2022全國智慧製造大數據分析競賽決賽
團隊測驗報告

報名序號:111011 (格式:111XXX)
團隊名稱:Urban2.0

註1:請用本PowerPoint 文件撰寫團隊測驗報告, 請轉成PDF檔案繳交。
註2:依據競賽須知第八條, 第5項規定:
決賽簡報之書面及口頭報告、服裝, 均不得使用學校系所標誌、提及學校系所、教授姓名及任何可供辨識參賽者身分的資料, 違者取消參賽資格, 或由主辦單位及評審會議決定處理方式
註3:請於11/19(六) 12:41前繳交團隊測驗報告及測驗結果, 至主辦單位指定網站。

【提醒】

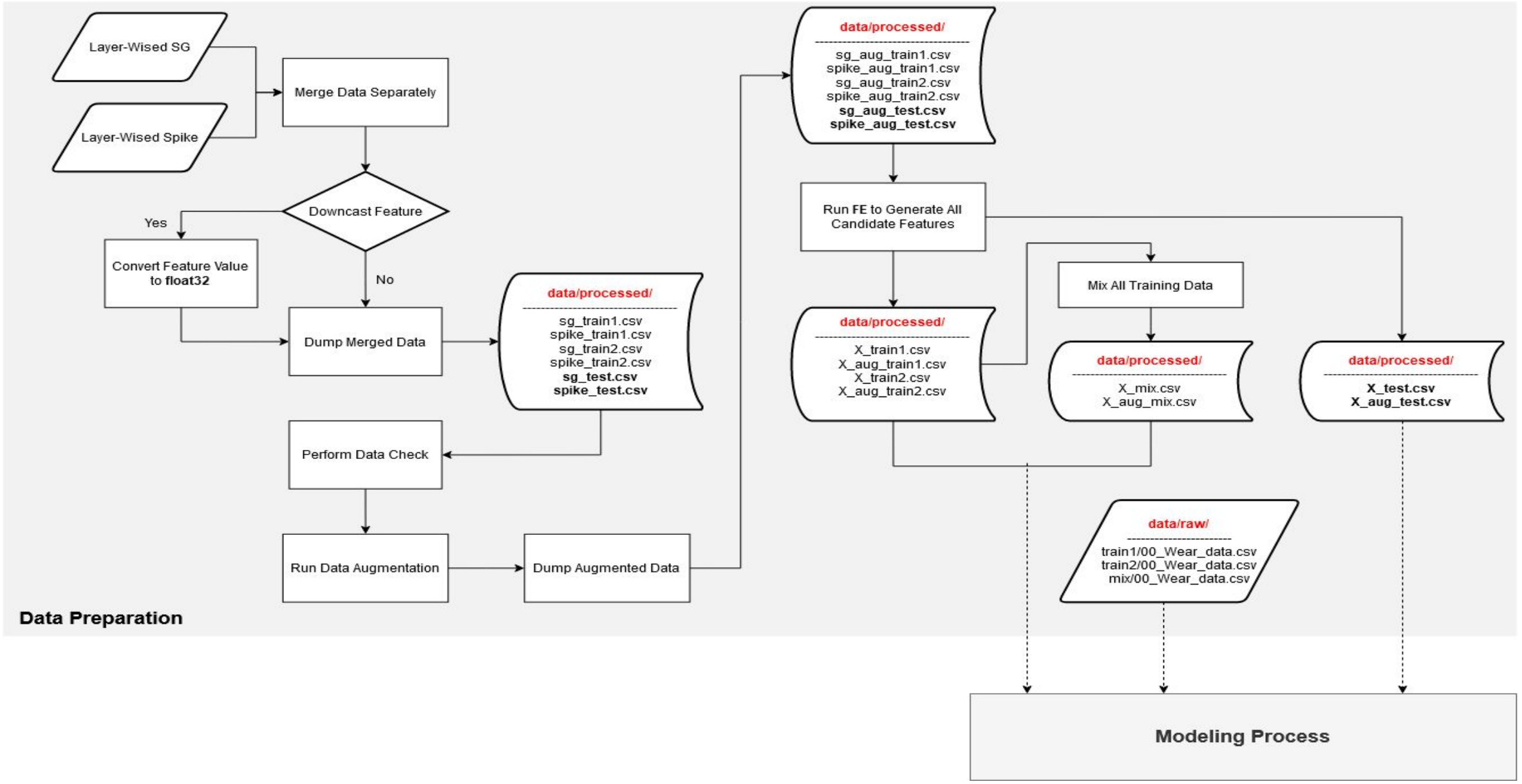
11/19(六)請繳交兩種檔案:

1. 簡報檔, 檔名命名規則如下, 使用英文命名:
- **ProjectA:報名序號_ projectA _report.pdf, 例如:111999_projectA_report.pdf**
- **ProjectB:報名序號_ projectB _report.pdf, 例如:111999_projectB_report.pdf**

2. 決賽測驗結果檔, 檔名命名規則如下, 使用英文命名:
- **ProjectA:報名序號_projectA_ans.csv, 例如:111999_projectA_ans.csv**
- **ProjectB:報名序號_projectB_ans.csv, 例如:111999_projectB_ans.csv**

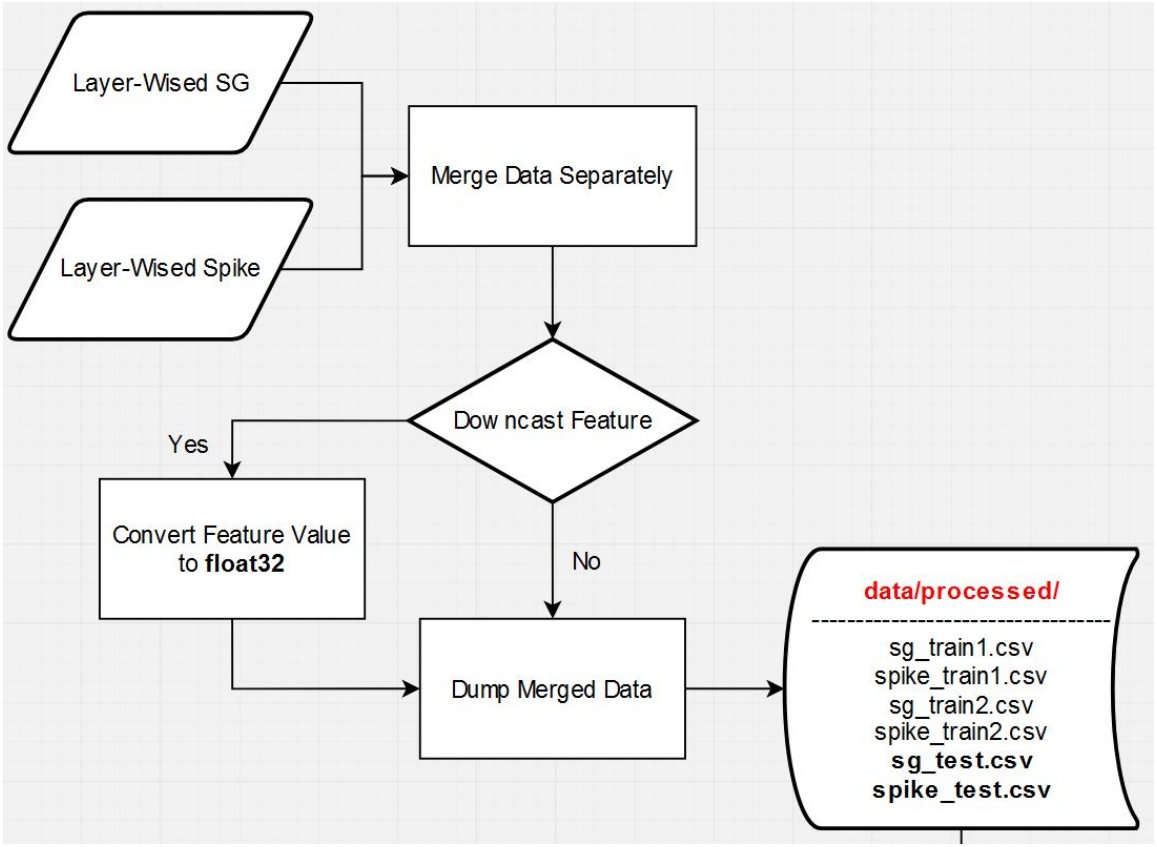
一、資料前處理(說明資料前處理過程)

Overview


Layer-Wised SG
Layer-Wised Spike
Merge Data Separately
Downcast Feature
Yes
Convert Feature Value to float32
No
Dump Merged Data
data/processed/
sg_train1.csv
spike_train1.csv
sg_train2.csv
spike_train2.csv
sg_test.csv
spike_test.csv
Perform Data Check
Run Data Augmentation
Dump Augmented Data
data/processed/
sg_aug_train1.csv
spike_aug_train1.csv
sg_aug_train2.csv
spike_aug_train2.csv
sg_aug_test.csv
spike_aug_test.csv
Run FE to Generate All Candidate Features
data/processed/
X_train1.csv
X_aug_train1.csv
X_train2.csv
X_aug_train2.csv
Mix All Training Data
data/processed/
X_mix.csv
X_aug_mix.csv
data/processed/
X_test.csv
X_aug_test.csv
data/raw/
train1/00_Wear_data.csv
train2/00_Wear_data.csv
mix/00_Wear_data.csv
Modeling Process
Data Preparation

資料清洗與合併

- 合併layer-wise sg以及spike資訊
- 降低feature精度
  - Disk-efficient
  - RAM-efficient
- 修正時間資訊錯誤標記
  - 檢查時間是否為單調遞增
  - 修正train1 spike layer11的時間資訊
- 刪除train2 spike layer17多餘的資訊



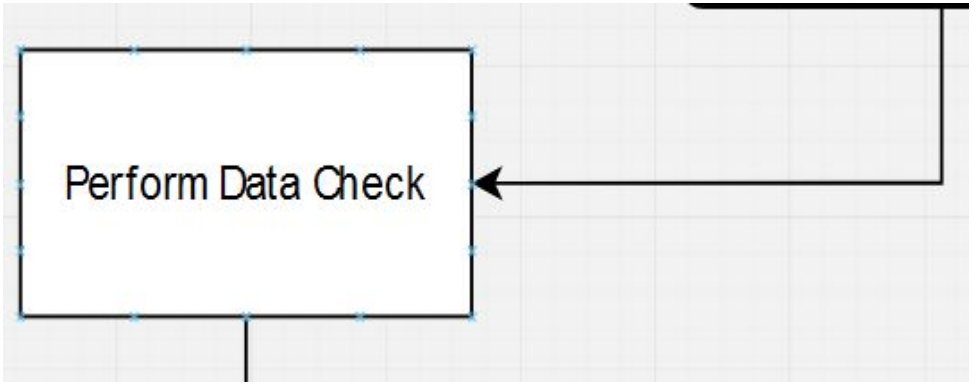
資料檢查

- 原始資料無資料缺失問題 (i.e., missing values)
- 同把智慧刀把的切削行為可被大略分為兩組 (train2僅一組)

```
Check if sg and spike information are sliced into two groups (like train1 26 / 20)...
(Directly check groups' boundary (min/max) and #layers per group (count).)
=====sg=====
layer          1       2       3       4   ...      43      44      45      46
n_samples  416001  416001  416001  416001   ...  284001  284001  284001  284001

[1 rows x 46 columns]
```

n_samples	min	max	count
284001	27	46	20
416001	1	26	26

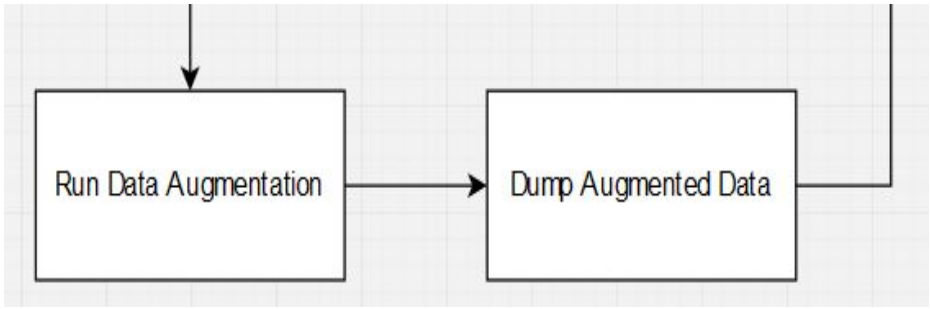


- 同一層資訊可以再切分為更細的行為單元

```
Check sg["e"].diff().sort_values() to help DA slicing...
Layer  1: -12.33 -12.29 -12.25 -12.21 -12.18 -11.66 -11.63 -11.56 -11.51 -11.37
           -7.68 -7.48 -7.28 -7.1 -7.07 -7.02 -6.99 -6.96 -6.95 -6.93
           -4.92 -4.81 -4.63 -4.61 -4.57 -4.56 -4.53 -4.52 -4.5 -4.49
```

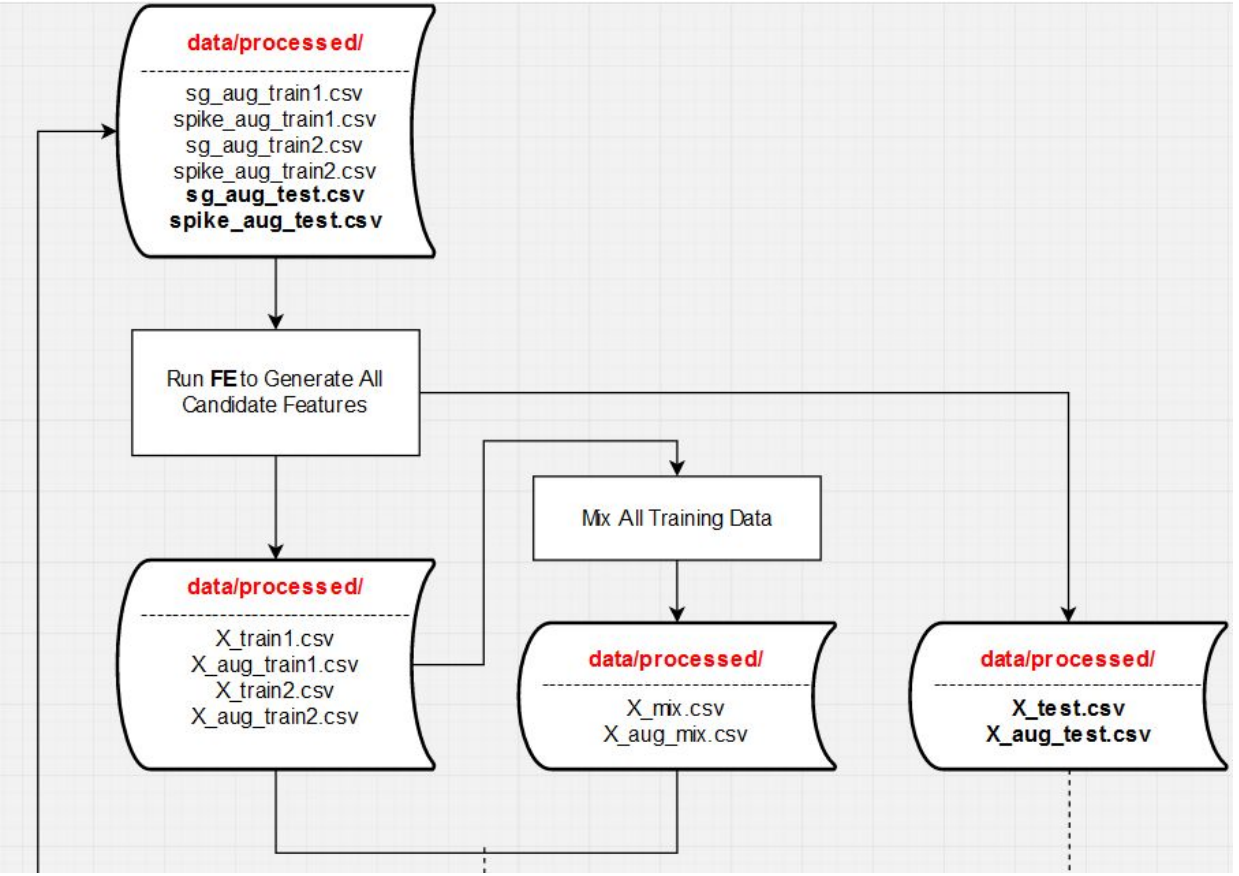
Data Augmentation - Fine-Grained Slicing

- 將更層資訊切分出更細緻的切削行為單元
  - 手動記錄sg中特徵E差值的負數極值時間點位並在本機驗證切分結果



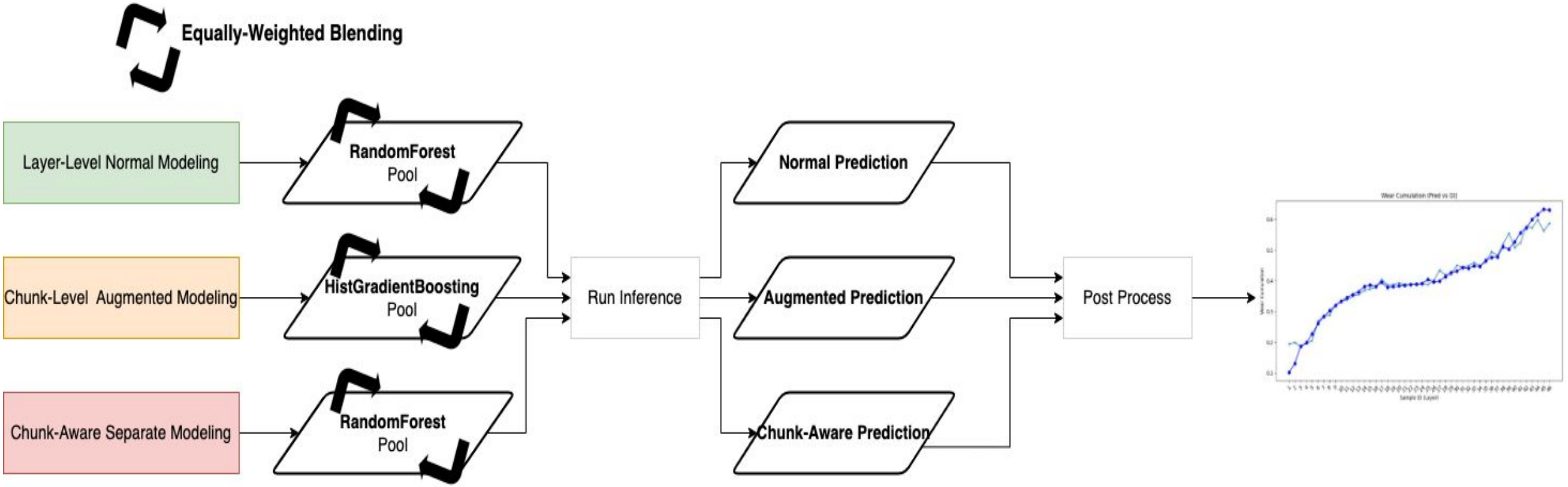
特徵工程

- 分兩種層次建立特徵
  - 以Layer為單位 - Coarse-grained
  - 以Chunk為單位 - Fine-grained
- 特徵分類 (共1584項特徵)
  - 簡單統計量
    - 平均值、標準差、分位數等
  - 數值趨勢
    - 建立簡易線性回歸模型
  - 數值變化量
    - 變化率、差值等



二、演算法和模型介紹(介紹方法細節)

Overview


Equally-Weighted Blending
Layer-Level Normal Modeling
RandomForest Pool
Chunk-Level Augmented Modeling
HistGradientBoosting Pool
Chunk-Aware Separate Modeling
RandomForest Pool
Run Inference
Normal Prediction
Augmented Prediction
Chunk-Aware Prediction
Post Process

Layer-Level Normal Modeling

- 特徵轉換與特徵選取
  - QuantileTransformer
  - VarianceThreshold(threhold=0.09)
  - SelectKBest(f_regression, k=100)
- 模型選擇
  - 輕量級RandomForestRegressor(n_estimators=50)
- Cross-Validation
  - KFold(n_splits=10) with 20 random seeds

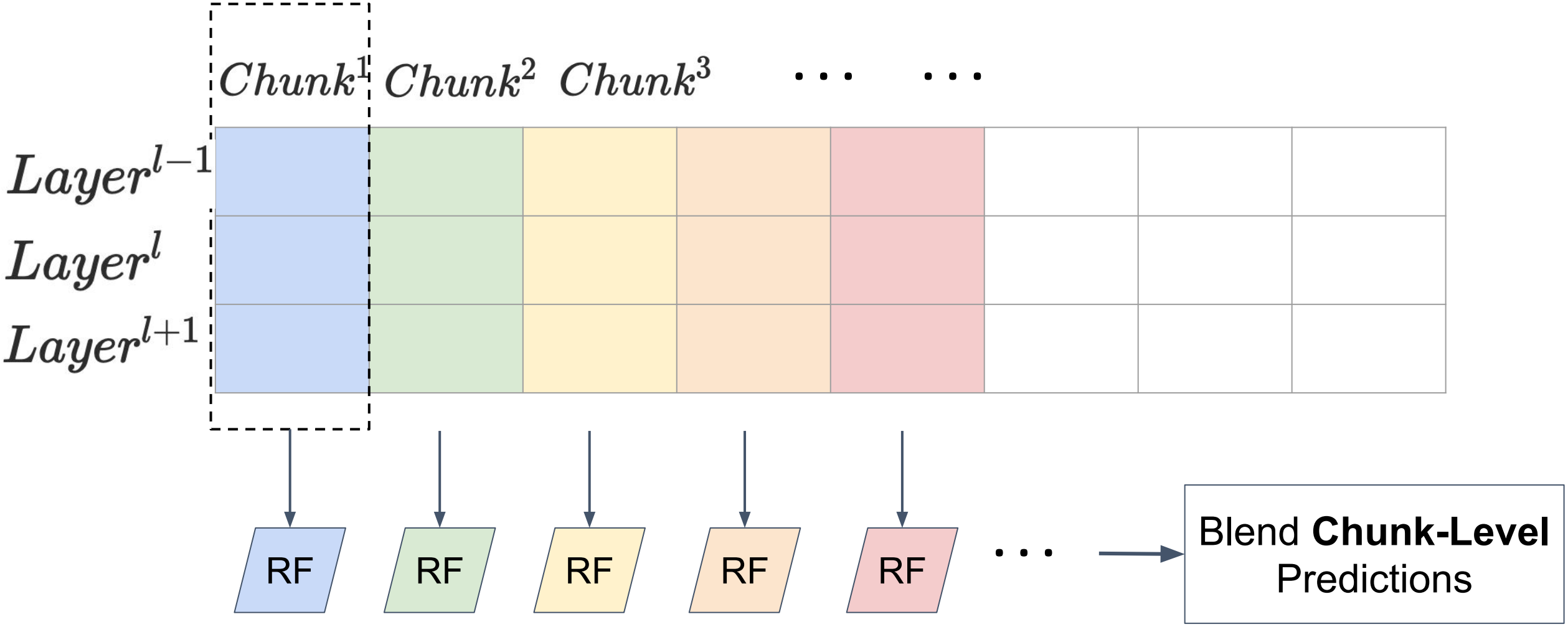
Chunk-Level Augmented Modeling

- 特徵轉換與特徵選取
  - QuantileTransformer
  - VarianceThreshold(threhold=0.085)
  - SelectKBest(f_regression, k=25)
- 模型選擇
  - HistGradientBoostingRegressor(max_leaf_nodes=8, min_samples_leaf=15)
- Cross-Validation
  - ShuffleGroupKFold(n_splits=10) with 20 random seeds

Chunk-Aware Separate Modeling

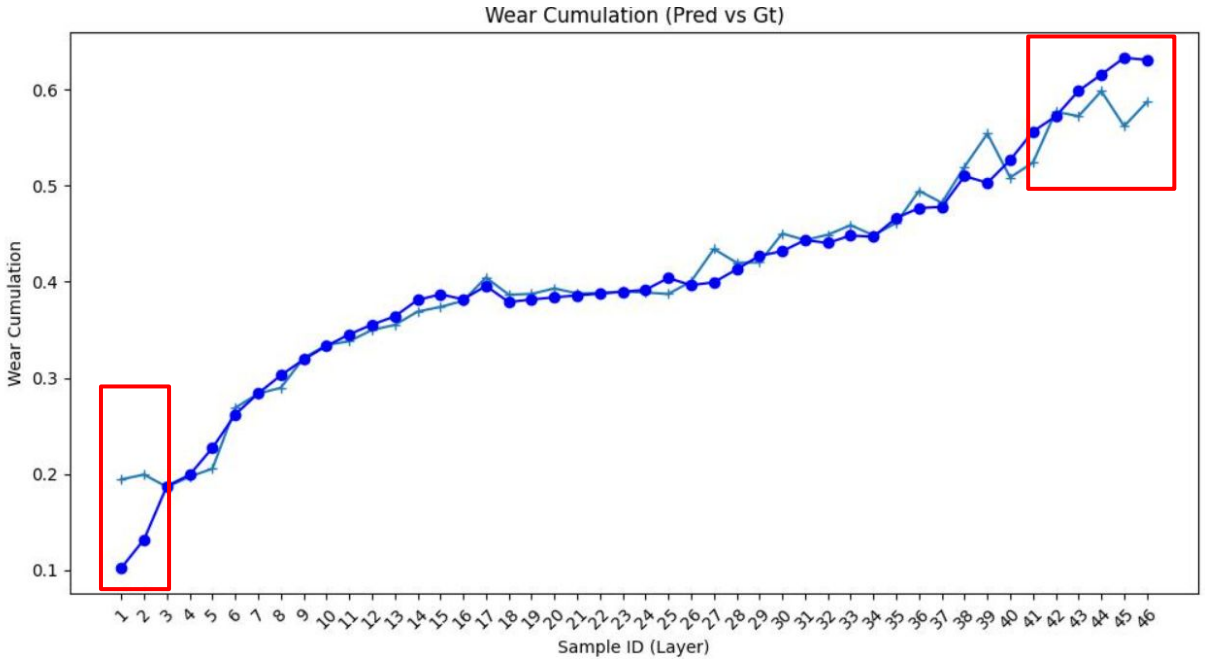
- 特徵轉換與特徵選取
  - QuantileTransformer
  - VarianceThreshold(threhold=0.085)
  - SelectKBest(f_regression, k=100)
- 模型選擇
  - 輕量級RandomForestRegressor(n_estimators=50)
- Cross-Validation
  - KFold(n_splits=10) with 20 random seeds
    - 每個chunk需分開建模

Chunk-Aware Separate Modeling (cont.)



Ensemble and Post-Processing

- **Ensemble – Inter-pool equally-weighted blending**
  - 三個model pool各自將預測值取平均
- **Post-Processing**
  - 修正頭端及尾端急遽磨耗的預測值



三、執行環境/套裝選擇/執行方式

執行環境與套件選擇

- 使用tf_keras環境
  - 實驗設置均用 *.yaml* 控制與調整
- 套件選擇
  - 主要使用pandas、numpy、scikit-learn、PyYAML、joblib等

執行方式

- **Data Preparation**
  - python -m data_preparation.clean_and_merge --dataset **<dataset>**
  - python -m data_preparation.check_data --dataset **<dataset>**
  - python -m data_preparation.run_da_slicing --dataset **<dataset>** --neg-peak-thres **<neg int>**
  - python -m data_preparation.run_fe --dataset **<dataset>** --data-type normal
  - python -m data_preparation.run_fe --dataset **<dataset>** --data-type aug
  - python -m data_preparation.mix_train

執行方式 (cont.)

- **Model Training**
  - Layer-Level Normal Modeling
    - python -m tools.train_eval --dataset **<dataset>** --data-type normal --mix-aug False --model-name rf --exp-id <n>
  - Chunk-Level Augmented Modeling
    - python -m tools.train_eval --dataset **<dataset>** --data-type aug --mix-aug True --model-name hgb --exp-id <n>
  - Chunk-Aware Separate Modeling
    - python -m tools.train_eval_chunk --dataset **<dataset>** --data-type aug --mix-aug False --model-name rf --exp-id <n>

執行方式 (cont.)

- **Inference**
  - python -m tools.infer --dataset **test** --exp-id **<要使用的model之exp_id>**

四、補充說明(或自行定義項目)

CV Reliability Study

- 頭段及尾段急遽磨耗為最具挑戰性的加工層