ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

# MÔN: PATTERN RECOGNITION

# AND MACHINE LEARING

Mid-term Project Report

# Mental attention states classification using EEG data


**Ngành:** Khoa học Dữ liệu

**Thành viên nhóm:**

Phan Bình Phương

Huỳnh Thảo Quỳnh

Nguyễn Ngọc Thanh Thư

Kiều Thị Ngọc Vui


***Thành phố Hồ Chí Minh, ngày 05 tháng 12 năm 2024***

# MỤC LỤC

# 1. Giới thiệu và Tổng quan về Dataset

## 1.1. Giới thiệu

Dự án này nhằm phân loại các trạng thái chú ý tinh thần — ***tập trung, không tập trung*** và ***buồn ngủ*** — dựa trên tín hiệu EEG từ thiết bị EMOTIV EEG. Dữ liệu được thu thập từ 34 thí nghiệm, mỗi thí nghiệm ghi nhận tín hiệu EEG từ các kênh 4 đến 17 với tần số mẫu 128 Hz. Mục tiêu là phát triển mô hình học máy chính xác để phân loại các trạng thái chú ý tinh thần.

## 1.2. Tổng quan về bộ dữ liệu EEG

Bộ dữ liệu EEG này được thu thập trong một thử nghiệm với 5 đối tượng (subjects), trong đó mỗi đối tượng thực hiện 7 thử nghiệm (trừ đối tượng thứ 5 chỉ thực hiện 6 thử nghiệm) để khảo sát ba trạng thái chú ý: ***tập trung, không tập trung,*** *và* ***buồn ngủ***. Mỗi đối tượng có 2 ngày đầu thử nghiệm và những ngày sau đó là thực nghiệm chính. **Vậy vậy, nhóm em quyết định lấy 5 ngày cuối (đối tượng 5 chỉ 4 ngày) để phân tích và huấn luyện mô hình.**

Mỗi trạng thái được ghi nhận trong các khoảng thời gian:

- **Tập trung**: Từ t = 0 đến t = 10 phút.
- **Không tập trung**: Từ t = 10 đến t = 20 phút.
- **Buồn ngủ**: Từ t = 20 phút trở đi.

**Cấu trúc và tính toán số lượng mẫu:** Dữ liệu EEG có tần số mẫu 128 Hz, tương đương với 128 * 60s (7680 mẫu/phút).

- **Tập trung**: Từ dòng 0 đến 76800.
- **Không tập trung**: Từ dòng 76800 đến 153600.
- **Buồn ngủ**: Từ dòng 153600 đến hết dữ liệu.

# 2. Data Loading & Preprocessing

Tín hiệu EEG chứa nhiều tần số khác nhau, và không phải tất cả các tần số đều có liên quan đến các hoạt động não mà chúng ta muốn phân tích. Vì vậy, sau khi tìm hiểu tổng quan về Dataset, nhóm em đề xuất các bước tiền xử lí dữ liệu như sau:

**Bước 1:** Tín hiệu EEG thô thường chứa nhiễu từ nhiều nguồn khác nhau. Để cải thiện chất lượng dữ liệu, nhóm em đã áp dụng 2 kỹ thuật lọc:

- **Lọc thông dải (Bandpass filter):** Kỹ thuật này loại bỏ các tần số ngoài dải từ **0.5Hz** đến **30Hz**, giúp tín hiệu EEG chỉ chứa các tần số liên quan đến hoạt động sóng não. Lọc thông dải loại bỏ nhiễu tần số thấp (như chuyển động mắt, cơ) và tần số cao (như nhiễu từ thiết bị điện tử), giúp cải thiện chất lượng tín hiệu và độ chính xác trong phân tích.

- **Lọc notch (Notch filter):** Bộ lọc này loại bỏ nhiễu tần số **50Hz** (hoặc **60Hz**), thường phát sinh từ nguồn điện lưới. Lọc notch giúp giảm thiểu ảnh hưởng của nhiễu điện từ môi trường, đảm bảo tín hiệu EEG chính xác và sạch hơn.

**Bước 2:** Sau khi thực hiện các bước lọc thông dải và lọc notch, nhóm em đã tiến hành trực quan hóa tín hiệu từ tất cả các kênh EEG trong một khoảng thời gian nhất định.

- Dựa trên các biểu đồ quan sát được, nhóm em lựa chọn các kênh có dao động mạnh mẽ và rõ ràng, đồng thời loại bỏ những kênh có dao động yếu và mờ nhạt.
- Các kênh hữu ích được chọn bao gồm: **F7**, **F3**, **P7**, **O1**, **O2**, **P8**, và **AF4**. Những kênh này có biên độ EEG dao động mạnh trong khoảng thời gian khảo sát, phù hợp để tiến hành phân tích sâu hơn.

**Bước 3:** Nhóm em tiếp tục áp dụng kỹ thuật **Phân tích thành phần độc lập (ICA)** để tách các thành phần tín hiệu độc lập từ dữ liệu EEG. Mục tiêu của bước này là loại bỏ các nhiễu không mong muốn, đặc biệt là nhiễu do chuyển động mắt (EOG) và hoạt động tim (ECG), qua đó giúp cải thiện chất lượng tín hiệu EEG và nâng cao độ chính xác trong phân tích sóng não.

Chi tiết quá trình thực hiện ICA:

1. *Áp dụng ICA*: Sử dụng thư viện MNE để áp dụng thuật toán ICA, tách tín hiệu EEG thành các thành phần độc lập giúp phân biệt tín hiệu não với các thành phần nhiễu không mong muốn, đồng thời giữ lại các thành phần có liên quan đến hoạt động não.
2. *Gán hệ thống Montage 10-20*: Để chuẩn hóa và xác định chính xác vị trí của các kênh EEG, nhóm em đã sử dụng hệ thống montage 10-20. Hệ thống này giúp xác định vị trí chính xác của các điện cực trên đầu người tham gia nghiên cứu và chuẩn hóa dữ liệu EEG. Montage 10-20 không chỉ đảm bảo tính chính xác trong quá trình phân tích, mà còn hỗ trợ việc tái tạo kết quả nghiên cứu.
3. *Phát hiện nhiễu EOG/ECG*: Trong quá trình xử lý tín hiệu EEG, nhiễu từ các yếu tố bên ngoài như chuyển động mắt (EOG) và hoạt động tim (ECG) có thể làm méo mó tín hiệu não và gây khó khăn trong việc phân tích các đặc trưng của sóng não. Để cải thiện chất lượng tín hiệu EEG, nhóm em đã áp dụng phương pháp ICA để phát hiện và loại bỏ các nhiễu này. Các kênh được chọn để phát hiện nhiễu liên quan đến mắt và tim được xác định dựa trên đặc điểm sinh lý học của chúng. Các kênh phổ biến liên quan đến chuyển động mắt bao gồm:
    - **AF4**: Kênh gần vùng trán
    - **F7**: Kênh gần vùng thái dương
    - **F3**: Kênh gần khu vực vỏ não phía trước

    Những kênh này được chọn vì chúng gần các vùng có thể bị ảnh hưởng trực tiếp bởi sự chuyển động của mắt, do đó giúp phát hiện và loại bỏ nhiễu EOG hiệu quả hơn.

    Nhóm em nhận thấy rằng các kênh được chọn không có sự can thiệp đáng kể từ hoạt động của tim trong quá trình khảo sát các kênh EEG.

    Sau khi áp dụng phương pháp ICA để phát hiện nhiễu EOG và ECG từ các kênh EEG, kết quả thu được như sau:

    - **Nhiễu EOG**: Không có chỉ số ICA nào bị nhiễu EOG trong các kênh EEG đã

chọn (F7, F3, P7, O1, O2, P8, AF4). Điều này cho thấy các kênh này ít bị ảnh hưởng bởi chuyển động mắt, giúp tín hiệu sạch hơn và phân tích sóng não chính xác hơn.

- **Nhiễu ECG**: Các kênh EEG cũng không bị nhiễu ECG đáng kể, duy trì chất lượng tín hiệu ổn định và không bị ảnh hưởng bởi tín hiệu tim.

4. Tiếp theo, nhóm em đã sử dụng phương pháp loại trừ các thành phần ICA không mong muốn.
    - Với đầu vào là danh sách các đối tượng ICA và Raw, cũng như các chỉ số thành phần ICA bị nhiễu do EOG và ECG.
    - Sau khi tổng hợp các chỉ số ICA bị nhiễu từ tín hiệu mắt và tim, thì tiến hành loại trừ các thành phần này khỏi quá trình phân tích.

**Tuy nhiên, đối với trường hợp này thì không có chỉ số ICA nào bị nhiễu EOG và ECG.**

## 3. Feature Engineering

Feature Engineering là một bước quan trọng trong quy trình phân tích dữ liệu EEG, nhằm trích xuất các đặc trưng quan trọng từ tín hiệu để phục vụ cho việc xây dựng các mô hình học máy.

### 3.1. Thực hiện đánh nhãn các trạng thái (focused, unfocused, drowsy)

Để phân loại các trạng thái chú ý tinh thần từ tín hiệu EEG, đầu tiên nhóm em phân chia dữ liệu thành ba trạng thái: *tập trung (focused)*, *không tập trung (unfocused)*, và *buồn ngủ (drowsy)* dựa trên khoảng thời gian thu thập dữ liệu. Mỗi trạng thái được gán với một khoảng thời gian cụ thể trong quá trình thử nghiệm.

Ở bước này nhóm em tổ chức lại dữ liệu cho từng đối tượng (subject) và trạng thái. Kết quả là một từ điển chứa dữ liệu của từng trạng thái chú ý, từ đó hỗ trợ cho quá trình phân tích và xây dựng mô hình.

### 3.2. Feature Extraction

Sau các bước tìm hiểu và tiền xử lí dữ liệu, nhóm em quyết định trích xuất đặc trưng là 4 dải tần ***delta***, ***theta***, ***alpha***, ***beta*** theo từng kênh (có 7 kênh) sẽ thu được 28 đặc trưng dựa trên sự liên quan của các dải tần này với các trạng thái tinh thần bằng cách phân tích công suất phổ (PSD).

#### 3.2.1. Các bước thực hiện

**Bước 1: Phân tích Short-Time Fourier Transform (STFT)**

- STFT là phương pháp phân tích tín hiệu trong cả không gian thời gian và tần số, cải tiến so với FFT bằng cách chia tín hiệu thành các đoạn nhỏ (epochs) và áp dụng FFT cho từng đoạn. Điều này giúp quan sát sự thay đổi phổ tần số theo thời gian, từ đó nắm bắt được các biến động trong hoạt động não bộ, đặc biệt là sự chuyển đổi giữa các trạng thái chú ý như tập trung, không tập trung, và

buồn ngủ.

- Để cải thiện độ chính xác, nhóm em sử dụng cửa sổ **Blackman-Harris**, giúp giảm thiểu hiện tượng *spectral leakage* (rò rỉ phổ) khi tín hiệu bị cắt đột ngột. Cửa sổ này làm mượt biên tín hiệu, giảm nhiễu và cải thiện độ chính xác của phép biến đổi Fourier, giúp phân tích tần số chính xác hơn.
- Tiếp theo, STFT được tính trên tín hiệu EEG để phân tích các thành phần tần số. Công suất phổ (power spectral density) được tính toán từ STFT, giúp phản ánh mức năng lượng của tín hiệu tại các tần số khác nhau. Đây là bước quan trọng để tiếp theo trích xuất các đặc trưng tần số phục vụ cho việc phân loại các trạng thái chú ý.

**Bước 2: Tính toán công suất trung bình theo dải tần số**

- Tín hiệu EEG được phân chia thành các dải tần số quan trọng: ***delta*** (0.5–4 Hz), ***theta*** (4–8 Hz), ***alpha*** (8–13 Hz), và ***beta*** (13–30 Hz), mỗi dải tần phản ánh một hoạt động não bộ đặc trưng và có mối liên hệ chặt chẽ với các trạng thái chú ý tinh thần.
- Để trích xuất đặc trưng từ các dải tần này, công suất trung bình của từng dải tần được tính toán cho mỗi kênh EEG. Quá trình này giúp phân tích mức năng lượng trong các dải tần số và làm nổi bật các đặc trưng liên quan đến từng trạng thái chú ý của người tham gia thử nghiệm.

**Bước 3: Áp dụng trung bình di động (Moving average window)**

- Bằng cách áp dụng trung bình di động với cửa sổ thời gian 15 giây (kết quả này là hợp lý nhất sau khi thử nghiệm trên nhiều window size khác nhau), tín hiệu EEG được làm mượt, giúp giảm nhiễu ngẫu nhiên và các biến động ngắn hạn trong tín hiệu được giảm bớt, giúp tín hiệu trở nên ổn định hơn và dễ dàng phân tích.

**Bước 4: Tạo Vector đặc trưng**

- Các đặc trưng công suất sau khi qua xử lý được chuyển thành vector đặc trưng, giúp mô hình học máy phân biệt các trạng thái chú ý. Phép logarit được áp dụng để giảm ảnh hưởng của các giá trị ngoại lai và đồng đều hóa đặc trưng.

### 3.2.2. So sánh các đặc trưng

Để so sánh các đặc trưng, nhóm em sử dụng ANOVA để kiểm tra sự khác biệt giữa các trạng thái, và T-test để so sánh cặp giữa các trạng thái chú ý.

- **ANOVA** cho thấy sự khác biệt có ý nghĩa thống kê giữa các trạng thái Focus, Unfocus, Drowsy ở tất cả các dải tần (p-value = 0.00000).
- **T-test**:
    - Focus vs Unfocus: Tất cả các dải tần (delta, theta, alpha, beta) có sự khác biệt rõ rệt (p-value = 0.00000), với hoạt động sóng não trong trạng thái Focus cao hơn.

- Focus vs Drowsy: Dải tần delta, theta, alpha có sự khác biệt (p-value < 0.05), nhưng beta không có sự khác biệt rõ rệt (p-value = 0.05888).
- Unfocus vs Drowsy: Tất cả các dải tần có sự khác biệt có ý nghĩa (p-value = 0.00000), với hoạt động sóng não trong trạng thái Unfocus thấp hơn.

**Kết luận:**

- Các trạng thái Focus, Unfocus, Drowsy có ảnh hưởng rõ rệt đến sóng não ở các dải tần.
- Dải tần Beta không phân biệt rõ giữa Focus và Drowsy, có thể do đặc điểm sóng não riêng biệt của từng người.

Ngoài ra, nhóm em còn sử dụng biểu đồ Topo map thể hiện mức độ hoạt động của các dải tần (Delta, Theta, Alpha, Beta) trên toàn bộ vùng đầu, với các khu vực màu vàng thể hiện công suất cao hơn, và màu xanh lá thể hiện công suất thấp.

- **Delta**: Phân bố mạnh ở phía bên trái đầu, liên quan đến trạng thái **buồn ngủ** hoặc **mệt mỏi**.
- **Theta**: Hoạt động mạnh ở vùng trước trán và vùng đỉnh có thể liên quan đến sự phân tán chú ý hoặc **buồn ngủ**.
- **Alpha**: Công suất cao ở vùng chẩm, thể hiện trạng thái **thư giãn** hoặc **không tập trung**.
- **Beta**: Rõ nhất ở vùng giữa trán đôi khi mở rộng sang vùng trán bên, liên quan đến trạng thái **tập trung cao độ**.

# 4. Model Development

## 4.1. Modeling và phát hiện ra vấn đề (Insight của data)

Sau khi hoàn tất việc chuẩn bị dữ liệu, nhóm em tiến hành bước huấn luyện mô hình và đánh giá hiệu quả. Quy trình thực hiện như sau:

- **Bước 1:** Chia dữ liệu thành tập huấn luyện (train) và tập kiểm tra (test).
- **Bước 2:** Scale dữ liệu của tập huấn luyện, sau đó áp dụng cùng một bộ chuẩn hóa cho tập kiểm tra.
- **Bước 3:** Áp dụng mô hình (ở đây nhóm chọn Random Forest để thử) và đánh giá các chỉ số hiệu suất.

Ban đầu, nhóm em gộp tất cả dữ liệu từ các file vào một bộ dữ liệu lớn, sau đó chia ngẫu nhiên thành các tập train và test. Sau khi scale dữ liệu, sử dụng mô hình Random Forest để đánh giá, và kết quả cho thấy **avg accuracy = 0.99**. Tuy nhiên, kết quả này khiến nhóm em đặt ra nghi vấn về khả năng *data leakage* hoặc *overfitting*.

Khi xem xét lại quy trình, nhóm nhận thấy một vấn đề trong bước chia dữ liệu. Cụ thể, khi sử dụng trung bình di động với cửa sổ 15 giây, mỗi điểm dữ liệu bị ảnh hưởng bởi các phân đoạn liền kề. Nếu chia dữ liệu một cách ngẫu nhiên, có thể dẫn đến việc **chồng lấn thông**

**tin** giữa tập train và test, làm rò rỉ dữ liệu từ tập huấn luyện sang tập kiểm tra.

Để giải quyết vấn đề này, nhóm em quyết định điều chỉnh cách chia dữ liệu sao cho các tập train và test **hoàn toàn độc lập**.

**(*) Cách thực hiện mới:**

- Với mỗi người, lấy **file cuối** của họ gộp lại làm **tập kiểm tra**, và các **file còn lại** gộp lại làm **tập huấn luyện**.
- Một cách khác, huấn luyện riêng biệt cho mỗi người (4 file cho tập huấn luyện và file cuối cho tập đánh giá), bảo đảm mô hình học chính xác đặc trưng của từng cá nhân.

Như vậy, dữ liệu trong tập huấn luyện và kiểm tra sẽ không bị trùng lặp, giúp đảm bảo tính chính xác của quá trình huấn luyện và đánh giá.

## 4.2. Model theory & Model building

Việc phân loại trạng thái chú ý tinh thần từ tín hiệu EEG đóng vai trò quan trọng trong nghiên cứu chức năng não bộ và các ứng dụng y sinh. Ở đây, nhóm em sử dụng hai mô hình học máy: **Support Vector Machine (SVM)** và **Random Forest**.

### 4.2.1. Suppot Vector Machine (SVM)

**SVM** là một thuật toán học có giám sát, phổ biến trong các bài toán phân loại và hồi quy. SVM hoạt động bằng cách tìm kiếm một *siêu phẳng (hyperplane)* tối ưu trong không gian nhiều chiều để phân tách các điểm dữ liệu thành các lớp khác nhau. Mục tiêu của SVM là tối đa hóa *lề (margin)*, tức là khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp. Việc tối đa hóa lề giúp cải thiện khả năng khái quát của mô hình và giảm thiểu lỗi phân loại.

Với khả năng sử dụng **kernel tricks**, SVM có thể xử lý dữ liệu **phi tuyến tính**, điều này rất quan trọng khi phân loại tín hiệu EEG, nơi các mối quan hệ giữa các đặc trưng của tín hiệu và các trạng thái chú ý có thể không tuyến tính. Đặc biệt, SVM rất thích hợp khi các trạng thái chú ý có sự **chồng lấn** về đặc trưng, ví dụ khi sóng *alpha* và *beta* có sự giao thoa trong tín hiệu EEG. SVM có thể sử dụng các kernel để ánh xạ tín hiệu EEG vào không gian đặc trưng cao hơn, nơi các lớp dữ liệu có thể được phân tách rõ ràng hơn.

### 4.2.2. Random Forest

Random Forest là một thuật toán học máy kết hợp (ensemble learning), sử dụng nhiều cây quyết định (decision trees) để tạo ra một mô hình dự đoán mạnh mẽ. Mỗi cây quyết định phân chia dữ liệu dựa trên các đặc trưng PSD của tín hiệu EEG và kết quả cuối cùng được đưa ra bằng cách kết hợp các dự đoán từ tất cả các cây quyết định.

Thuật toán giảm thiểu overfitting nhờ kết hợp ngẫu nhiên các cây quyết định, đồng thời đánh giá tầm quan trọng của các đặc trưng PSD, giúp phân tích rõ hơn vai trò của các dải tần trong phân loại trạng thái chú ý.

Quan trọng hơn, Random Forest có ưu điểm đặc biệt trong việc xử lý dữ liệu **mất cân bằng**. Khi các lớp không phân bổ đều, thuật toán có thể sử dụng **kỹ thuật bỏ phiếu đa**

**số** để cân bằng ảnh hưởng của mỗi lớp trong quá trình ra quyết định. Với khả năng tự điều chỉnh trọng số giữa các lớp, Random Forest giúp mô hình phân loại tốt hơn các lớp ít gặp, như lớp *tập trung* và *không tập trung*, trong khi vẫn giữ được độ chính xác khi phân loại lớp *buồn ngủ* có tỷ lệ lớn hơn.

### 4.3. Thực hiện modeling với dữ liệu imbalance áp dụng phương pháp chia tập train, test (*) như đã đề cập ở trên

**Phương pháp 1:**

Chia thành 2 model phân loại (ở đây nhóm em sử dụng SVM RBF Kernel) với:

- Model 1: Phân loại Drowsy và Not-Drowsy (Focused, Unfocused)
- Model 2: Phân loại Focused và Unfocused

Cách tiếp cận này giúp giảm thiểu tác động của việc mất cân bằng dữ liệu (khi lớp Drowsy có thể chiếm ưu thế hơn). Việc chia nhỏ bài toán thành các mô hình độc lập giúp mỗi mô hình tập trung vào một phân loại rõ ràng hơn. Đồng thời, phương pháp này giúp tối ưu hóa kết quả trên từng nhóm lớp mà không bị ảnh hưởng bởi sự chênh lệch số lượng giữa các lớp.

**Phương pháp 2**

Giảm số lượng mẫu của lớp Drowsy để tạo ra một phân bố cân bằng giữa các lớp, số lượng data giảm được kiểm chứng qua nhiều lần test. Việc này giúp tránh hiện tượng mô hình thiên lệch vào lớp chiếm ưu thế (Drowsy), đảm bảo rằng mô hình không bị ảnh hưởng quá nhiều bởi lớp này khi học và phân loại. Sau khi cân bằng, chúng em áp dụng hai mô hình phân loại sau để so sánh hiệu suất của từng mô hình

- SVM với RBF Kernel
- Random Forest

**Đánh giá metrics của các mô hình ở cả 2 phương pháp trên**

| **Metrics** | **2 – SVM (PP1)** | **SVM** | **Random Forest** |
|---|---|---|---|
| **Accuracy** | 0.56 | 0.63 | 0.66 |
| **F1-score** | 0.51 | 0.60 | 0.62 |
| **Precision** | 0.51 | 0.60 | 0.64 |
| **Recall** | 0.54 | 0.61 | 0.63 |

**Nhận xét:**

Random Forest vượt trội so với SVM và 2-SVM trên cả bốn chỉ số chính (accuracy, F1-score, precision, recall), đặc biệt là trong độ chính xác và F1-score. Với dữ liệu EEG trong bài toán phân loại trạng thái chú ý (focused, unfocused, drowsy), Random Forest có ưu điểm sau:

1. *Xử lý dữ liệu phi tuyến:* Dữ liệu EEG phức tạp và có nhiễu, với các biên phân cách không rõ ràng. Random Forest học tốt hơn SVM nhờ khả năng xử lý mối quan hệ phi tuyến.

2. *Xử lý sự khác biệt cá nhân*: EEG có sự khác biệt lớn giữa các cá nhân, và Random Forest không bị phụ thuộc vào các đặc trưng riêng biệt như SVM, giúp xử lý dữ liệu đa dạng tốt hơn.
3. *Khả năng xử lý imbalance*: Dữ liệu EEG có sự phân bố không cân bằng, và Random Forest có thể tự điều chỉnh để xử lý vấn đề này mà không bị thiên lệch.
4. *Khả năng tổng quát*: Random Forest tổng quát tốt hơn và giảm thiểu overfitting, sử dụng nhiều đặc trưng để ra quyết định, trong khi SVM và 2-SVM cần điều chỉnh tham số phức tạp.

Như vậy, Random Forest hiệu quả hơn SVM và 2-SVM trong phân loại trạng thái chú ý nhờ khả năng xử lý dữ liệu phức tạp và mất cân bằng.

## 5. Đề xuất phương pháp cải thiện độ chính xác

Sau khi nghiên cứu kỹ hơn về bộ dữ liệu EEG, nhóm em nhận thấy rằng vì đây là dữ liệu từ BCI (Brain-Computer Interface), mỗi cá nhân có đặc điểm sóng não khác nhau, do đó cách tiếp cận trong việc phân chia tập train và tập test cần phải được điều chỉnh sao cho phù hợp với đặc tính này. Thay vì chia dữ liệu ngẫu nhiên thành các tập train và test, một phương pháp phổ biến trong nghiên cứu với dữ liệu EEG là sử dụng dữ liệu của nhiều buổi thu thập sóng não của cùng một cá nhân.

Cụ thể, phương pháp này thường bao gồm các bước sau:

- Sử dụng 5 models riêng biệt cho từng subject khác nhau
- Chuẩn bị data train, test cho mỗi model:
  - Data cho train: Lấy dữ liệu từ 4 buổi thử nghiệm đầu tiên của subject tương ứng (người cuối cùng là 3 buổi thử nghiệm đầu tiên)
  - Data cho test: Sử dụng dữ liệu từ buổi thử nghiệm cuối cùng của subject đó.

Với cách chia tập này, nhóm em đã cải thiện một phần khả năng tổng quát của mô hình, giúp mô hình hoạt động tốt hơn trên các dữ liệu thu thập sau này và giảm thiểu sự thiên lệch do việc huấn luyện trên nhiều subject khác nhau.

## 6. Kết luận

Qua dự án, nhóm đã phân tích và so sánh hiệu suất của các mô hình học máy trên dữ liệu EEG, nhằm phân loại các trạng thái chú ý tinh thần (tập trung, không tập trung, buồn ngủ). Kết quả cho thấy mô hình Random Forest vượt trội hơn SVM nhờ khả năng xử lý dữ liệu phi tuyến và đối phó tốt với sự mất cân bằng trong dữ liệu. Tuy nhiên việc lựa chọn mô hình cần cân nhắc giữa độ chính xác và hiệu suất xử lý, tùy thuộc vào mục tiêu cụ thể của ứng dụng. Việc tiếp tục nghiên cứu và tối ưu hóa các đặc trưng có thể giúp cải thiện thêm khả năng phân loại, đặc biệt là với các lớp có sự tương đồng trong dữ liệu EEG. Những kết quả này hứa hẹn đóng góp tích cực vào các ứng dụng y sinh và giao diện não-máy tính (BCI).

## 7. Tài liệu tham khảo


[1] [Machine Learning] Sử dụng sóng não EEG để nhận biết và kiểm soát giấc ngủ của con người!

[2] Sci-Hub | Distinguishing mental attention states of humans via an EEG-based passive BCI using Machine Learning Methods. Expert Systems with Applications | 10.1016/j.eswa.2019.05.057

[3] EEG_Protocol_3_States

[4] Repairing artifacts with ICA — MNE 1.8.0 documentation

[5] Qingkai's Blog: Signal Processing: Why do we need taper in FFT