Machine Learning Final Project

0416303 楊博凱 0416323 黃右萱 0416305 王柏堯

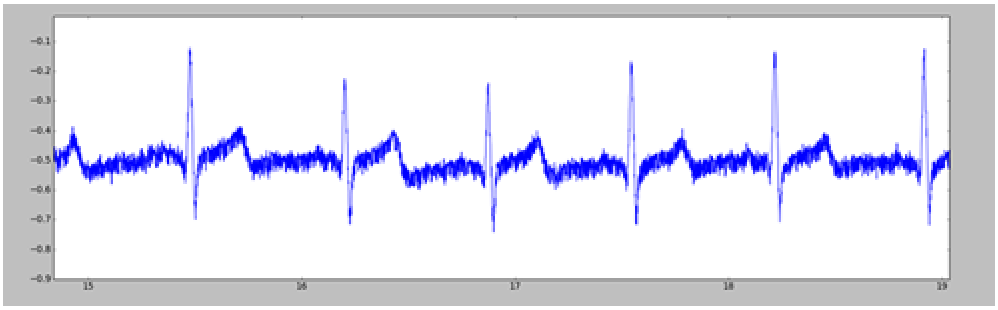
Scheme

特徵抽取→預測模型→綜合判斷

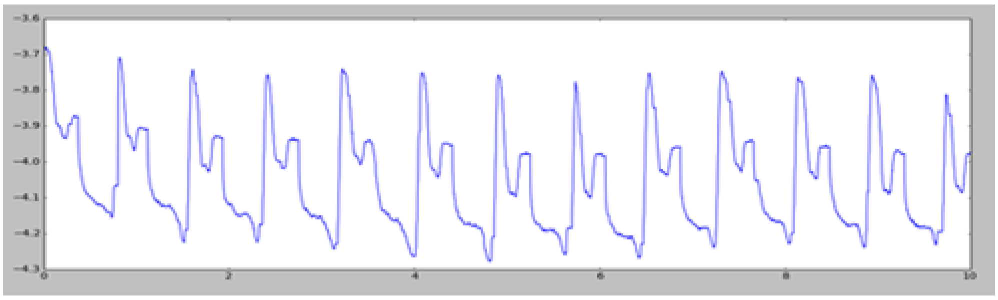
資料來源

訊號與系統實驗專題的資料，經過 IRB 認證
在身上配戴 ECG，脈搏，胸與腹的大小的測量裝置，測量觀看恐怖的影片與興奮的影片，以及平靜時的身體狀態，為了簡化計算量，我們就將這些訊號歸類為 1.害怕 2.興奮 3.平靜 來當作 target。
總共收集到 5 位同學願意協助給予資料，本次專案將專注於 ECG 與脈搏之分析。

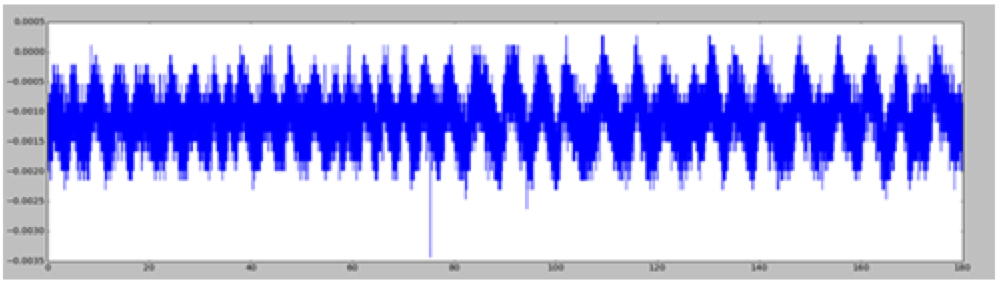
- ECG



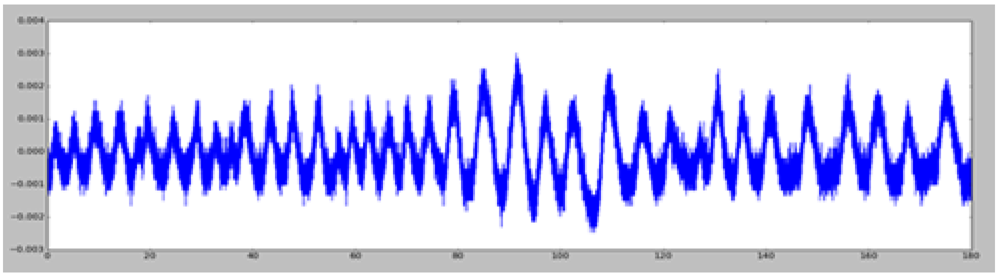
- Blood Pulse



- Chest



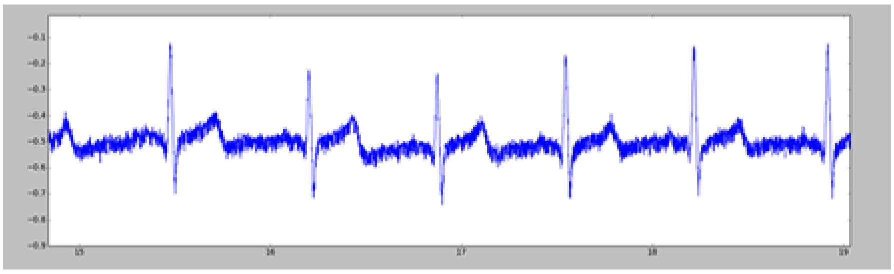
- Abdomen



特徵抽取

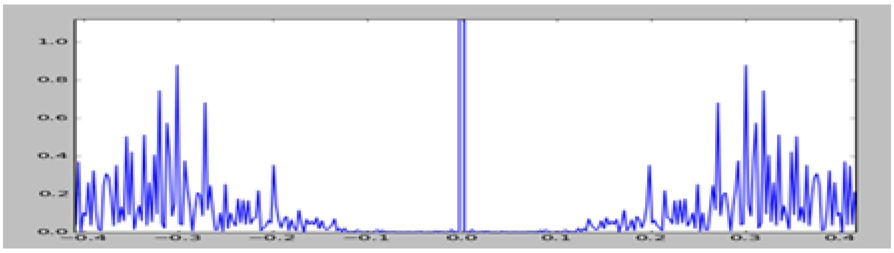
Naive

- 對時間做切割，直接作為 descriptive feature



Fourier Transform

- 對時間做若干切割，然後做快速傅立葉轉換，取其 Amplitude Spectrum 作為 descriptive Feature



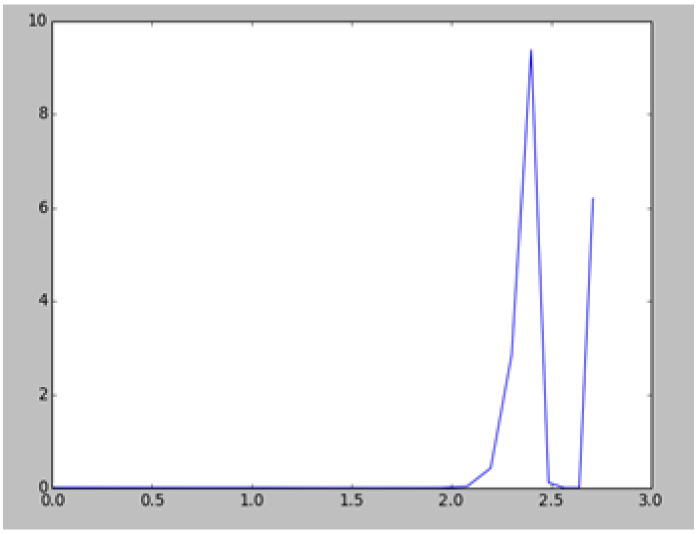
Empirical Mode Decomposition (EMD)

- Goal：
  - 找出非穩態訊號的不同趨勢訊號
  - 找出滿足 Hilbert 假設的趨勢訊號
    - 零均值
    - 局部對稱

- Scheme:
  - 迭代找出若干個 Intrinsic Mode Function(IMF)
    - 找平均包絡線
      - 上下包絡線之平均
      - 包絡線由對區域極值找 Cubic Spline(RK4)求得
    - 訊號鉗位
      - 當前訊號扣除平均包絡線
      - 直到滿足終止條件(可以訊號標準差等界定)
  - 找出 IMF 後將原本訊號扣除 IMF 得到 Residual Signal
    - 若 Residual Signal 為大致單調，則結束 EMD
    - 否則對 Residual Signal 求 IMF

Power Analysis

- Concept：找出重要的 IMF 分量進行分析
- 如下圖，各個 x 值對應到一個 IMF，我們將挑選出 peaked power IMF(s)



Down Sampling:

- Goal:
  - 增加可用 Feature 個數
  - 加速 EMD 拆解速度
    - 實作上原本數分鐘的處理能在秒至毫秒內完成
  - Concept：
    - 將訊號拆解成若干較低取樣精度的訊號
  - Description：

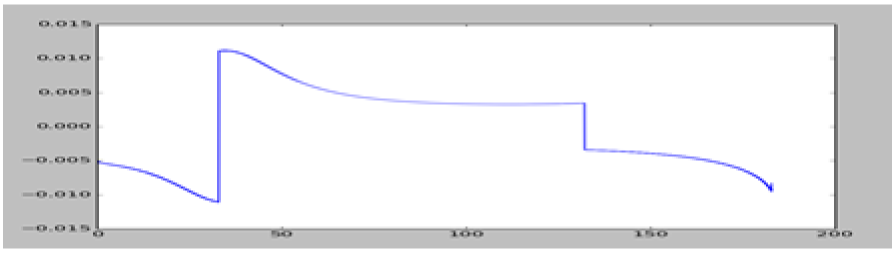
給定序列 $< a_k >$, $k \in [0, nm)$
取分割子序列集 $\{< b_j^i >: i \in [0, n)\}$ where $b_j^i = a_{j \cdot m+i}$

Hilbert Huang Transform (Modified for Feature Extraction)

- Hilbert Huang Transform :
  - Scheme :
    → Down Sampling
    → EMD
    → Power Analysis
    → Hilbert Transform
    → Obtain Instantaneous Phase & Frequency (IF)
    → Down Sampling
  - Hilbert Transform :

$$H\{u\}(t) = p.v. \int_{-\infty}^{\infty} \frac{u(\tau)}{\pi \cdot (t-\tau)} d\tau$$

  - From Hilbert to IF :
    - $\bar{u}(t) = H\{u\}(t)$
    - $s(t) = u(t) + j \cdot \bar{u}(t)$
    - $s(t) = m(t)e^{\theta(t)}$, for real valued $m(t), \theta(t)$
      - Phase Determination (Multiple Ways):
        - use arctan transformation
        - use absolute value
    - $f(t) = \frac{\omega(t)}{2\pi}$ where $\omega(t) = \frac{d\theta(t)}{dt}$

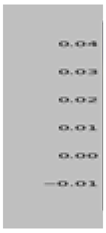


Direct Quadrature (Modified for Feature Extraction)

- Direct Quadrature :
  - Scheme :
    → Down Samplin
    → EMD
    → Power Analysis
    → Normalize

→ Obtain Instantaneous Phase & Frequency (IF)
→ Down Sampling

- Concept：
  - 對訊號標準化，並將之視為弦波函數取 IF
- Normalize：
  - 找出 IMF 絕對極值
  - 以 RK4 找出其上包絡線
  - 將整個訊號除去上包絡線
- 找 IF：



  - 考慮信號$u(t)$標準化後為$\bar{u}(t)$
  - 取出瞬時相位：$\theta(t) = tan^{-1}(\frac{\sqrt{1-\bar{u}^2(t)}}{\bar{u}(t)})$
  - 得到瞬時頻率：$f(t) = \frac{d\theta(t)}{dt}$

Learning Model

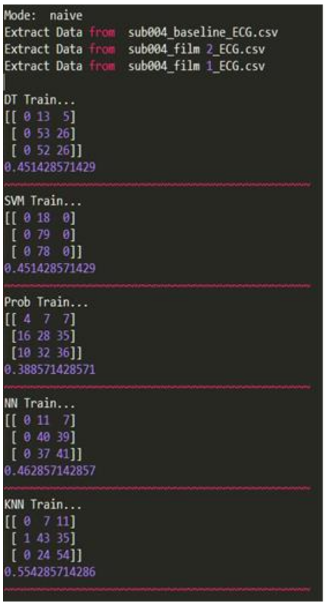
- 決策樹 (Decision Tree)：我們採用 Random Forest 對多個 Decision Tree 的結論做出綜合決策
  - n_estimators (Number of trees in the forest) = 5
  - max_depth (Maximum depth of the tree) = 3
- 支持向量機 (Support Vector Machine)：我們運用 sklearn.svm.SVC 來實作 SVM
- 高斯機率模型 (Gaussian Naive Bayssian)：我們運用 sklearn.naive_bayes.GaussianNB 來實作 Probability
- 類神經網絡 (Neural Network)：我們運用 sklearn.neural_network.MLPClassifier 來實作 NN
  - Solver (Solver for weight optimization) = 'lbfgs' (an optimizer of quasi-Newton methods)
  - Activation (Activation function for the hidden layer) = 'tanh' (returns f(x) = tanh(x))
- K-近鄰算法 (K-Nearest Neighbor)：我們運用 sklearn.neighbors.KNeighborsClassifier 來實作 KNN
  - n_neighbors (Number of neighbors) = 10
  - metric (Distance metric to use for the tree) = 'manhattan'
  - weights = 'distance' (Closer neighbors of a point will have greater influence than those which are far away)

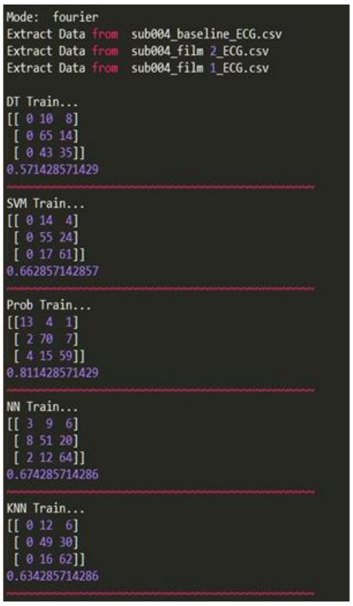
最後，實際使用時，我們製作了一個模組，傳入生理信息，將自動抽取出特徵，並做多次預測，最後以 voting 的方式決定最後預測結果。

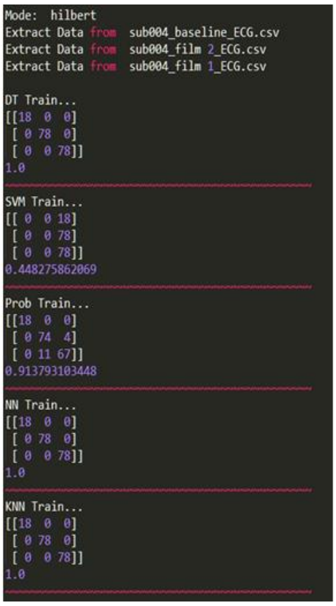
Validation Result

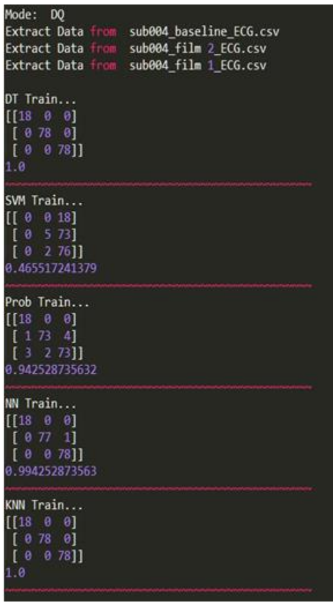
單人數據交叉審驗

- 採用同一人的生理信號，搗亂後分割為訓練以及測試數據集
- 10 折交叉審驗
- ECG

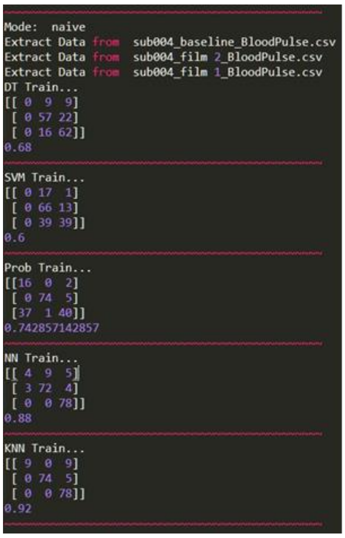


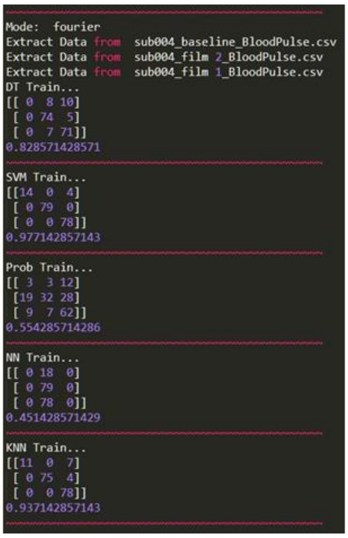


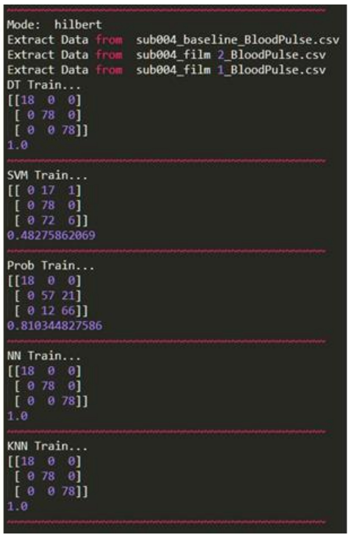


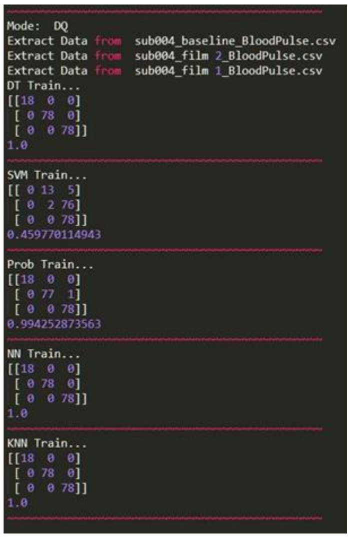


- 脈搏





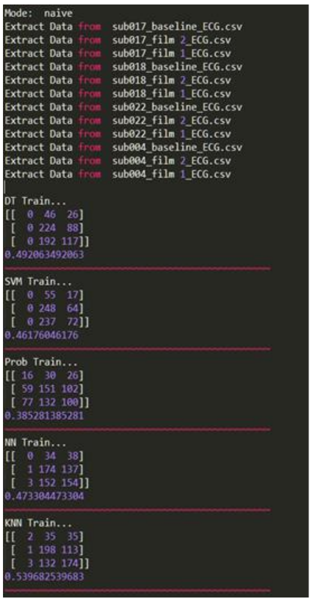
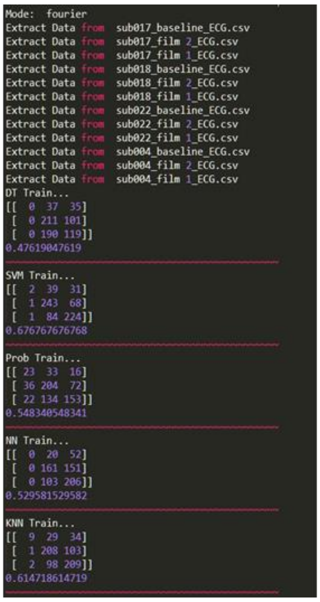
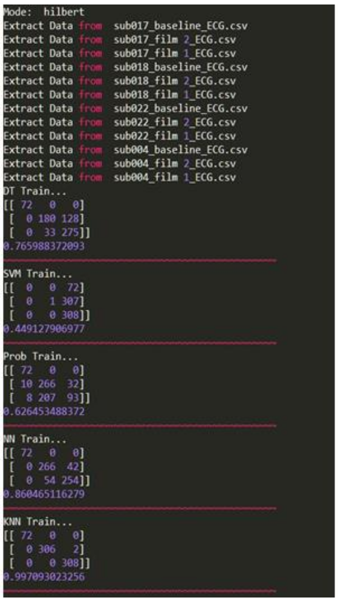
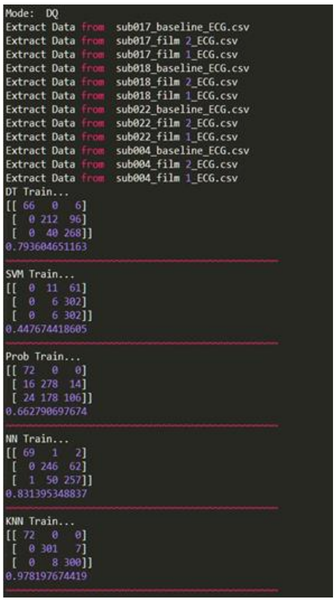




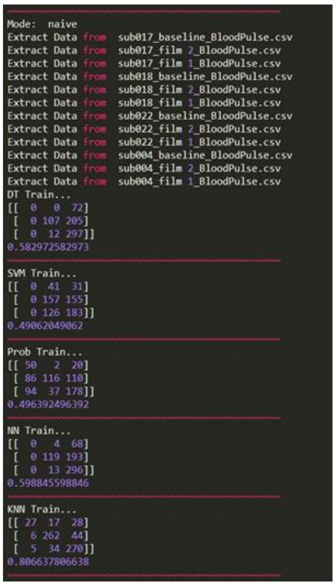
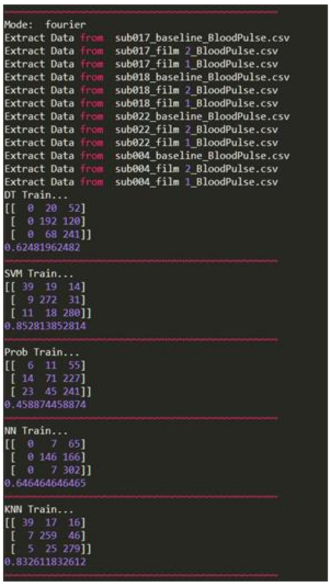
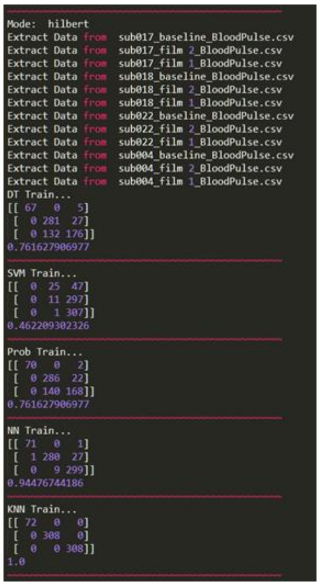
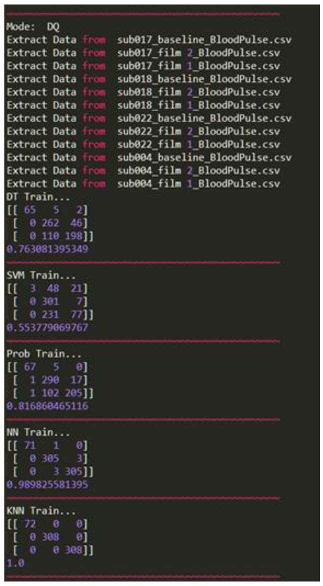
多人數據交叉審驗

- 採用多人的生理數據，搗亂後分割為訓練以及測試數據集
- 10 折交叉審驗

- ECG

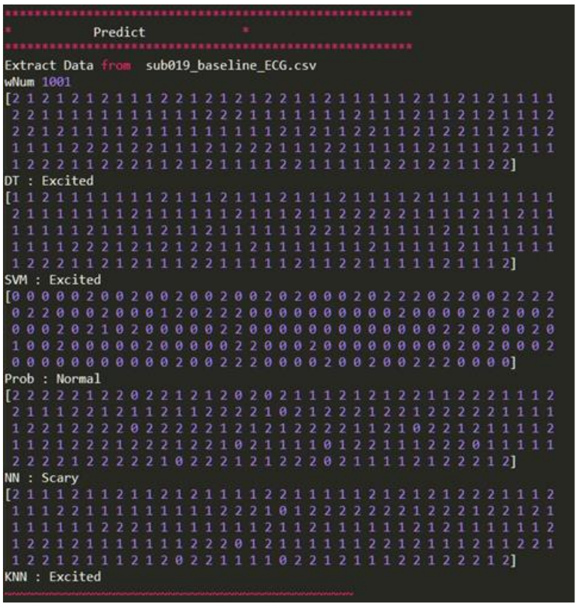
- 脈搏

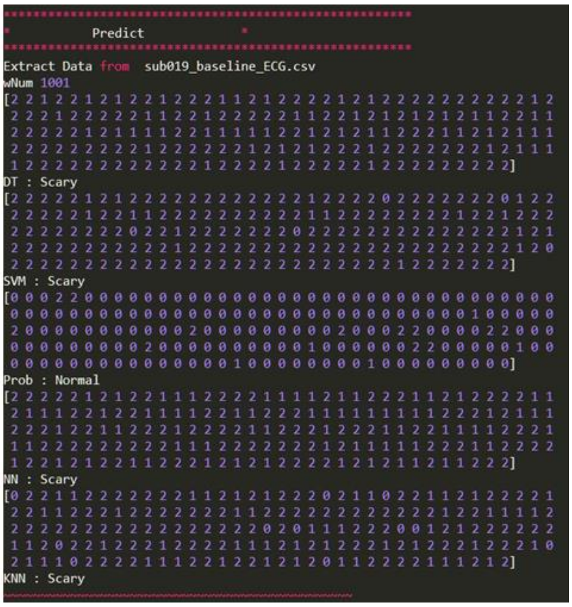





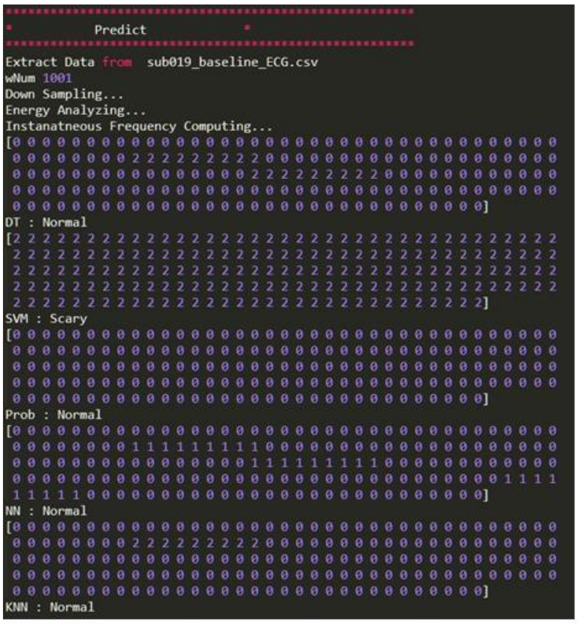
多人數據分割審驗

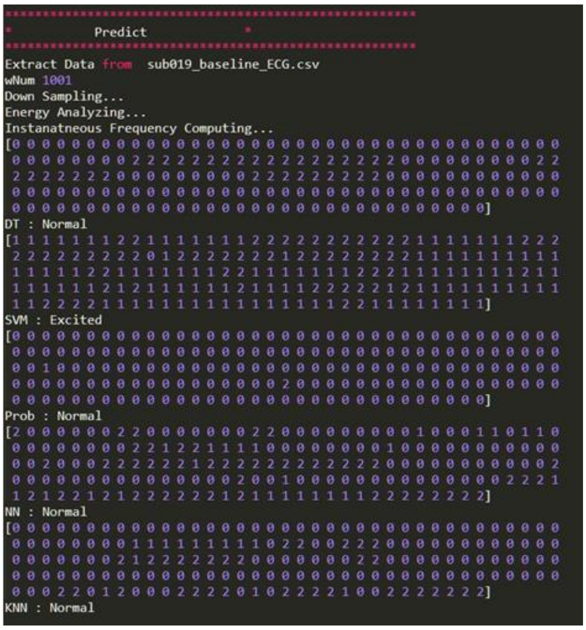
- 採用多人的生理數據訓練，並採用另外一人作為測試數據
- 此測試皆採用平靜狀態之數據做為測試

- ECG

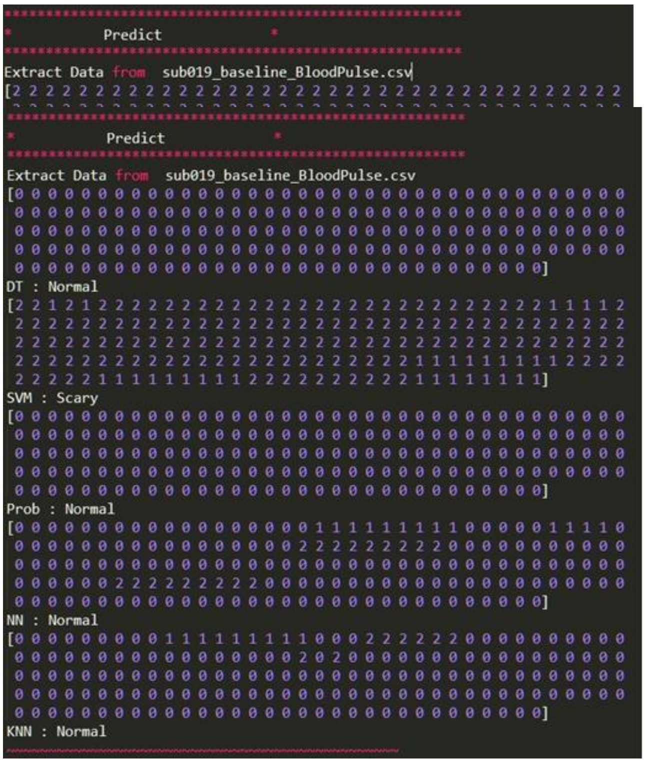

Predict
Extract Data from sub019_baseline_ECG.csv
wNum 1001
DT : Excited
SVM : Excited
Prob : Normal
NN : Scary
KNN : Excited

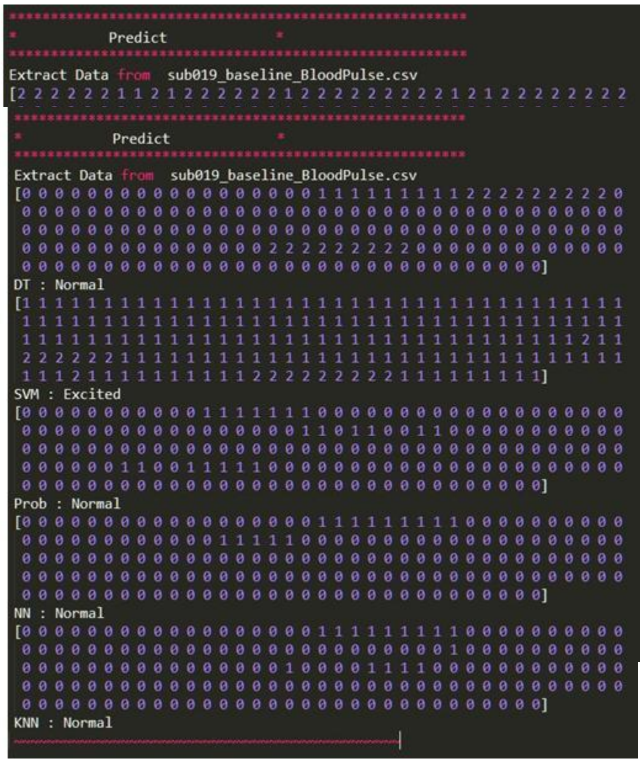

Predict
Extract Data from sub019_baseline_ECG.csv
wNum 1001
DT : Scary
SVM : Scary
Prob : Normal
NN : Scary
KNN : Scary


Predict
Extract Data from sub019_baseline_ECG.csv
wNum 1001
Down Sampling...
Energy Analyzing...
Instanatneous Frequency Computing...
DT : Normal
SVM : Scary
Prob : Normal
NN : Normal
KNN : Normal


Predict
Extract Data from sub019_baseline_ECG.csv
wNum 1001
Down Sampling...
Energy Analyzing...
Instanatneous Frequency Computing...
DT : Normal
SVM : Excited
Prob : Normal
NN : Normal
KNN : Normal

- 脈搏


Predict
Extract Data from sub019_baseline_BloodPulse.csv
Predict
Extract Data from sub019_baseline_BloodPulse.csv
DT : Normal
SVM : Scary
Prob : Normal
NN : Normal
KNN : Normal


Predict
Extract Data from sub019_baseline_BloodPulse.csv
Predict
Extract Data from sub019_baseline_BloodPulse.csv
DT : Normal
SVM : Excited
Prob : Normal
NN : Normal
KNN : Normal

結論

Feature Extraction

人體是非穩態系統，並不會滿足 Fourier 的週期性假設，所以想要單純使用傅立葉分析是不切實際的；再者，原訊號中存在許多測量誤差，以及干擾信號，所以也不應該是 naive；取出重要信息分量，而後對於當下的頻率去做分析才是合理的，所以 HHT 與 DQ 取出的瞬時頻率會有最好的結果。

Model

- DT 能有良好的結果，推測是因為 target 只有做簡單的分類，所以在選擇上較不會出錯
- SVM 的結果不好可能是這些訊號並沒有辦法用超平面去切成三區不同的類別
- probability 的結果為普通，可能因為 naive 的 model 又把一些誤差帶回來，使得運算加快，但準確率降低
- NN 能夠處理複雜的 feature，所以這次表現算是優良，但因為有多層神經網絡的關係，使得運算非常緩慢
- KNN 效果非常良好，推測原因為特徵空間中同狀態分類的訊號相似程度高

總結來說，我們要先用 HHT 或 DQ 處理訊號，使得準確率大幅上升，再配合 KNN 或 DT 來當作 model(速度較快)，但若要進一步使用，當 target 變多變複雜時，就要使用 NN 了。

*** 由於我們的資料部分無法公開，所以部分資料將不放到 zip 檔中 ***

****************************** 組員分工 ******************************

Learning Model & PPT：0416303 楊博凱

Feature Extraction & PPT：0416323 黃右萱

兩邊協調 & 上台報告：0416305 王柏堯