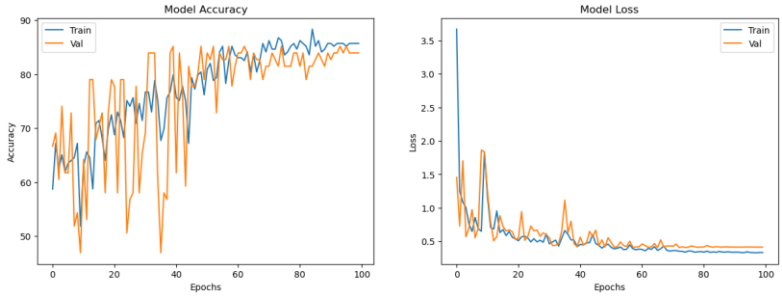
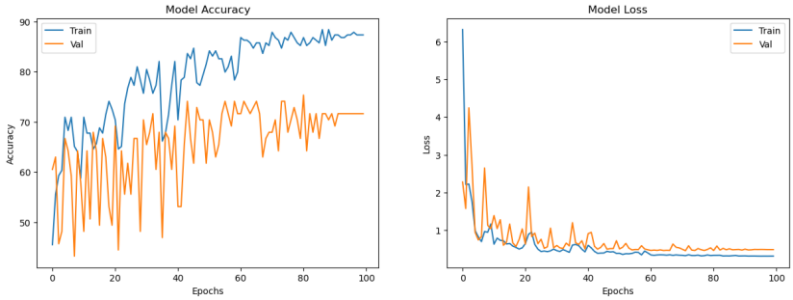
國立陽明交通大學 312704004 郭盈均

1. Select 2 hyper-parameters of the artificial neural network used in Lab 2, and set 3 different values for each. Perform experiments to compare the effects of varying these hyper-parameters on the loss and accuracy metrics across the training, validation, and test datasets. Present your findings with appropriate tables.

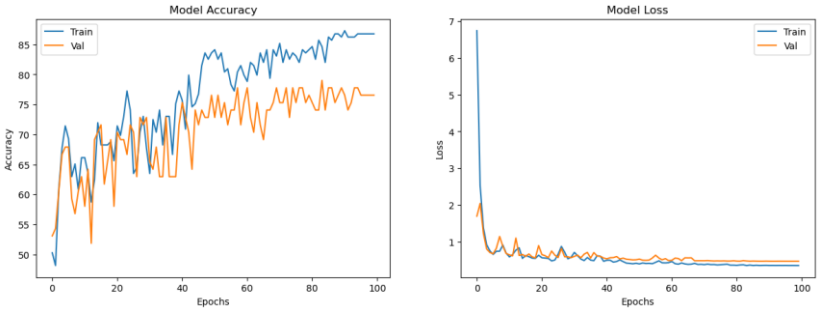
   i. Batch size = 8, 16, 32




Batch size = 8



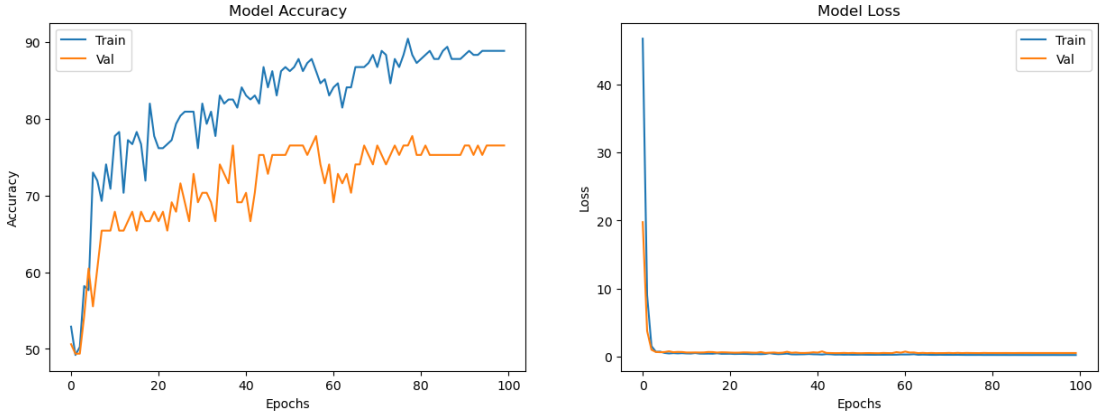

Batch size = 16



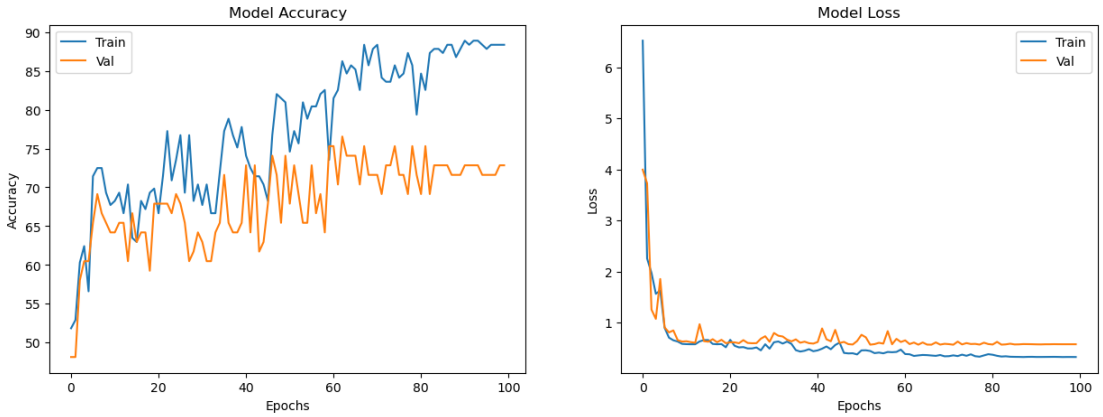

Batch size = 32

| Batch size | 8 | 16 | 32 |
|---|---|---|---|
| Accuracy | 67.74% | 80.65% | 77.51% |

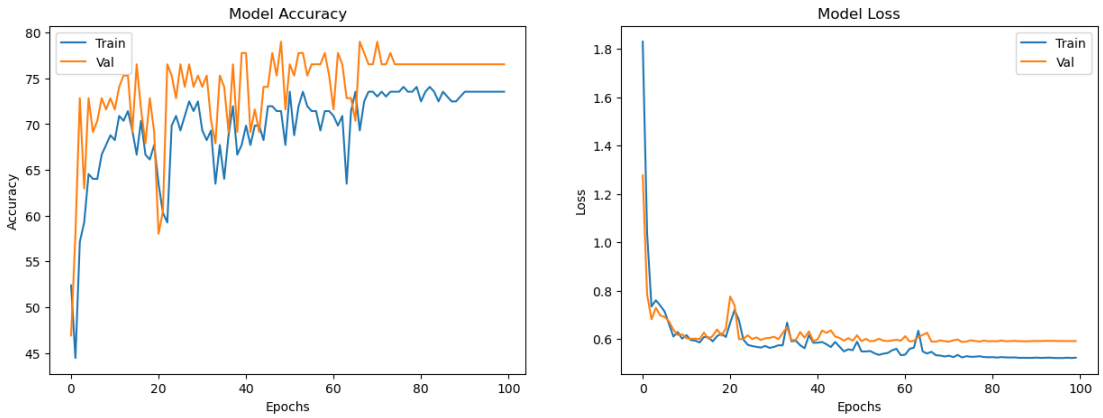
ii. Learning Rate = 0.01, 0.001, 0.0001




Learning Rate = 0.01




Learning Rate = 0.001



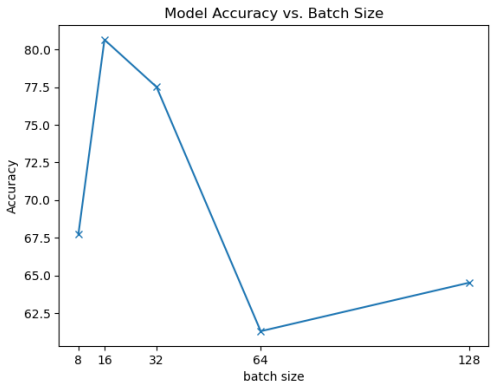

Learning Rate = 0.0001

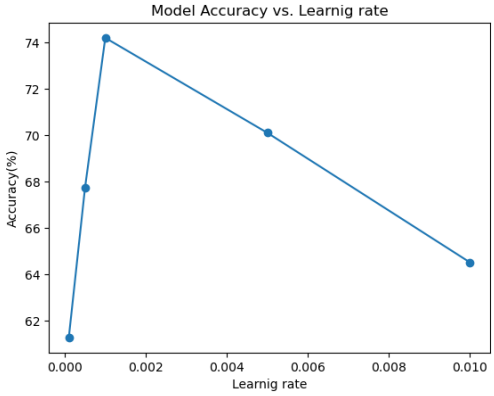
| Learning rate | 0.01 | 0.001 | 0.0001 |
|---|---|---|---|
| Accuracy | 64.52% | 74.19% | 61.29% |

2. Based on your experiments in Question 1, analyze the outcomes. What differences do you observe with the changes in hyper-parameters? Discuss whether these adjustments contributed to improvements in model performance, you can use plots to support your points.

我嘗試改變 Batch size 及 Leaning rate 這項超參數，超參數的改變確實會影響性能。當 Batch size =16 時性能最好，不論是調大或調小，性能都會降低。另外，當超參數 Learning rate =0.001 時，性能優於其他比較值。

Batch size 較大者可以使梯度計算時較為穩定，而小批量訓練過程更新較為頻繁，相對有機會更快收斂。Learning Rate 指得是學習率，主要是決定每次更新參數時的步長大小，當步長太大，容易無法收斂，步長太小則更新太慢，花上更多計算時間。




3. In Lab 2, you may have noticed a discrepancy in accuracy between the training and test datasets. What do you think causes this occurrence? Discuss potential reasons for the gap in accuracy

i. 我認為例題中的訓練資料不夠多，可能使得在訓練的過程學習效果有限

ii. 超參數的設置會影響模型訓練的過程，不適當的超參數設置可能導致過度擬合或欠擬合，進而影響模型在測試數據上的準確率。

iii. 由於在資料前處理過程，有將非數值型資料轉換成數值型，若是此種非數值型特徵沒有被好好捕捉，意即訓練模型沒辦法好好學習該特徵帶來的訊息，可能也會造成測試數據時的性能表現不佳。

4. Discuss methodologies for selecting relevant features in a tabular dataset for machine learning models. Highlight the importance of feature selection and how it can impact model performance. You are encouraged to consult external resources to support your arguments. Please cite any sources you refer to.

特徵提取不僅可以改善預測準確度，還可以提高模型的解釋性和執行效率。常見三種特徵提取方法分別為過濾法、包裝法及嵌入法。過濾法中常見為卡方過濾法，計算每個特徵與目標變量之間的卡方統計量，然後根據這些統計量來衡量特徵與目標之間的相關性。通常卡方統計量越高，特徵與目標之間的相關性就越大。排名較高的特徵適合作為特徵子集。

包裝法中常見的為遞迴特徵消除法(RFE)，透過反覆訓練模型並消除相對不重要的特徵，從而找到最佳的特徵子集。RFE 通常與某些監督學習模型結合使用。

Ref/ Medium-資料前處理-特徵工程、Medium-特徵篩選 — 基本介紹

5. While artificial neural networks (ANNs) are versatile, they may not always be the most efficient choice for handling tabular data. Identify and describe an alternative deep learning model that is better suited for tabular datasets. Explain the rationale behind its design specifically for tabular data, including its key features and advantages. Ensure to reference any external sources you consult.

TabNet 是 Google Cloud 在 2019 年提出的專門針對表格型資料設計的神經網路結構，結合了樹模型和 DNN 的優勢。TabNet 採用順序多步架構，每一步驟會根據前一步驟的訊息來決定要使用那些特徵，自動選取重要特徵後，將處理完後的特徵聚合到總體的決策中，此外導入了 Encoder-Decoder 框架來實現自監督學習，提高了模型可解釋性。

Ref/人工智能选股--最适合股票表格数据的神经网络模型