中原大學

電機資訊學院學士班

專題報告

基於機器學習的鋼琴指法預測

組長 電資四 11020107 蘇伯勳

組員 電資四 11020143 吳柏賢

組員 電資四 11020137 關翔謙

指導教授 莊秀敏 教授

2024/11/16

摘要

標記合適的鋼琴指法對於演奏者而言至關重要，不正確的指法不僅會導致錯誤的音色與詮釋，更嚴重甚至會對手指造成傷害。要讓使用者快速且正確標記出合適的指法需要大量的經驗累積，而若能透過機器學習建立出合宜的指法標記模型，不僅對剛開始學習的學生有幫助，對於演奏者而言也會多出幾種不同的演奏方式的選擇。本專題之目的為對讀入的 MIDI 鋼琴樂曲檔案預測出每個音符對應的合理指法。我們建立了兩種 RNN 模型，並使用不同來源的資料進行驗證與比較。

關鍵詞

Automatic Piano Fingering, Music Processing, MIDI Music

目錄

圖目錄

表目錄

第壹章、緒論

第一節、研究動機

在當今網路媒體中，不同於傳統的露臉網紅，近年新出現的新形態虛擬直播主(Virtual YouTuber)隨著網紅偶像化的潮流會推出與角色相應的原創曲。而我們的小組組長身為一個鋼琴演奏者、編曲者，同時也是虛擬網紅的粉絲，欲將他們的原創曲之鋼琴改編曲套用虛擬直播主的人物模型產生演奏動畫。而在這個目標的最重要環節就是如何透過鋼琴樂曲取得人物模型對應的合理手指動作。並在身為鋼琴演奏者時，在練習新樂曲的第一件事就是先觀察並標記出合理的指法以便演奏，因此我們便想要透過機器學習的方式探索自動標記合理鋼琴指法的可能性。

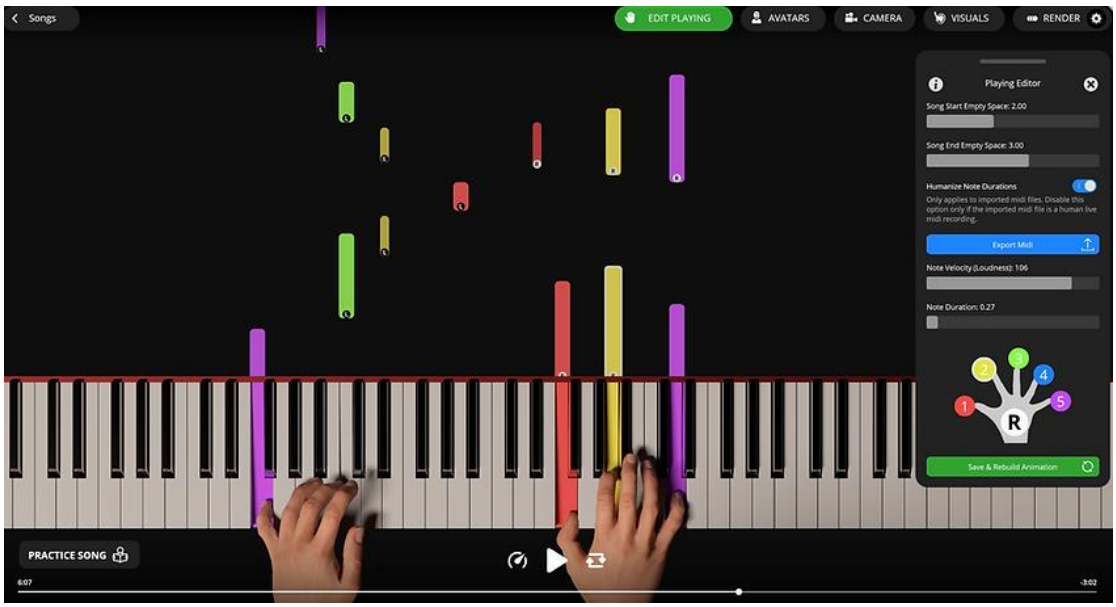

圖 壹-一 Concert Creator 指法標註介面

第二節、研究目的

本專題之目的為在有限的資料中透過機器學習得出能標記符合人體工學之鋼琴指法，並在讀取由製譜軟體 MuseScore 產生的 MIDI 鋼琴獨奏檔案後產生每一個音符對應的合理指法。

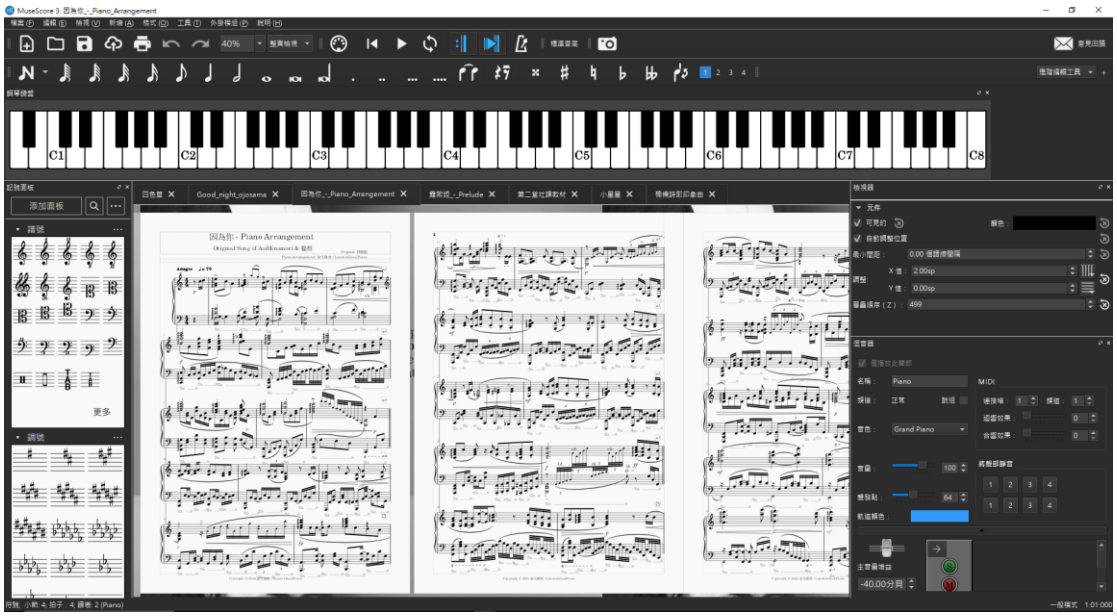

圖 壹-二 MuseScore 3 製譜軟體

第三節、研究之重要性

對於鋼琴演奏者而言，標記指法時往往需要透過自身經驗的累積，以及大量樂曲與演奏技巧的練習，才能在短時間內得出適合的指法。不同的指法能夠產生不同的音色、情感與音樂性，因此沒有絕對正確的指法，只有最適合作曲家原意，或是最能表現演奏者心境的指法，因此我們希望本專題只要能得到符合人體工學的指法即可。此外，正確標註的指法也有助於增加演奏者的肌肉記憶，並提供詮釋樂曲的思路，在後續遇到全新的樂曲時在短時間內找到合適的指法。

如下圖樂曲，李斯特的超技練習曲第四號「馬采巴」，在中間聲部的相同音型使用了相同的指法，但此指法較不符合人體工學，因為需要在極短的時間使用相同的指法演奏不同的連續的多個音型，反而若是使用其它手指就能演奏出連續的效果。這表示作曲家期望一個短促、唐突的不連續之音色，並且這便是沒有絕對正確，只有符合樂曲音樂性的合理指法之證明。

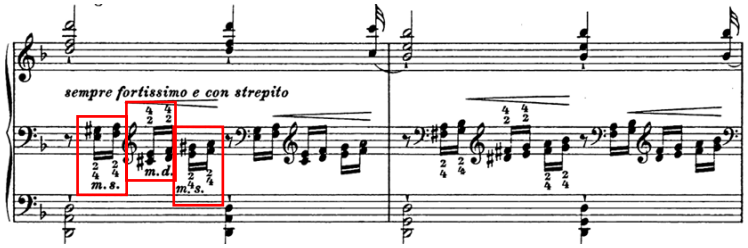

圖 壹-三 F.Liszt - Transcendental Etude S.139 No.4 in D minor, “Mazeppa”

第貳章、文獻探討

早在 1997 年，Parncutt 等人[1]提出條件約束型(constraint-based)模型使用預定義規則與動態規劃來估算指法，強調手指與音符之間的舒適互動距離，並且需要手動調整規則權重以達最佳效果。此後也有研究進行了改良[2-7]。

Yonebayashi 等人[8]首次提出統計型(statistical-based)模型，使用隱馬可夫模型(HMM)建立指法序列，特定指法的發生機率取決於前一時間的指法狀態與當前的音符。Nakamura 等人[9]提出了一種「合併 HMM」以分離不同聲部與標記雙手指法；李強等人[10]結合 constraint-based 與判斷函數改進 Viterbi 演算法中的最佳化規則，而 Nakamura 等人[11]提出了高階 HMM 模型、兩個深度神經網路模型(前饋網路與長短期記憶)，以及首次公開本專題採用的鋼琴指法資料集(PIG Dataset)。

Kim 等人[12]提到在音樂教育背景下對鋼琴演奏進行自動評估的方法(APPA)，而 Jeong 等人[13]以圖神經網路著重對付有音樂性的鋼琴演奏進行建模。

第參章、研究方法

第一節、研究對象

本專題採用 Nakamura 等人[11]發表的 PIG Dataset，並對琴鍵位置進行特徵提取，而我們將樂曲中每個音符對應的音高、力度、時間等特徵統稱為此音符對應的「音符事件」。PIG Dataset 中每個 txt 檔案都是樂曲的片段，一個音符有對應的起始指法與結束指法，而由於起始指法與結束指法不同，即「同音換指」的情況只佔 0.3%左右，因此我們僅針對起始指法進行研究；並且在研究過程中發現此資料集中存在大量不符合人體工學的指法序列，但又因為它是目前唯一可使用的鋼琴指法資料集，因此最終還是決定使用此資料集進行研究。實驗中我們將訓練集：驗證集：測試集之比例定為 8:1:1。

Table 1: The subsets of the fingering dataset.

Subset	Composers	Pieces (bars; notes)	Pieces (bars; notes) with different fingerings
Bach	1	10 (218; 3,657)	40 (872; 14,628)
Mozart	1	10 (185; 2,546)	60 (1,110; 15,276)
Chopin	1	10 (244; 4,022)	50 (1,220; 20,110)
Miscellaneous	24	120 (2,533; 38,501)	159 (3,355; 50,030)
All	24	150 (3,180; 48,726)	309 (6,557; 100,044)

圖 參-一 PIG Dataset 之組成

另外，我們使用 MuseScore 創作了三首有不同音符事件數量的完整樂曲，將樂曲以 MIDI 輸出，並按照 PIG Dataset 的格式進行特徵擷取與指法標註，以測試經由有缺陷的 PIG Dataset 得出的模型在面對完整樂曲的標註準確率。

表 參-一 自作三首樂曲之音符事件數量

	左手音符事件數量	右手音符事件數量	音符事件數量總計
樂曲 A	11	30	41
樂曲 B	94	42	136
樂曲 C	557	569	1126

第二節、資料處理與模型設計

PIG Dataset 中原本有的 7 個特徵：音符 ID、起始與結束時間（秒）、音高、起始與結束力度、MIDI 軌道，而我們首先對資料進行前處理。由於 MIDI 軌道在不同製作者的處理下有所差異，因此移除 MIDI 軌道此特徵；再根據 Nakamura 等人提出的網格法對鍵盤進行二維座標化，增加了琴鍵的 x 與 y 之座標兩個特徵。而指法的表示，PIG Dataset 將左手由拇指到小指依序標記為-1~-5，右手拇指到小指依序為+1~+5。由於左右手的人體結構具有對稱性，因此我們訓練了兩類模型：單手模型與雙手模型，而單手模型中又分為左手與右手。並且，因為手指對於單手只會有 5 類，雙手只有 10 類，因此我們對指法採取最簡單的 One-Hot Encoding。

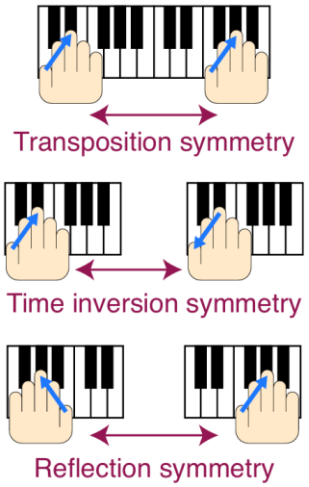

圖 參-二 三種不同的對稱性

受限於 PIG Dataset 提供的資料，我們最終使用傳統的 RNN 模型；單手與雙手結構皆採用 BiLSTM，並且每次輸入都設為 5 個時間步長組成，因此在每首樂曲的開頭會先插 4 筆空值。接著，在不破壞單一樂曲片段內的音符順序之前提下，將所有樂曲的順序隨機打亂，並在打亂後將所有樂曲結合成一個音符序列，拆分進行批次訓練。最後，繪製 Accuracy 與 Loss 的訓練曲線，並使用前述提到的測試集與自行編寫的三首完整樂曲進行測試，繪製測試的混淆矩陣。我們透過實驗對隱藏層的層數及位置改動，最終決定了模型的結構如下：

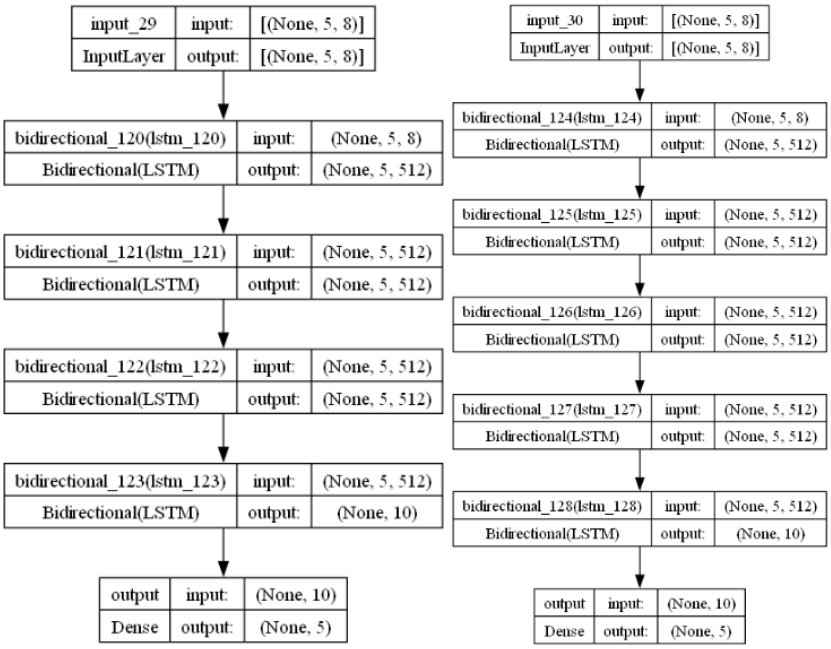

圖 參-三 左手與右手之單手模型架構

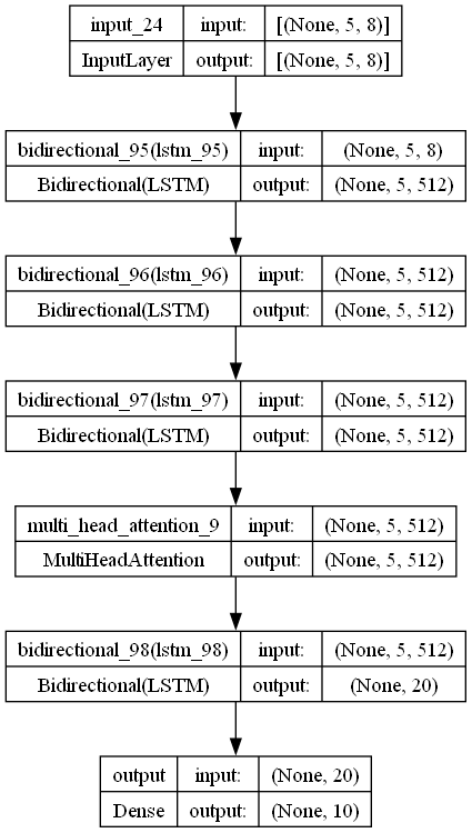

圖 參-四 雙手模型之架構

Activation function 與 Loss function 分別使用 Softmax 與 Categorical crossentropy，因為決定要使用哪隻手指本質上是分類任務；Optimizer 選擇 Adam，並且設定提前停止的 patience 為 3。本專題使用 Python 語言、Jupyter Notebook 結合 VScode 開發環境、Keras 神經網路函式庫，以及 cuDNN GPU 加速庫。

第肆章、研究結果與討論

我們使用 accuracy 與 loss 隨著 epochs 的變化繪製訓練曲線，並使用混淆矩陣呈現模型的預測結果。對於單手模型，可以發現左手模型的驗證集之 accuracy 比右手模型的高，而 loss 也比右手模型的低：

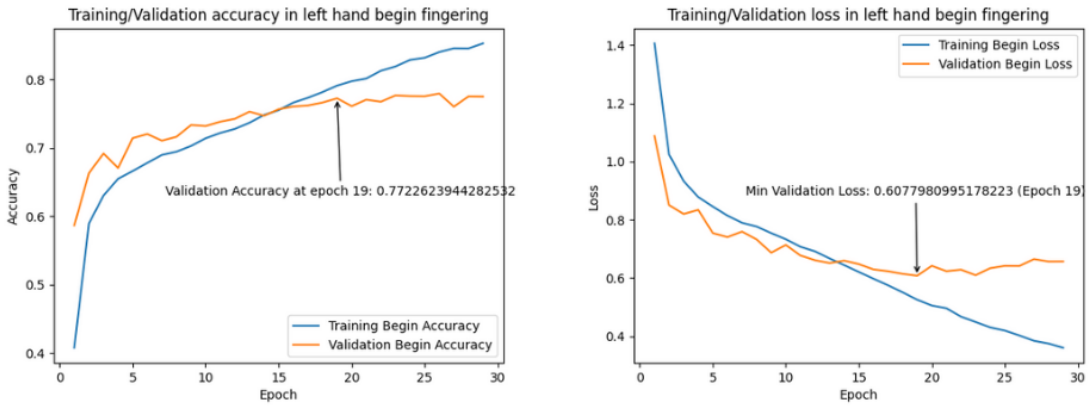

圖 肆-一 左手模型之 accuracy 與 loss 隨 epochs 變化曲線

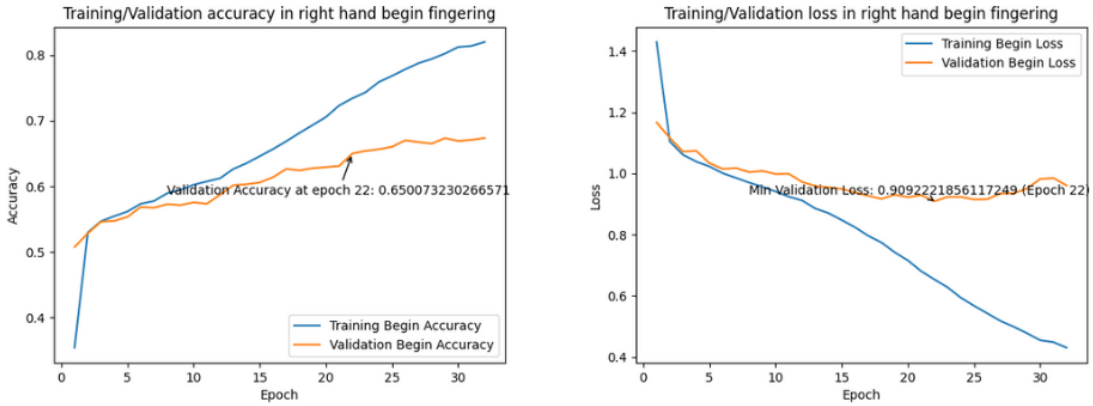

圖 肆-二 右手模型之 accuracy 與 loss 隨 epochs 變化曲線

對於雙手模型，可以發現它比右手模型的表現還要差。

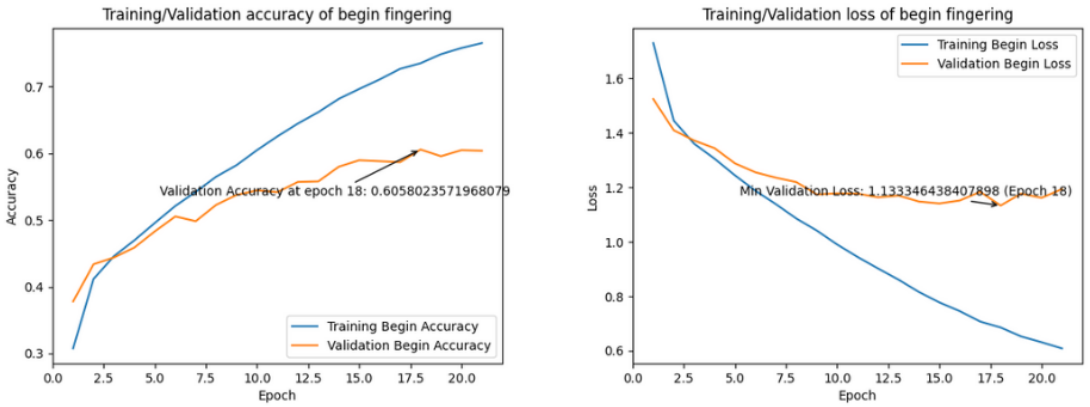

圖 肆-三 雙手模型之 accuracy 與 loss 隨 epochs 變化曲線

因為設定了提前停止，因此在 validation loss 連續三次都沒有變低後就會結束訓練。左手、右手及雙手分別經過了 19、22 及 18 個 epochs，並且 loss

都高於 0.5，初步推斷在 PIG Dataset 這個由樂曲片段構成且資料存在不合理分配，並且還存在多首相同樂曲不同指法的資料集上，模型只能大致學會小部分的合理指法序列。而使用訓練後的模型對測試集進行預測的結果如下：

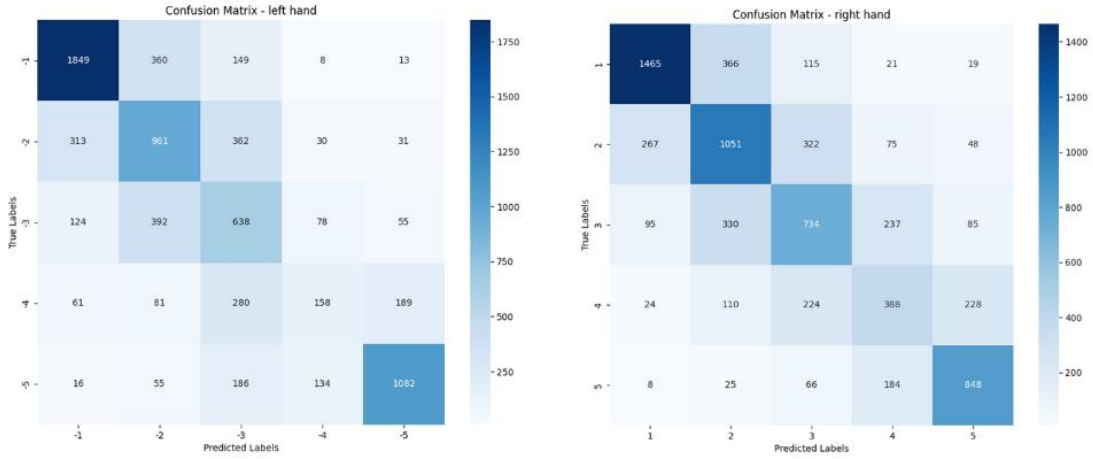

圖 肆-四 左右手模型的 PIG 測試集之混淆矩陣

觀察單手模型的指法分佈後可以發現，最常使用的是最外側的手指（拇指與小指），其次是靠近拇指的內側手指（二指與三指），最少使用的是人體相對較難控制的四指。而評估的指標我們選用 Accuracy, Loss, Precision score, Recall score 與 F1 score 五項：

表 肆-一 左右手模型的 PIG 測試集評估指標

	左手模型	右手模型
Accuracy	0.616	0.611
Loss	1.020	1.036
Precision score	0.619	0.612
Recall score	0.616	0.611
F1 score	0.613	0.611

雖然 accuracy 勉強高於六成，但 loss 依然皆高於 1，而相應的其它三項指標也都與 accuracy 相近。而雙手模型的預測表現如下：

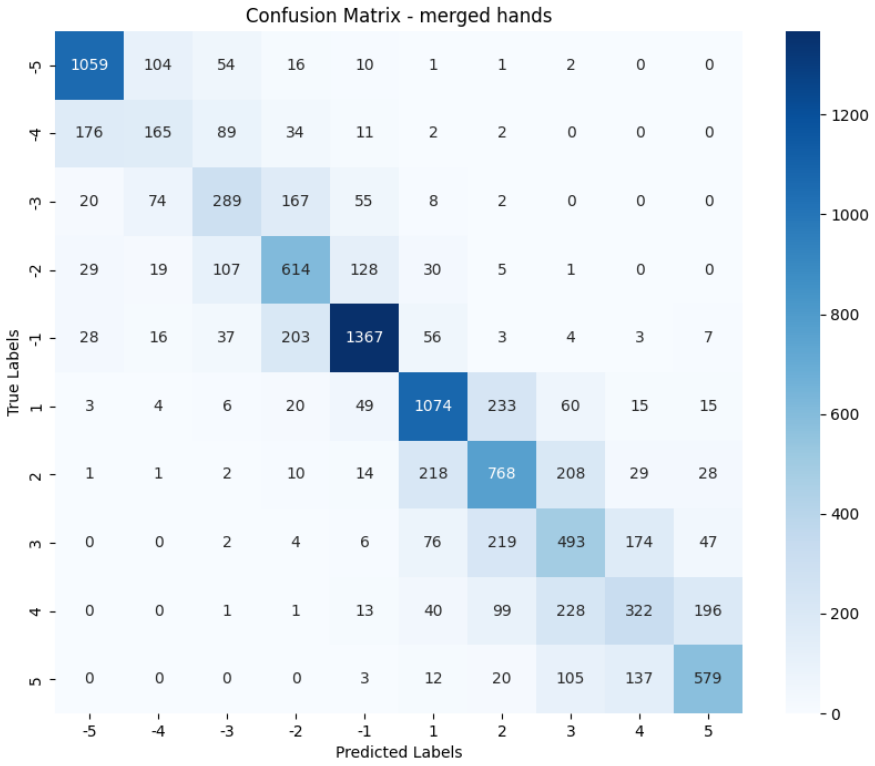

圖 肆-五 雙手模型的 PIG 測試集之混淆矩陣

雙手模型的混淆矩陣可以觀察到更多訊息。首先，它與單手模型「拇指與小指優先使用，二三指其次，最後為四指」這個按照手指的常用優先度由高至低選擇，可以看到合併兩手後的模型沒有遺失這一單手模型的特性；其次，對「應為某隻手的指法卻預測到另一隻手」的跨手錯誤預測情況，兩手不同手指的錯誤預測數量皆是由最內側的拇指到最外側的小指遞減。這表明模型在距離較近的跨手手指上預測錯誤的機率愈高，或是有不同符合人體的指法選擇愈多。例如左手拇指預測成右手拇指，因為這兩跨手手指在總共十指中距離最近，說明有些左手拇指的音符可能可以讓給右手拇指演奏。

表 肆-二 雙手模型的 PIG 測試集評估指標

	雙手模型
Accuracy	0.638
Loss	1.037
Precision score	0.635
Recall score	0.638
F1 score	0.635

而五個評估指標相較單手模型而言，除了 loss 以外都高了 0.02 左右，loss 則與右手模型相近。推測是因為雙手模型因為同時考慮兩手，而琴鍵位置這個特徵能使模型對不合理之指法相較於單手模型進行更多迴避。

而當我們使用完整的獨立樂曲依次讓模型進行預測時，發現在面對這些具有重複音型的樂曲，PIG Dataset 中篩選掉重複音型的樂曲片段使模型的準確率大幅下降，泛化能力極差。三首樂曲的評估指標選用 Accuracy, Loss 與 F1 score 三項。

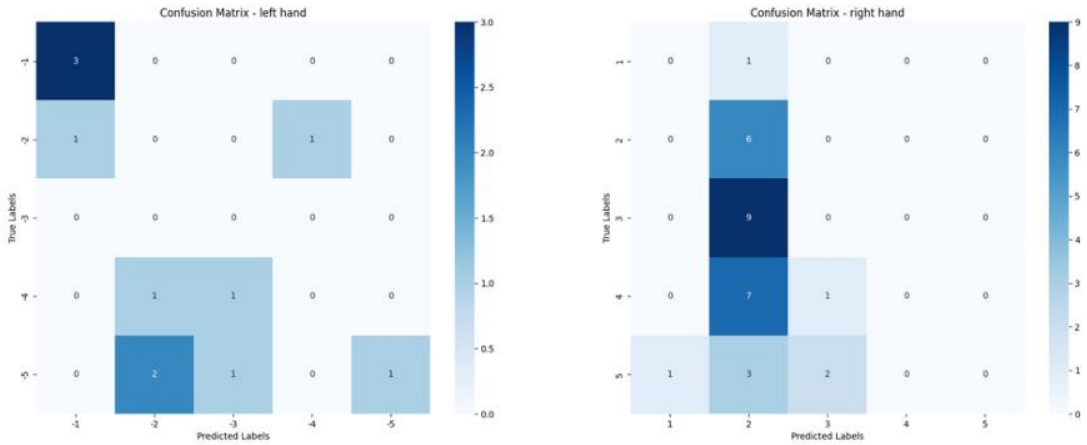

圖 肆-六 左右手模型預測樂曲 A 之混淆矩陣

表 肆-三 左右手模型預測樂曲 A 之評估指標

	樂曲 A 左手模型	樂曲 B 左手模型
Accuracy	0.364	0.2
Loss	2.069	3.513
F1 score	0.379	0.075

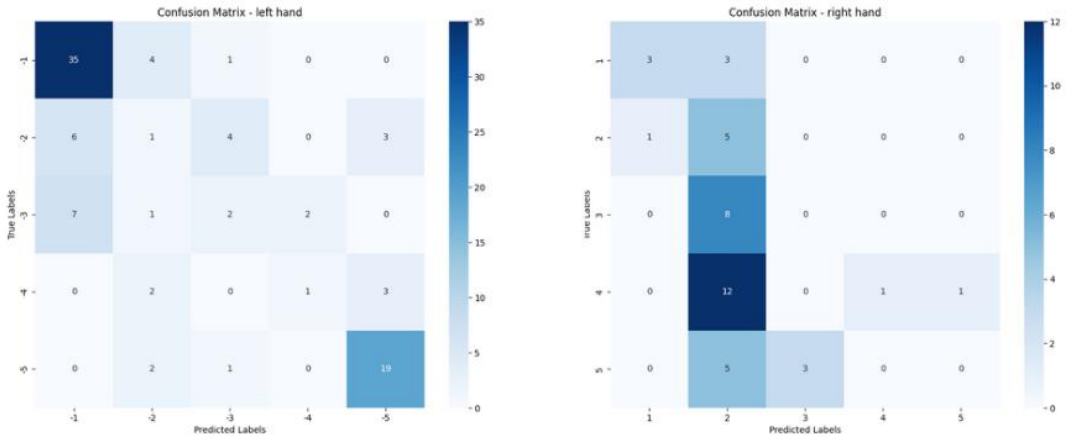

圖 肆-七 左右手模型預測樂曲 B 之混淆矩陣

表 肆-四 左右手模型預測樂曲 B 之評估指標

	樂曲 B 左手模型	樂曲 B 右手模型
Accuracy	0.617	0.214
Loss	1.069	3.680
F1 score	0.580	0.167

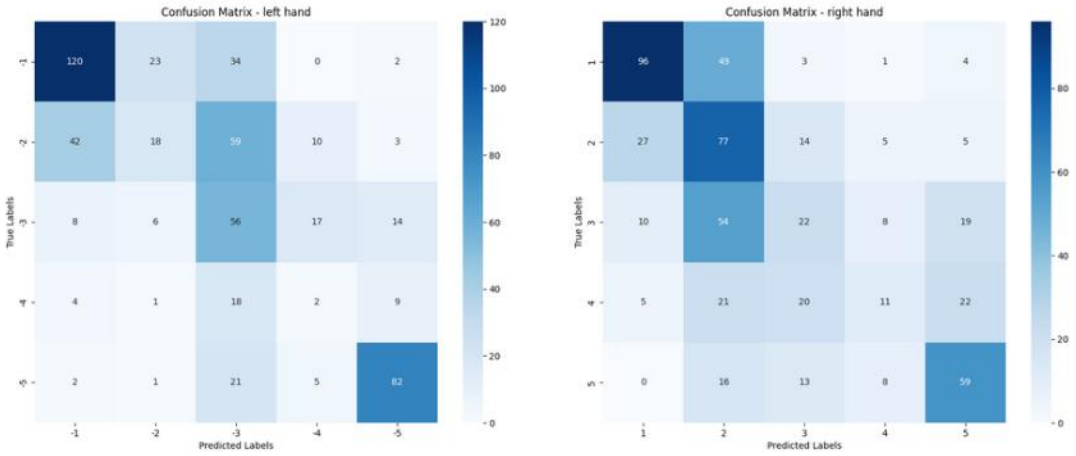

圖 肆-八 左右手模型預測樂曲 C 之混淆矩陣

表 肆-五 左右手模型預測樂曲 C 之評估指標

	樂曲 C 左手模型	樂曲 C 右手模型
Accuracy	0.499	0.466
Loss	1.225	1.526
F1 score	0.486	0.499

綜合單手模型的表現，可先推斷當音符事件數量與模型表現大致成正比，但還需考慮重複音型的存在。通常音符事件數量愈少，重複音型就愈少出現，旋律也愈簡單。其次，我們發現左手模型在錯誤預測時傾向預測成二指與三指，右手模型亦同且錯誤預測成二指的次數最多，而右手出現此現象可能與大部分樂曲都是右手演奏主旋律左手演奏伴奏，使右手二指較經常使用有關。接著，來看雙手模型之表現。

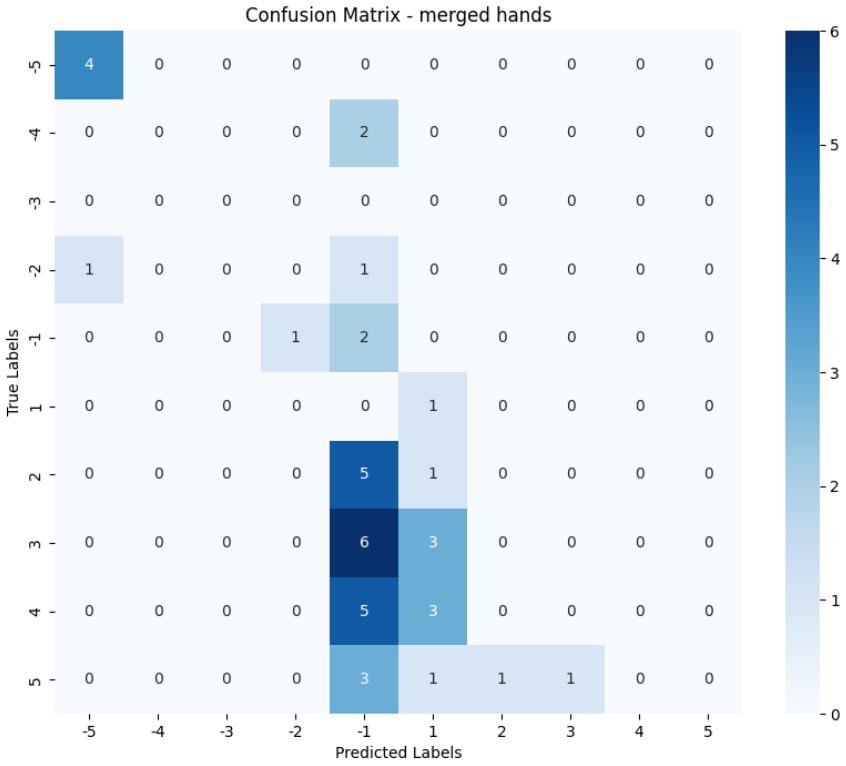

圖 肆-九 雙手模型預測樂曲 A 之混淆矩陣

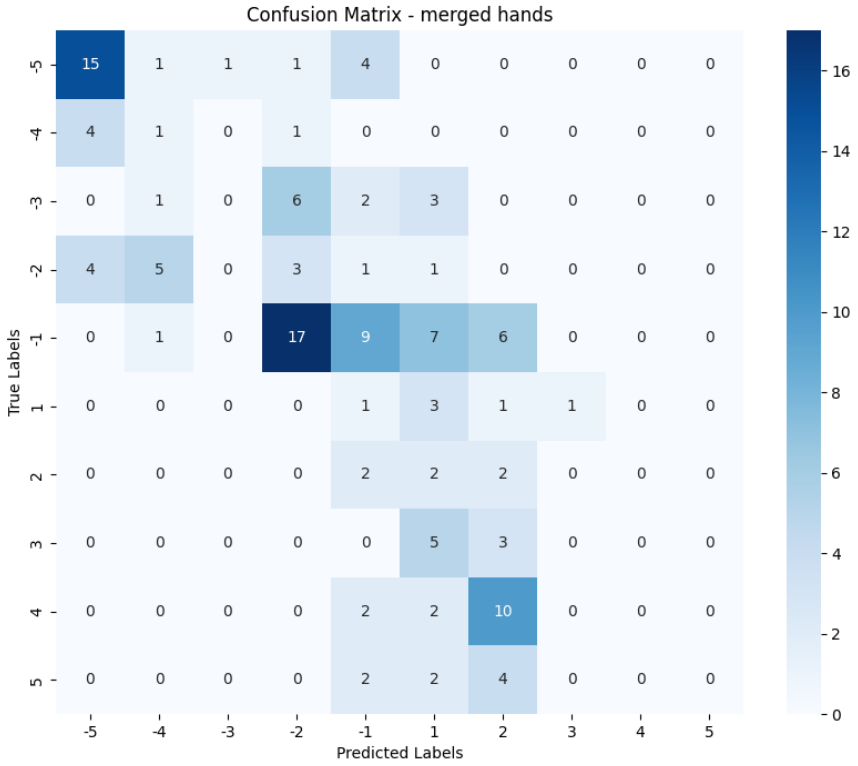

圖 肆-十 雙手模型預測樂曲 B 之混淆矩陣

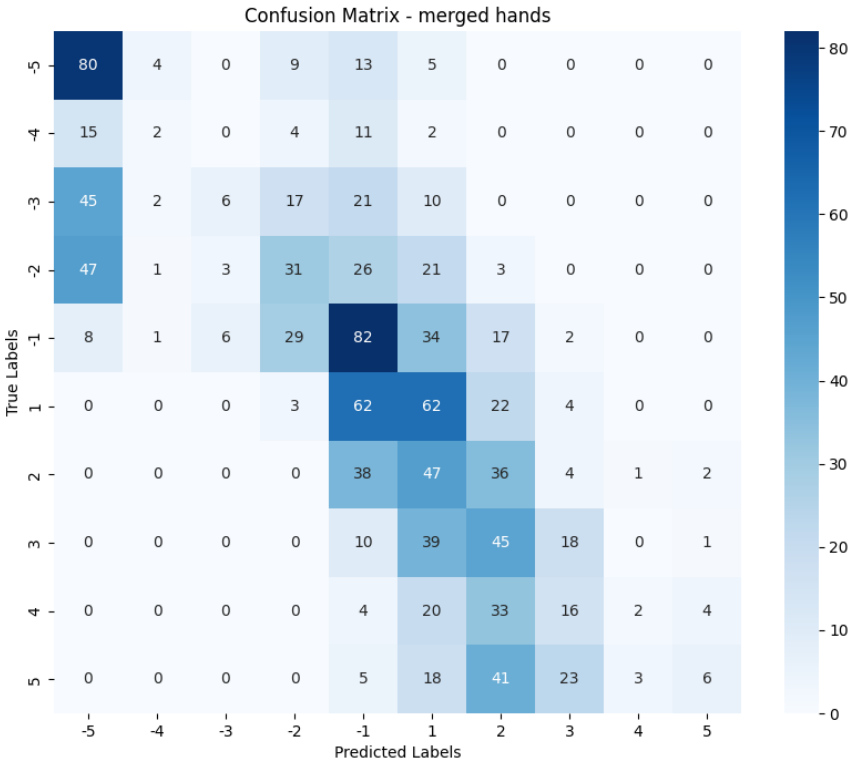

圖 肆-十一 雙手模型預測樂曲 C 之混淆矩陣

表 肆-六 雙手模型預測樂曲 ABC 之評估指標

	樂曲 A 雙手模型	樂曲 B 雙手模型	樂曲 C 雙手模型
Accuracy	0.171	0.243	0.289
Loss	5.190	2.860	2.782
F1 score	0.102	0.227	0.253

單看雙手模型，在樂曲 A 的表現最差，且 Loss 為所有實驗中最高，但在樂曲 B 與樂曲 C 卻又降低了將近一半，由此可見音符事件數量的影響。並且，Accuracy 與 F1 score 都低於三成，但觀察三個混淆矩陣的指法分佈，發現錯誤預測的手指都與真實手指的距離相近；並且觀察樂曲 C 更能發現拇指被錯誤預測的數量最多，並且錯誤預測數量有著隨手指向外側而遞減的趨勢。

綜合單手與雙手模型的表現，我們可以總結為幾點：

1. 所有模型正確預測合理指法的機率隨著音符事件的數量增加而提高。
2. 所有模型在錯誤預測時，對單一手指的錯誤預測數量隨著手指由內向外而遞減。
3. 錯誤預測時，與真實手指相近的其它手指被使用的機率最高。
4. 慣用且對人體最易於控制的手指，相較於最難控制的四指在錯誤預測時被使用的機率最高。
5. 縱使基於樂曲片段所訓練的模型在完整樂曲的表現不好，但還是能找到符合人體工學的部分規則。

由於 PIG Dataset 不僅只使用樂曲片段，並且還存在不合理的指法，當預測單一完整樂曲時準確率大幅下降；並且，單個樂曲片段存在至多 4 種不同合理指法的分佈也會使模型的準確率下降。若是再考慮同音換指問題，模型效能必然會再更進一步下跌。因此，我們可以採取幾項措施：

1. 建立自己的資料集，由完整且不重複的樂曲組成，並使用有公認指法的樂曲，例如哈農、徹爾尼等作曲家的作品。
2. 資料集中選用跨時代的作品，如巴洛克時期、後浪漫主義時期、近現代樂派、無調性音樂等等，使模型對不同樂派與風格的指法標註之泛化能力能夠在各方面都盡量涵蓋。
3. 由於樂曲的長度及音符事件數量必定不是固定的，因此若想將音符事件數量統一，必須使用填充等技巧，而填充的最大音符事件數量之選擇將會是關鍵，因為若資料集中單一樂曲的音符事件數量與欲預測的單一樂曲之音符事

件數量相差過大（如資料集中只有鋼琴前奏曲，但欲預測一整套鋼琴協奏曲），將會有更多影響指法的因素，例如重複音型、跨聲部琶音、同聲部換手等問題出現。

4. 在自行標註與產生資料集後，方可採用不同的資料前處理與模型架構，如將 MIDI 轉為訊號、影像或圖，模型採用大型語言模型常用的 Transformer、今年橫空出世的 KAN 等等。

第伍章、結論與未來研究

本專題基於 PIG Dataset 與簡易的 BiLSTM 架構建立了針對左手、右手與雙手三個鋼琴指法自動標註模型，並找出了樂曲與指法之間的關聯性。並且比對單手模型與雙手模型後，發現單獨處理左右手比起合併處理兩手的準確性較高；雖然受限於 PIG Dataset 的不準確性、不完善性，與這三個模型都遠遠不及實用，不過對未來如何準備資料、建立模型等等規劃提供了較為明確的思路及經驗。

鋼琴指法自動標註的技術必定沒有正確解答，但手指的合理運動規則是確實存在的。若能找出此規則，不僅對音樂教育、演奏有極大的幫助，在使用動態捕捉技術時亦能節省成本，還能幫助機械手臂參與高精度加工。只不過相對其它領域而言，此議題的相關研究只佔少數，仍需一段時間的發展，才可一窺手指運動的機制。

第陸章、參考文獻

[1] Richard Parncutt, John A Sloboda, Eric F Clarke, Matti Raekallio, and Peter Desain. 1997. An Ergonomie Model of Keyboard Fingering for Melodic Fragments Sibelius Academy of Music , Helsinski. Music perception: An Interdisciplinary Journal 14, 4 (1997), 341–382.

[2] Martin Gellrich and Richard Parncutt. 1998. Piano technique and fingering in the eighteenth and nineteenth centuries: Bringing a forgotten method back to life. British Journal of Music Education 15, 1 (1998), 5–23.

[3] J. Sloboda, E. Clarke, R. Parncutt, and M. Raekallio, “Determinants of finger

choice in piano sight-reading," Journal of Experimental Psychology: Human Perception and Performance, vol. 24, pp. 185–203, 02 1998

[4] M. Hart, R. Bosch, E. Tsai, Finding optimal piano fingerings. UMAP J. 21(2), 167–177 (2000)

[5] J. P. Jacobs, Refinements to the ergonomic model for keyboard fingering of Parncutt, Sloboda, Clarke, Raekallio, and Desain. Music. Percept. 18(4), 505–511 (2001)

[6] C. C. Lin, D.S.M. Liu, An intelligent virtual piano tutor, Proceedings of the 2006 ACM International Conference on virtual reality continuum and its applications. (ACM, Hong Kong, pp. 353–356

[7] A. Al Kasimi, E. Nichols, C. Raphael, in Paper presented at the 8th International Conference on Music Information Retrieval, ICMIR 2007. A simple algorithm for automatic generation of polyphonic piano fingerings, (Vienna, 2007), pp. 23–27

[8] Y. Yonebayashi, H. Kameoka, S. Sagayama, in Paper presented at the 20th International Joint Conference on Artificial Intelligence. Automatic decision of piano fingering based on hidden Markov models, (Hyderabad, 2007), pp. 6–12

[9] E. Nakamura, N. Ono, S. Sagayama, in Paper presented at the 15th International Society for Music Information Retrieval Conference, ISMIR 2014. Merged-output HMM for piano fingering of both hands, (Taipei, 2014), pp. 27–31

[10] L. Qiang, L. Chenxi, G. Xin, Automatic fingering annotation for piano score via judgement-HMM and improved viterbi. J Tianjin Univ. (Sci. Technol.) 53(08), 2020

[11] E. Nakamura, Y. Saito, K. Yoshii, Statistical learning and estimation of piano fingering. Inf. Sci. 517, 68–85 (2020)

[12] H. Kim, P. Ramoneda, M. Miron, and X. Serra. 2022. An Overview of

Automatic Piano Performance Assessment within the Music Education Context. In Proceedings of the 14th International Conference on Computer Supported Education.
[13] Dasaem Jeong, Taegyun Kwon, Yoojin Kim, and Juhan Nam. 2019. Graph neural network for music score data and modeling expressive piano performance. In International Conference on Machine Learning. PMLR, 3060–3070.