VSB Power Line Fault Detection

110034023 翼瑄卉　110034103 陳詩凱

Content

Introduction

01 Introduction

Purpose

To detect partial discharge patterns in signals acquired from these power lines with a new meter designed at the ENET Centre at VSB. Effective classifiers using this data will make it possible to continuously monitor power lines for faults.

Evaluation

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

Where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

01 Introduction

Submission File

For each signal in the test set, you must predict a binary prediction for the target variable. The file should contain a header and have the following format:

```
signal_id,target
0,0
1,1
2,0
etc.
```

Prize

1st Place: $ 12,000
2nd Place: $ 8,000
3rd Place: $ 5,000

Dataset

02 Dataset

Data description

Each signal contains 800,000 measurements of a power line's voltage, taken over 20 milliseconds. As the underlying electric grid operates at 50 Hz, this means each signal covers a single complete grid cycle. The grid itself operates on a 3-phase power scheme, and all three phases are measured simultaneously.

02 Dataset

File description

- **id_measurement：**
  the ID code for a trio of signals recorded at the same time.
- **signal_id：**
  the foreign key for the signal data. Each signal ID is unique across both train and test, so the first ID in train is '0' but the first ID in test is '8712'.
- **phase：** the phase ID code within the signal trio. The phases may or may not all be impacted by a fault on the line.
- **target：** 0 if the power line is undamaged, 1 if there is a fault.
- **[train/test].parquet ：** The signal data.

signal_id	id_measurement	phase	target
0 – 8711	0 – 2903	0 – 2	0 – 1
0	0	0	0
1	0	1	0
2	0	2	0
3	1	0	1
4	1	1	1
5	1	2	1
6	2	0	0
7	2	1	0

Best Solution Overview

Reference Team: mark4h

03-1 Best solution overview

1. Pre-processing

Peak separation, denoise,
find the noise floor.

2. Feature modeling

Previous and next
peak comparison.

3. Model

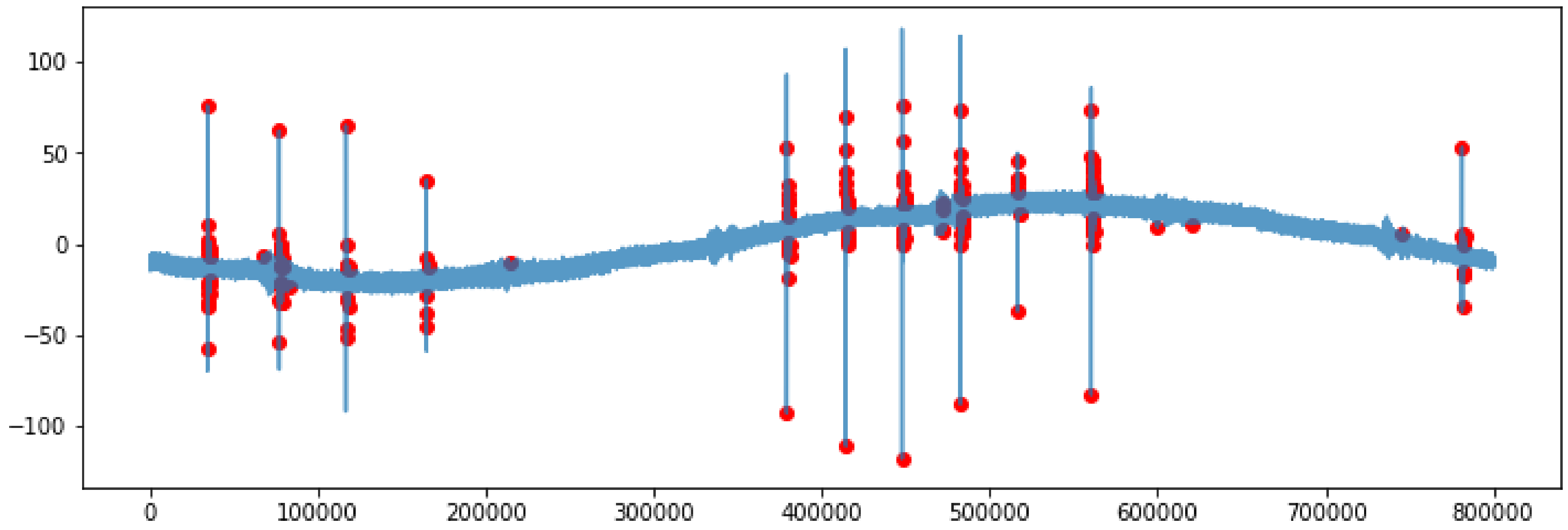
LightGBM

4. Analysis

Prediction errors,
feature importance

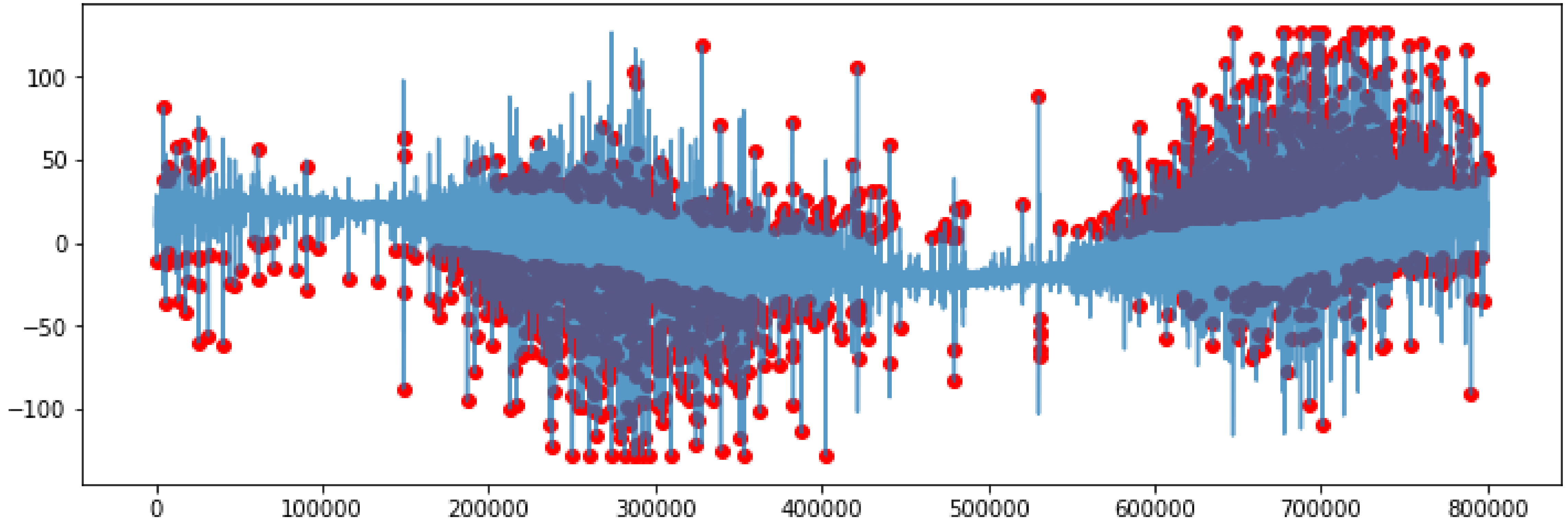
03-2 Preprocessing

Finding the noise of peak



Sigid = 10

Mark the noise of the peak produced by PD

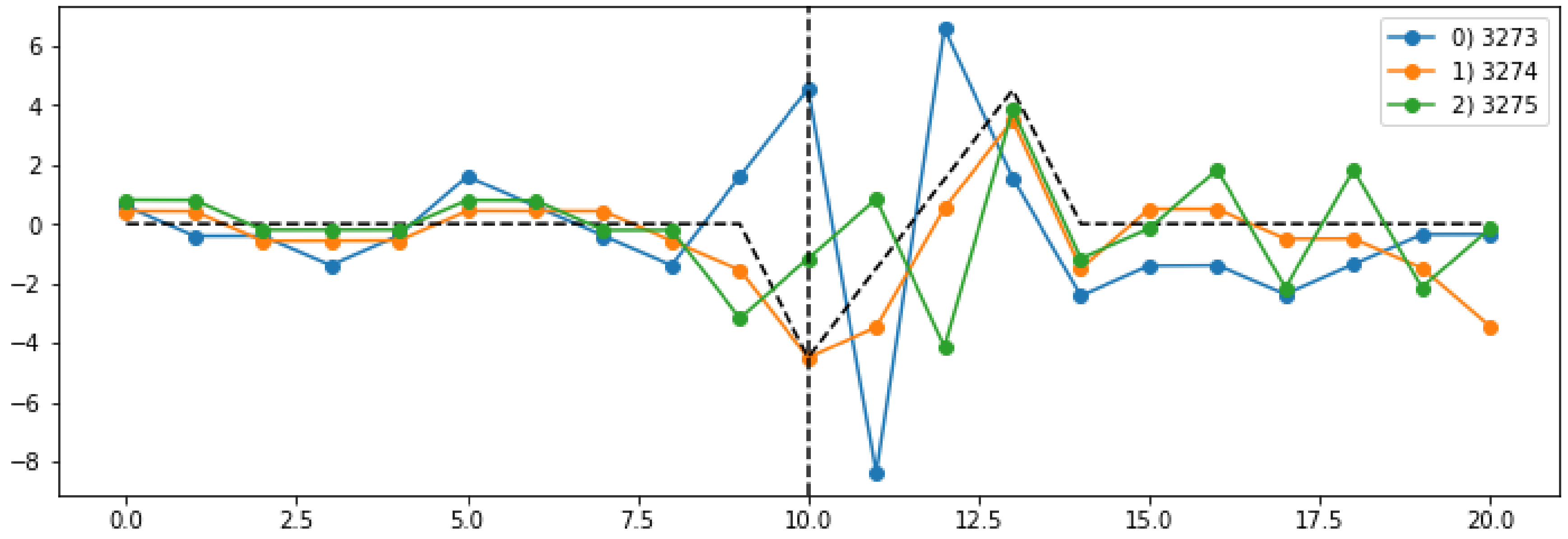


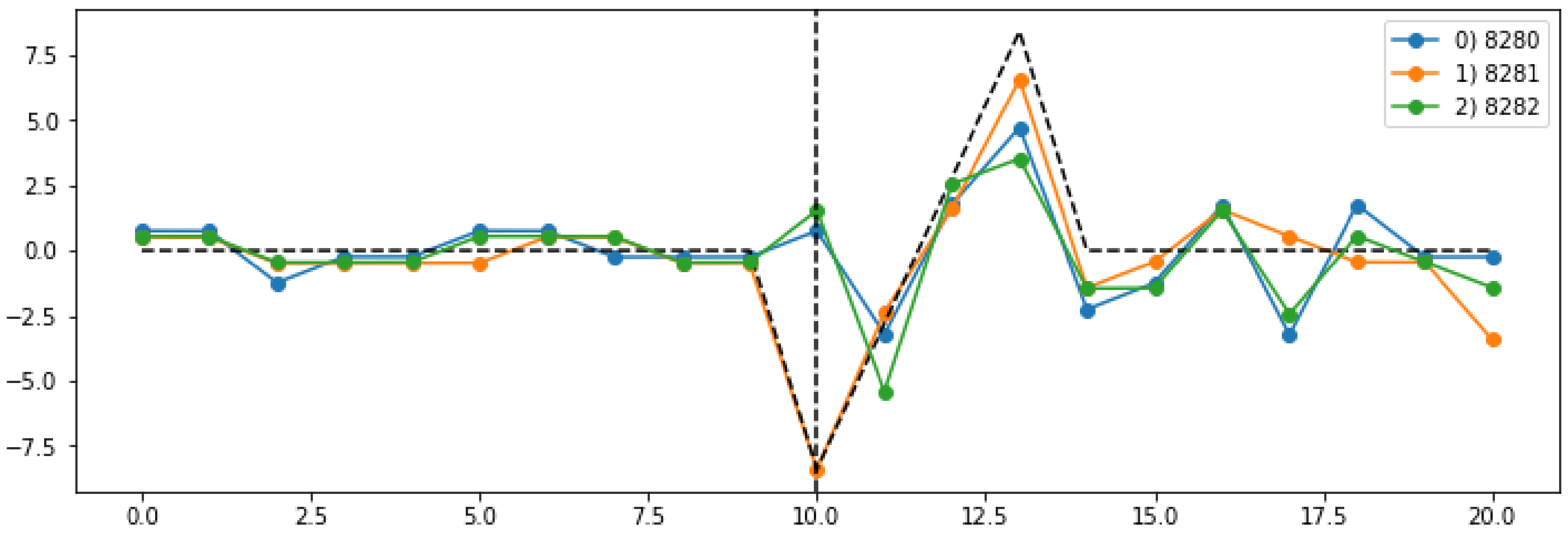
Sigid = 4225

Mark the noise of the peak produced by PD

03-2 Preprocessing

After Removing the noise





去除雜訊的峰值後，對剩餘的峰值計算特徵。

1.執行了一個稱為 calculate_peak_features 的 function

2. sawtooth_rmse 特徵：
針對每個剩餘峰值做鋸齒波形的模板之間的均方根誤差（RMSE）的計算。

03-3 Feature

- The absolute **height of the peak**.
- The **RMSE between the peak and a sawtooth shaped template**
  ➡ This sort of shape was common in traces marked as faulty.
- The absolute **ratio** of the **peak to the next data point**.
- The absolute **ratio** of the **peak to the previous data point**.
- The **distance to the maximum**, of opposite polarity to the peak, within a window of 5 either side of the peak.

03-3 Feature

peak_count_Q02

peak_count_total

peak_count_Q13

height_mean_Q02

height_std_Q02

ratio_prev_mean_Q02

ratio_next_mean_Q02

abs_small_dist_to_min_mean_02

sawtooth_rmse_mean_Q02

03-4 Model

- Model training：
  - Measurement_ID aggregates the corresponding peak features, possibly extracted from multiple.
  - Model training is based on measurement IDs, not signal IDs.

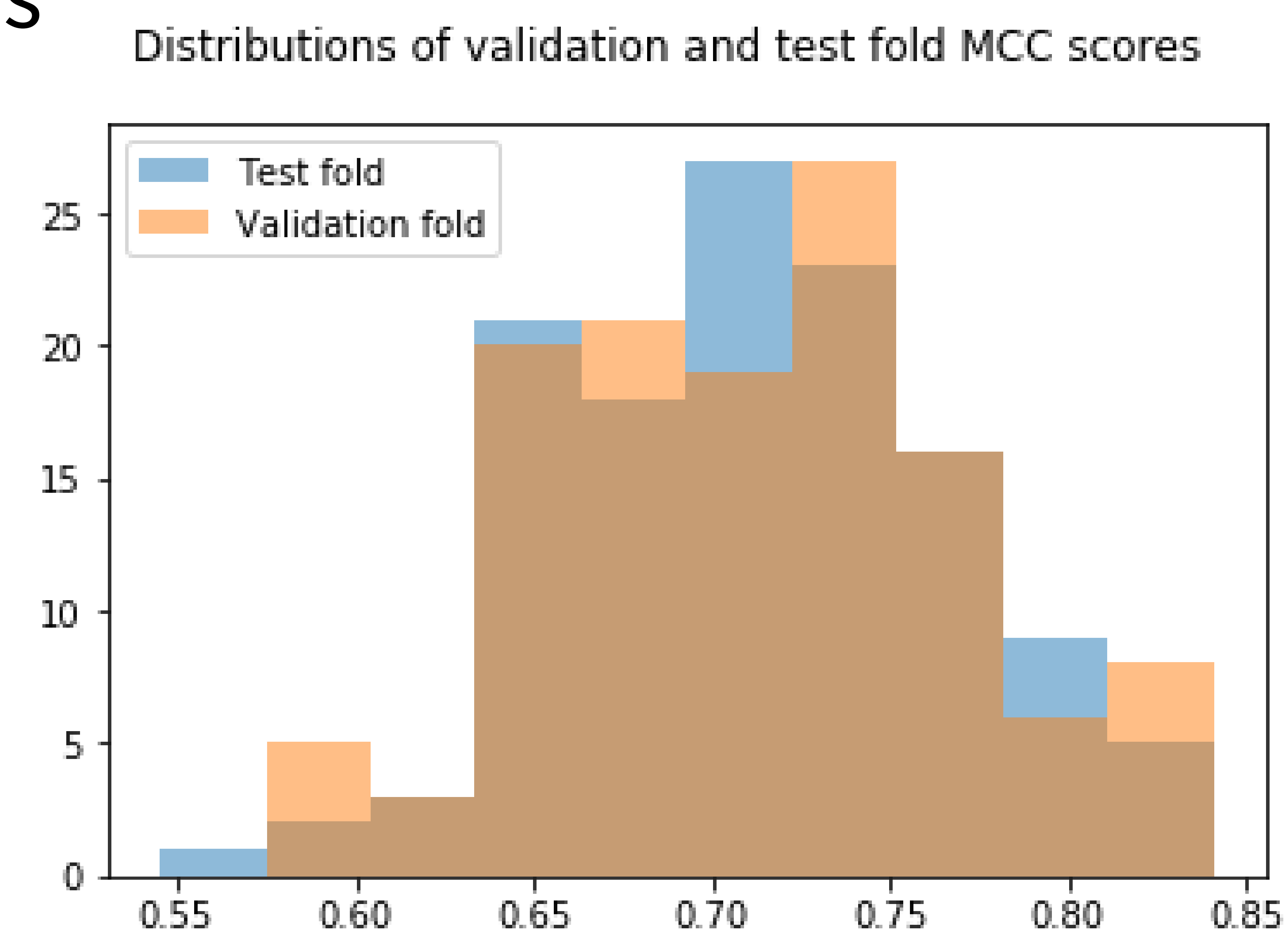
- The authors trained the model in units of measurement ID by **aggregating and processing the peak features of the signal**, and specifically processed the signal phase to **extract and use a range of features**.

03-4 Model

- Using LightBGM：
  - Model training
    - The LightGBM model is trained using the specified parameters.
  - Performance Metric Calculation
    - Calculates cross-validated Log Loss
    - Calculate Matthews correlation coefficient
  - Performance Metric Output
  - Model Prediction Visualization

03-4 Model

- Using LightBGM：
  - Model Prediction Visualization
    - Validation Fold MCC Scores
    - Test Fold MCC Scores



- Bin = 0.7 has better performance.

03-4 Model

1 **高效率**

可以處理具有數百萬個特徵的大型資料集。

3 **可擴展性**

可以快速擴展到數百萬個訓練範例和特徵。

2 **學習速度快**

透過執行更少的計算來減少找到最佳解。

4 **特殊處理**

可以處理稀疏或包含缺失值或離群值的資料。

2 **Hyperparameters**

有幾個重要的可能很難最佳化。

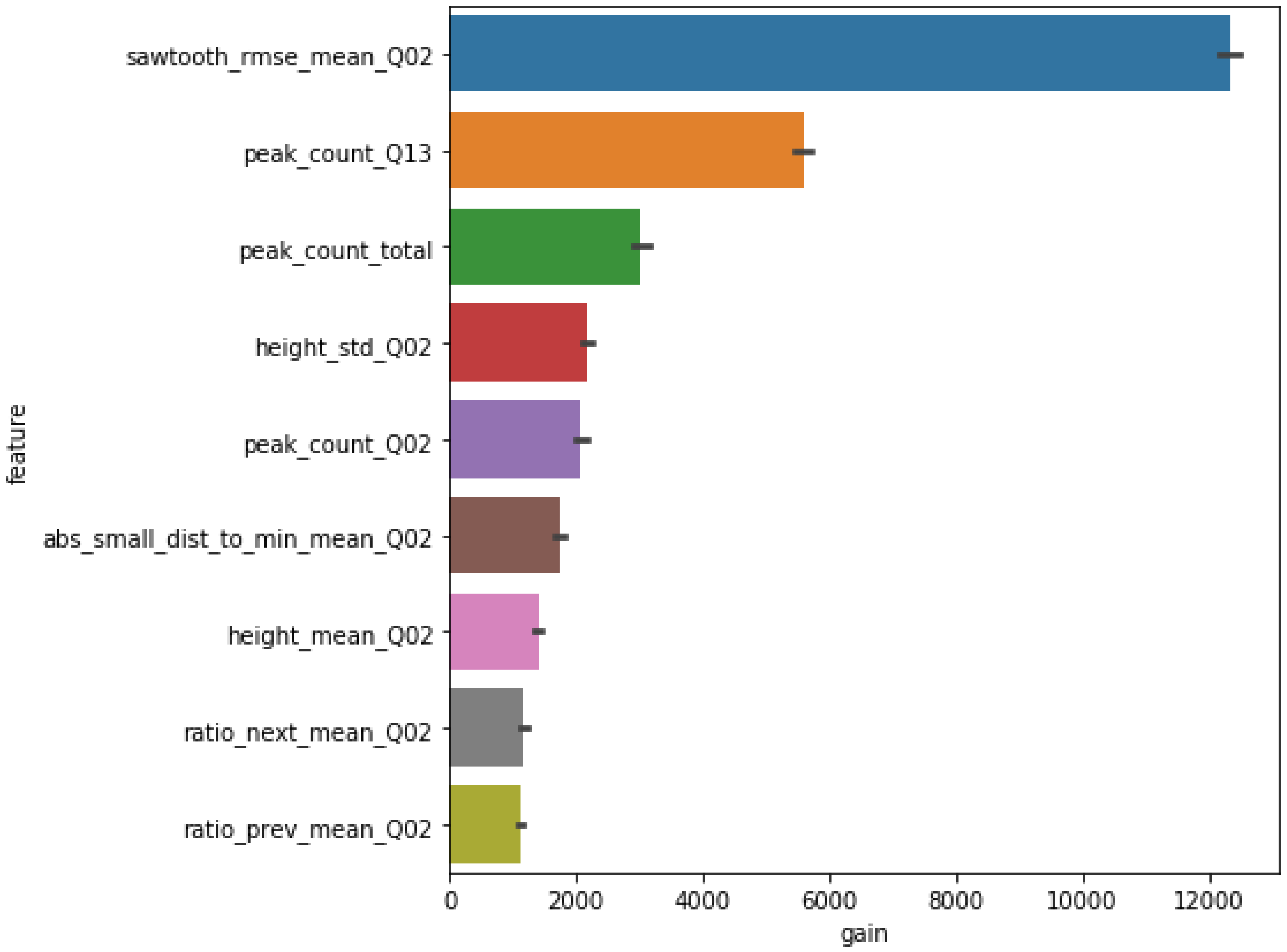
4 **時間和記憶體限制**

花費過多的時間和記憶體來運行。

03-5 Analysis

Feature importance

- Feature Importance Analysis
- Distinguishing Different Folds
- Integration of Results
- Saving the Results



03-5 Analysis

Predictions error

measurement_id	**error**	measurement_id	**error**
1380	0.004017	126	0.645305
2718	0.004069	1103	0.686100
2499	0.004094	1010	0.708625
2723	0.004114	1981	0.721519
244	0.004118	774	0.791745

Keys & Ideas

04-1 Understanding the data

- Learn to study the pattern
  - Partial discharge (PD)
  - Random pulses interference (RPI)
  - Discrete spectral interference (DSI)
- Various noise interference

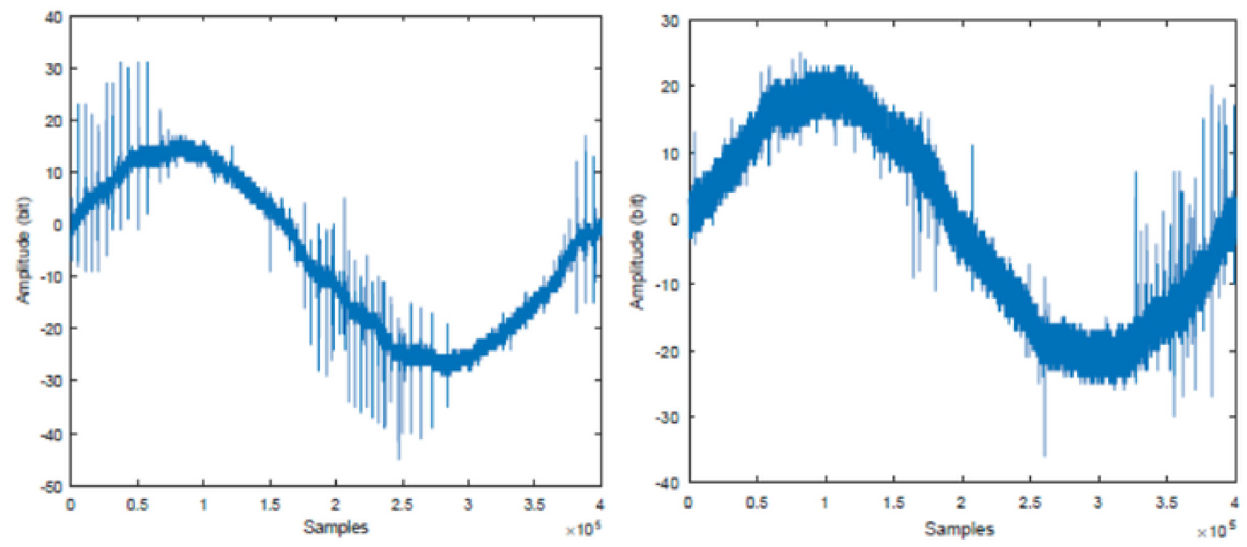


Figure 1.2: Measured signal snapshots in a raw state. Failure-free signal with corona discharge pulses (left) possess annotation 0 and fault indicating signal with higher EBN (right) possess annotation 6.

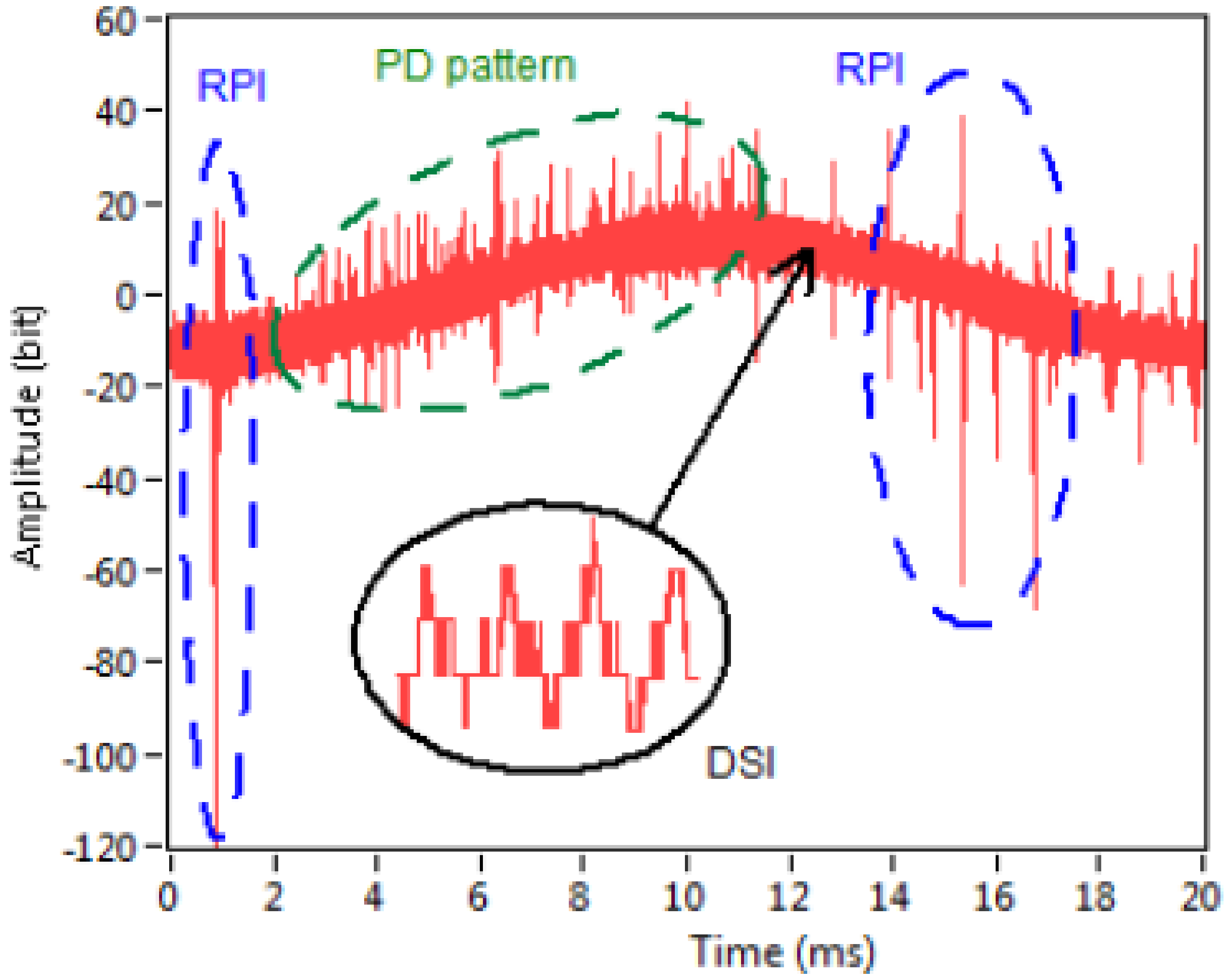
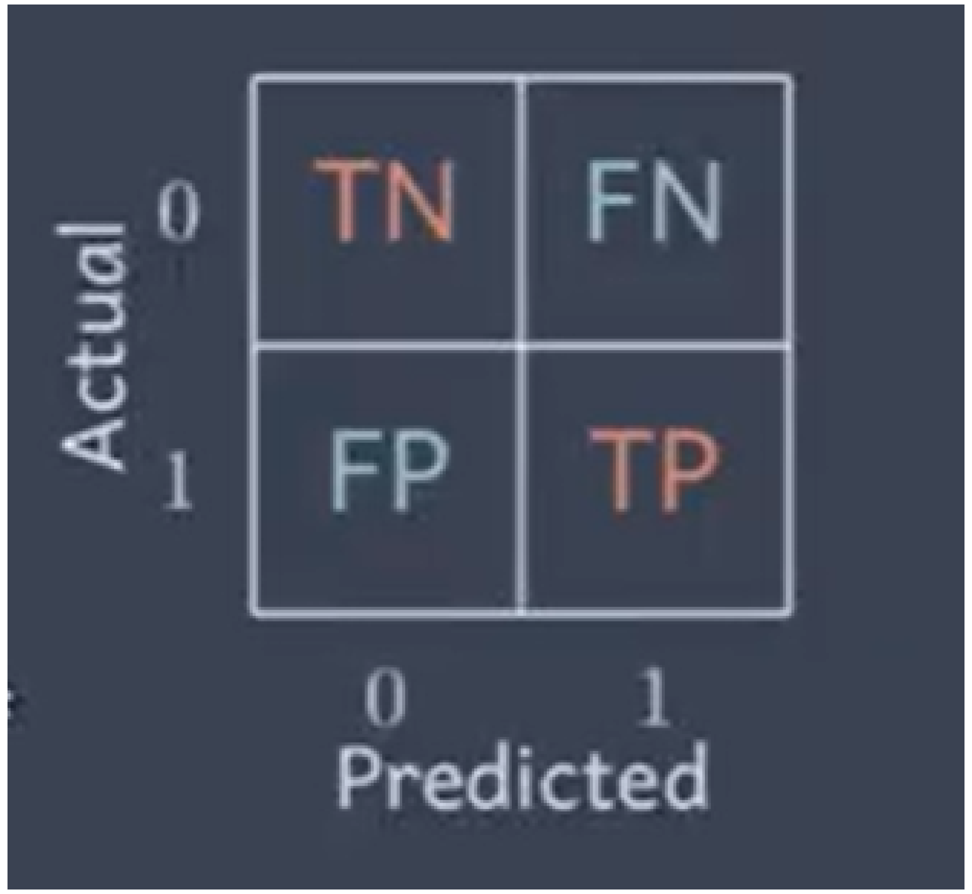


Figure 3. PD-pattern with noise during CC fault.

04-2 Brief Introduction to MCC

- Matthews correlation coefficient (MCC)
- A single-value metric that summarizes the confusion matrix.

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$



MCC=1

$FP = FN = 0$
$TP \neq 0$
$TN \neq 0$

MCC=0

$TP \times TN = FP \times FN$

MCC=-1

$TP = TN = 0$
$FP \neq 0$
$FN \neq 0$

04-3 Why MCC?

- Most used when in **binary classification**.
- Four entries are more **equally considered**.

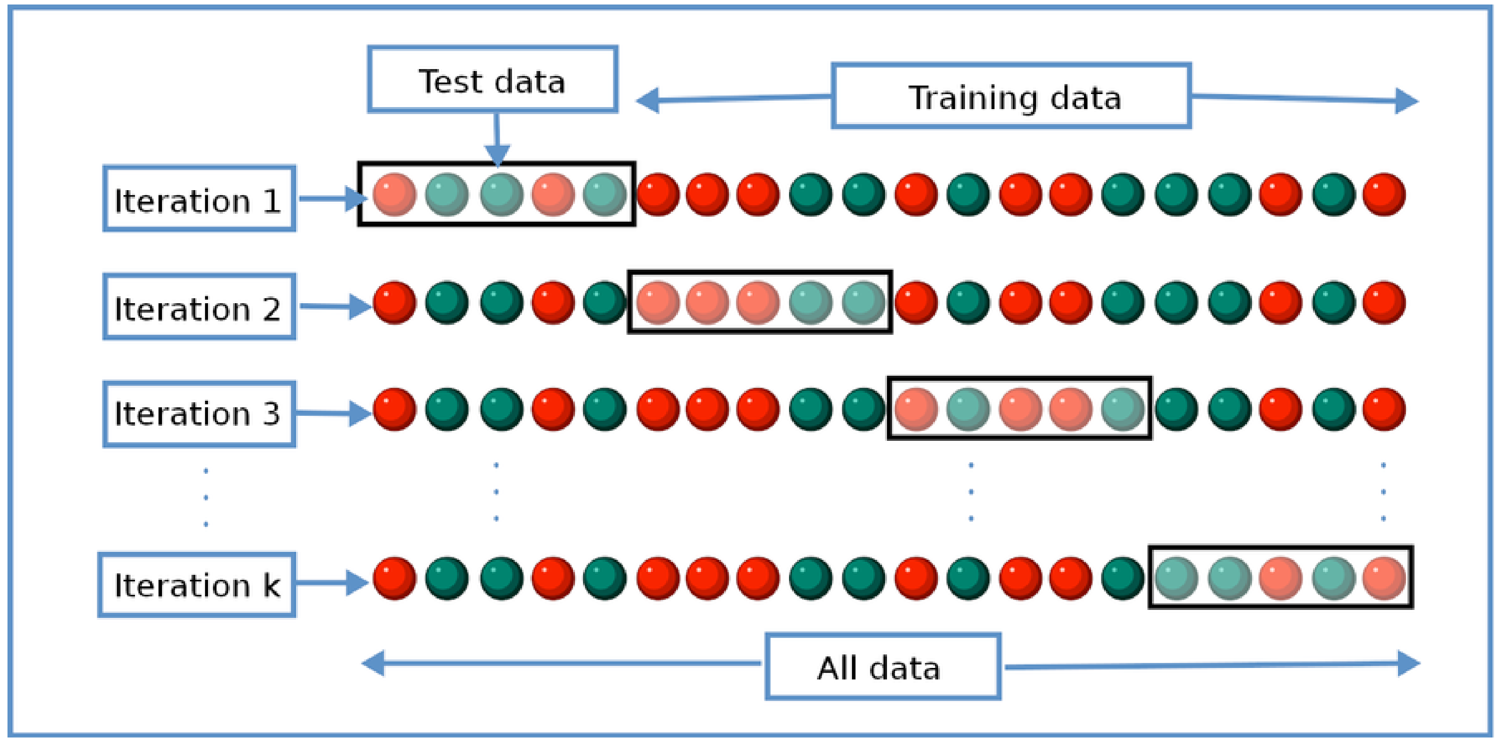
$$F1 \text{ score} = \frac{2TP}{2TP + FN + FP}$$

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

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

04-4 Choice of Validation

- **K-Fold Cross-Validation explained**



- **5-fold cross-validation**
  - Data set size is **large**, lower folds for **higher efficiency**
  - **Cheeper** than Leave One Out Cross-Validation (LOOCV)

04-5 Things to Be Aware Of

- Low sampling rate in the patended device therefore **publicly available dataset not recommended.**
- Labeling is hard, may cause **inaccurate “target” data.**
- Electrics knowledge required.

05 Reference-1

VSB Power Line Fault Detection：
https://www.kaggle.com/code/mark4h/vsb-1st-place-solution/notebook#Features
Champions’ method - mark4h：
https://www.kaggle.com/code/mark4h/vsb-1st-place-solution/notebook#Features
5 fold CV advantages：
https://blog.csdn.net/weixin_44299786/article/details/133085930
LightBGM algorithm：
https://dataaspirant.com/lightgbm-algorithm/

05 Reference-2

Matthew Correlation Coefficient:
https://towardsdatascience.com/matthews-correlation-coefficient-when-to-use-it-and-when-to-avoid-it-310b3c923f7e

The Definitive Guide to the Matthews Correlation Coefficient:
https://www.youtube.com/watch?v=u-Ez7trpNrM&t=516s

K-fold cross-validation:
https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fwhat-is-k-fold-cross-validation-5a7bb241d82f&psig=AOvVaw1y2DTyxjMsJGUhZaSEUK7n&ust=1701973382749000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCMDR44e3-4IDFQAAAAAdAAAAABAD

謝謝聆聽！