統

從準確率 0.9742 可以看出在基準系統中 XGBoost 分類器表現良好，能夠將 ASVSPOOF 英文資料集中大部分的深度偽造音訊正確分類為 Spoof 或 bonafide。從接收者操作特徵曲線下面積 0.9006 已經很接近上限的數值 1 來看，基準系統在達成 ASVSPOOF 英文資料集的深度偽造音訊檢測的任務中是不錯的系統。值得注意的是，雖然所選擇的 ASVSPOOF 英文資料集欺騙音訊與真實音訊的標籤數量不平衡，但是基準系統的 XGBoost 分類器檢測效能仍然很好。

2. 增強系統

ASVSPOOF 基準系統加上語音情緒辨識系統後成為增強系統，各項評估指標均有上升：準確率上升值達 0.59%、精確度上升值達 4.12%、召回率上升值達 2.83%、F1 分數上升值達 3.42%、接收者操作特徵曲線下面積上升值達 1.6%。接收者操作特徵曲線也能看

出增強系統曲線已靠往左上角最佳模型。我們推斷是因為真實語音相較具有更多情緒，而深度偽造音訊則是像情緒較少的機器人。其次，因為語音情緒辨識系統使用英文的瑞爾森情緒語言與歌曲視聽資料集的因素，導致同樣為英文的 ASVSPOOF 深度偽造音訊資料集的增強系統上有較多的接收者操作特徵曲線下面積增幅。

(二) CFAD

1. 基準系統

實驗結果顯示 CFAD 基準系統的效能良好，而值得注意的是，我們所選擇的 CFAD 中文資料集欺騙音訊與真實音訊的標籤數量幾近平衡，但是跟 ASVSPOOF 英文資料集所呈現的準確率差異並不大。

2. 增強系統

CFAD 基準系統加上語音情緒辨識系統後成為增強系統，雖然準確率、召回率、F1 分數和 ROC 下面積的數值略為提升，但是精確度卻是下降值達 0.15%。而在接收者操作特徵曲線也能看出增強系統曲線已靠往的左上角最佳模型，但是幅度略少於 ASVSPOOF 增強系統。我們認為是語音情緒辨識系統使用英文的瑞爾森情緒語言與歌曲視聽資料集的因素，導致語言為中文的 CFAD 深度偽造音訊資料集的增強系統上有較少的接收者操作特徵曲線下面積增幅。

(三) ASVSPOOF 混合 CFAD 後

1. 基準系統

從各項評估指標中可以看出將 ASVSPOOF 英文資料集和 CFAD 中文資料集所混合的深度偽造音訊資料集所得到的實驗結果介於單獨使用 ASVSPOOF 英文資料集和單獨使用 CFAD 中文資料集的基準系統實驗結果之間。

2. 增強系統

相似於基準系統，混合後的資料集在增強系統所獲得的實驗結果一樣介於單獨使用 ASVSPOOF 英文資料集或 CFAD 中文資料集的實驗結果之間。

伍、結論與建議

本文旨在研究以語音情緒辨識系統來增強深度偽造音訊檢測系統的可行性，實驗操作的深度偽造音訊資料集為 ASVSPOOF 英文資料集和 CFAD 中文資料集，語音情緒識別資料集為瑞爾森情緒語言與歌曲視聽資料集，實驗流程分為兩組系統：基準檢測系統是將深度偽造音訊資料庫的音訊檔案轉換為梅爾頻率倒譜係數後，匹配相對應的欺騙或真實標籤以輸入 XGBoost 分類器；增強系統是先

以瑞爾森情緒語言與歌曲視聽資料集輸入預訓練的 EfficientNet 模型進行遷移式學習，訓練出可以萃取情緒特徵的語音情緒辨識模型，其後將基準系統輸入 XGBoost 分類器之前的梅爾頻率倒譜係數加上所萃取的語音情緒特徵，再進行分類任務。實驗結果顯示：因為加入語音情緒辨識形成增強系統而提升了大部分深度偽造音訊檢測的評估指標，但是在 CFAD 中文資料集的實驗中卻得到精確度下降的結果。而在本研究中也發現以英文資料集所訓練出的語音情緒辨識系統應用在英文的深度偽造音訊檢測系統能有較好的增強效果，以及即使欺騙標籤音訊和真實標籤音訊的數量不平衡，對於 XGBoost 分類器的檢測效能所造成的影響並不大。

建議未來研究首先可以針對語音情緒辨識的泛用性進行發展，因為本研究於語音情緒辨識的效能評估時的驗證準確率為 0.635，仍有優化空間，所以可能透過了解語音情緒辨識與深度偽造音訊檢測之間的泛用性來優化模型；其次，研究不同語言資料集所訓練的語音情緒辨識模型與深度偽造音訊檢測模型之間的關聯性。

參考文獻

[1] P. S. Tomar, K. Mathur and U. Suman, “Unimodal approaches for emotion recognition: A systematic review,” *Cognitive Systems Research*, 2023, 77, pp. 94-109.

[2] 彭正鈦，2021，*中文語音文字線上情緒辨識及應用*，國立臺北科技大學資訊工程系碩士班碩士論文。

[3] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang and W. Zhang, “A systematic review on affective computing: emotion models, databases, and recent advances,” *Information Fusion*, 2022, pp. 19-52.

[4] Y. B. Singh, S. Goel, “A systematic literature review of speech emotion recognition approaches,” *Neurocomputing*, 2022, vol. 492, pp. 245-263.

[5] S. R. Livingstone, F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLoS ONE, 2018, vol. 13(5).

[6] S. Jothimani, K. Premalatha, “MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network,” *Chaos, Solitions and Fractals*, 2022, vol. 162, 112512.

[7] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," *IEEE Signal Processing Letters*, 2018, vol. 25, no. 10, pp. 1440-1444.

[8] S. Padi, S. O. Sadjadi, R. D. Sriam and D. Manocha, “Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation,” *Proceedings of the 2021 International Conference on Multimodal Interaction*,

2021, pp. 645-652.

[9] N. Vryzas, L. Vrysis, R. Kotsalis and C. Dimoulas, “A web crowdsourcing framework for transfer learning and personalized Speech Emotion Recognition, “ *Machine Learning with Applications*, 2021, vol. 6, 100132.

[10] M. Tan, Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” *International Conference on Machine Learning*, 2019, pp. 6105-6114.

[11] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 2023, vol. 31, pp. 2507-2522.

[12] H. MA, J. Yi, C. Wang, X. Yan, J. Tao, T. Wang, S. Wang and R. Fu, “CFAD: A Chinese Dataset for Fake Audio Detection,” arXiv:2207.12308, 2023.

[13] Bonvang, G. A.P., 2023, *Listening Beyond Words: Transfer Learning for Audio Deepfake Detection*, Aalborg University master’s thesis.

[14] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, N. Evans, “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112-119.

[15] L. Li, T. Lu, X. Ma, M. Yuan and D. Wan, “Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT,” *Applied Sciences*, 2023, vol. 13(14), 8488.

[16] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, S. Tubaro, "Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8962-8966.

[17] X. Wang, J. Yamagishi, “A Practical Guide to Logical Access Voice Presentation Attack Detection,” arXiv:2201.03321, 2022.

[18] S. Ahlawat, A. Choudary, “Hybrid CNN-SVM Classifier for Handwritten Digit Recognition,” *Procedia Computer Science*, 2020, vol. 167, pp. 2554-2560.