Báo cáo đồ án giữa kỳ
Mental Attention States Classification Using EEG Data

Ngày 5 tháng 12 năm 2024

Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia TP. Hồ Chí Minh

Khoa: Toán - Tin học

Môn học: Nhận dạng mẫu

Thông tin thành viên nhóm

STT	Họ và tên	MSSV
1	Võ Minh Thịnh	22280087
2	Nguyễn Trường Thịnh	22280086
3	Trần Bình Phương	22280071
4	Nguyễn Hồng Sơn	22280078

Học kỳ: HKI (2024-2045)

Ngày nộp báo cáo: Ngày 5 tháng 12 năm 2024

Mục lục

1 Giới thiệu về dữ liệu

1.1 Bối cảnh nghiên cứu bài toán

Trong bản cáo này là nghiên cứu về vấn đề phát hiện các sự thay đổi trạng thái tinh thần ở con người khi họ cần duy trì trạng thái tĩnh lặng hoặc thụ động, đồng thời phải duy trì mức độ tập trung đáng kể trong suốt quá trình. Một ví dụ của tình huống này có thể là giám sát lâu dài các phi công khi máy bay đang được điều khiển bởi hệ thống lái tự động. Trong tất cả các trường hợp này, việc giám sát không can thiệp vào quy trình là điều mong muốn, trong khi các cá nhân liên quan vẫn được yêu cầu phải có sự tỉnh táo và khả năng phản ứng nhanh trong mọi trường hợp có thể xảy ra.

Ở đây chúng ta sẽ hướng đến việc nhận diện các trạng thái tinh thần như tập trung, mất tập trung và trạng thái buồn ngủ thông qua tập dữ liệu EEG (Electroencephalography) và sử dụng các mô hình học máy để giải quyết vấn đề này.

1.2 Bộ dữ liệu EEG

1.2.1 Quy trình thu thập dữ liệu

Trong nghiên cứu này, ta có một bộ dữ liệu gốc EEG thu thập từ 5 người tình nguyện viên tham gia thực nghiệm với việc điều khiển một chuyến tàu mô phỏng trên máy tính bằng chương trình "Microsoft Train Simulator". Mỗi thí nghiệm yêu cầu người tham gia điều khiển tàu trong khoảng từ 35 đến 55 phút trên một tuyến đường chủ yếu không có đặc điểm nổi bật trong chương trình mô phỏng được đề cập.

Các trạng thái tinh thần cần được nghiên cứu là:

- Tập trung nhưng thụ động (focused but passive attention): Trạng thái này được hiểu là trạng thái giám sát chuyến tàu một cách thụ động trong khi vẫn duy trì được sự tỉnh táo và tập trung.
- Giám sát tách rời (disengaged supervision): Trạng thái này được hiểu là người tham gia không rơi vào trạng thái buồn ngủ nhưng họ không còn tập trung đến màn hình nữa.
- Buồn ngủ (explicit drowsing): Trạng thái này có thể hiểu là người tham gia có sự tập trung kém và không còn khả năng phản ứng nhanh với các biến đổi từ màn hình.

1.2.2 Một số thông tin về bộ dữ liệu EEG

Bộ dữ liệu này gồm có 34 thí nghiệm để giám sát trạng thái chú ý của người tham gia thí nghiệm. Mỗi tập tin .mat là một đối tượng chứa dữ liệu thu thập được từ thiết bị EMOTIV trong một thí nghiệm. Tất cả 34 thí nghiệm đều được thực hiện trong khoảng thời gian từ 7 giờ tối mỗi ngày để tạo trạng thái buồn ngủ phục vụ cho thí nghiệm. Các thí nghiệm này đều trải qua sự giám sát của người điều phối thí nghiệm đảm bảo người thực hiện thí nghiệm trong trạng thái mạnh khỏe và trong quá trình thí nghiệm, người tham gia không được nói chuyện hoặc di chuyển.

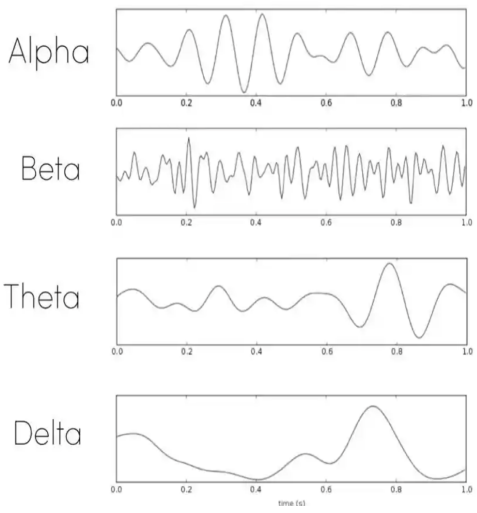
Mỗi người thực hiện thí nghiệm là một subject và mỗi người thực hiện 7 thí nghiệm trong 7 ngày với mỗi ngày một thí nghiệm. Nghĩa là một subject có 7 file .mat với 2 file đầu tiên dùng để làm quen và 5 file còn lại được dùng để làm dữ liệu huấn luyện mô hình. Như vậy thì các file .mat có số thứ tự là 1, 2, 8, 9, 15, 16, 22, 23, 29, 30 sẽ bị loại bỏ khỏi bộ dữ liệu huấn luyện.

Một file .mat có cấu trúc là một dictionary lưu trữ các thông tin như sau:

- Header: Gồm có các thông tin về phiên bản của MATLAB được sử dụng khi tạo file, nền tảng của hệ điều hành khi tạo file và ngày giờ khi file đó được tạo ra.
- Version: Lưu trữ thông tin của phiên bản định dạng của file.mat.
- Globals: Lưu trữ các biến toàn cục của bài toán.
- o: Một structure array lưu trữ các thông tin của thí nghiệm.

Một file .mat lưu các dữ liệu tín hiệu thu được từ 14 kênh dữ liệu là AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 với tần số lấy mẫu là 128Hz. Khoảng băng thông nằm trong khoảng từ 0.3Hz đến 30Hz. Mỗi kênh tín hiệu lại tồn tại nhiều loại sóng khác nhau như:

- **Delta**: 0.5–4 Hz, thường liên quan đến giấc ngủ sâu và trạng thái nghỉ ngơi.
- **Theta**: 4–8 Hz, liên quan đến trạng thái thư giãn hoặc thiền định.
- **Alpha**: 8–13 Hz, xuất hiện trong trạng thái tỉnh táo nhưng thư giãn.
- **Beta**: 13–30 Hz, liên quan đến sự tập trung và hoạt động nhận thức.
- **Gamma**: 30–50 Hz, gắn liền với xử lý thông tin phức tạp và hoạt động nhận thức cao cấp.



Hình 1: Các dải sóng phổ biến

2 Data Preprocessing

Trước hết vì các files trên có ý nghĩa theo thứ tự là 7 files đầu tiên tương đương với người thứ nhất và tương tự với 4 người còn lại nên ta cần phải sắp xếp lại tên các files và địa chỉ files trên để dễ dàng cho việc xử lý dữ liệu.

Thực hiện: theo như hiển thị các tên files và địa chỉ files ở phần trên ta có sự khác nhau giữa các files với nhau hoặc giữa các địa chỉ files với nhau là ở số thứ tự đằng sau record, trước .mat (VD: eeg_record1.mat) nên để sắp xếp lại loạt tên files theo thứ tự ta cần dùng lệnh với key là số đằng sau record được lấy từ hàm extract_number.

2.1 Data Restructuring

Trong thí nghiệm trên, mỗi record là 1 lần thử nghiệm, 7 lần thử nghiệm là 1 subject, có tất cả 5 subjects tức 35 records và mỗi subjects có 2 lần thử nghiệm đầu tiên là làm quen thí nghiệm.

Nhưng lần thử nghiệm cuối cùng của người cuối cùng không được hoàn thành nên sẽ không có record số 35 nên chúng ta có 34 records.

Chúng ta sẽ không sử dụng các records dùng để làm quen với thí nghiệm.

Mà record số 28 bị thiếu dữ liệu nên ta cũng không sử dụng record này.

⇒ Các record có index sau sẽ không được sử dụng: 1, 2, 8, 9, 15, 16, 22, 23, 28, 29, 30.

Để có thể thuận tiện cho việc xử lý dữ liệu của tất cả các records cùng lúc thì chúng ta sẽ tái cấu trúc dữ liệu của tất cả các records thành 3 dictionaries theo trạng thái tập trung, không tập trung và buồn ngủ như sau:

- Tạo 1 marker (dấu mốc) để có thể lấy mẫu dễ dàng hơn ;

- Ta có tần số lấy mẫu là 128 Hz (128 mẫu/ s) và thời gian chuyển đổi trạng thái giữa các trạng thái trong 1 record là 10 phút và thời gian ngủ là thời gian còn lại.

$$\texttt{marker} = 128 \times 60 \times 110(\text{mẫu})$$

Cuối cùng, ta sẽ có ba `dictionaries` với `keys` là tên của *records*, và trong mỗi `key` là các mảng có dạng n channels $\times$ n samples. Cụ thể:

$$\texttt{dictionary}_i = \{\text{key}_j : \mathbf{X}_j \,|\, \mathbf{X}_j \in \mathbb{R}^{n \times m}\}$$

Trong đó:

- key_j là tên của một *record*.
- $\mathbf{X}_j$ là một ma trận với n kênh (channels) và m mẫu (samples).
- $\mathbb{R}^{n \times m}$ chỉ ra rằng các ma trận có kích thước $n \times m$ (n kênh và m mẫu).

2.2 Band Pass Filter (BPF)

Quá trình thực hiện:

Trước hết ta tính toán phương sai của bộ dữ liệu theo từng kênh ta thấy được giá trị phương sai rất lớn như 8913, 4512, 12692, và 5636 cho thấy rằng các tín hiệu này có sự phân tán mạnh. Điều này có thể chỉ ra rằng tín hiệu chứa nhiều nhiễu hoặc có biến đổi lớn trong dữ liệu.

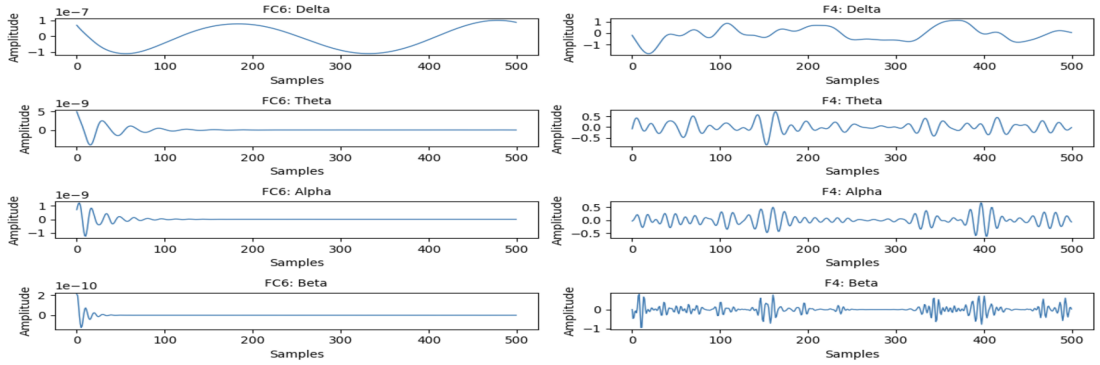
Trong nghiên cứu này, chúng ta lựa chọn phân tích các dải tần từ 0.3 Hz đến 30 Hz, bao gồm các dải **Delta**, **Theta**, **Alpha**, và **Beta**. Việc lựa chọn giới hạn từ 0.3 Hz nhằm giảm thiểu ảnh hưởng của nhiễu tần số thấp, đồng thời tập trung vào các hoạt động điện não có liên quan đến trạng thái tinh thần (*mental state*) của con người.

Những dải tần này được xem là phù hợp để phân tích trạng thái não bộ vì chúng phản ánh các hoạt động chức năng đặc trưng của não trong các tình huống khác nhau, từ nghỉ ngơi đến tập trung cao độ.

$\Rightarrow$ Quá trình lọc này sẽ được thực hiện với ngưỡng highpass = 0.3 Hz và lowpass = 30 Hz.

Sau khi thực hiện quá trình lọc dữ liệu:

- Phương sai tính theo từng kênh đã giảm đáng kể so với trước khi lọc.
- Tuy nhiên, khi vẽ biểu đồ dải tần, chúng ta nhận thấy rằng các loại sóng của kênh FC6 không đúng định dạng kỳ vọng, và có sự xuất hiện của các thành phần không phù hợp.



Hình 2: Hình dạng lạ của các dải sóng trong kênh FC6

Điều này chỉ ra rằng trong tín hiệu vẫn còn sự pha trộn giữa các tín hiệu nhiễu và tín hiệu thần kinh chân thực hoặc tín hiệu nhiễu đã lấn át tín hiệu gốc.

Để giải quyết vấn đề này, chúng ta sẽ tiến hành bước tiếp theo, đó là thực hiện Independent Component Analysis (ICA).

2.3 ICA - Independent Component Analysis

Trong trường hợp này, ICA được sử dụng để lọc các nguồn tín hiệu nhiễu như nháy mắt, tín hiệu cơ (EMG), và nhiễu từ thiết bị có xuất hiện trong tín hiệu EEG.

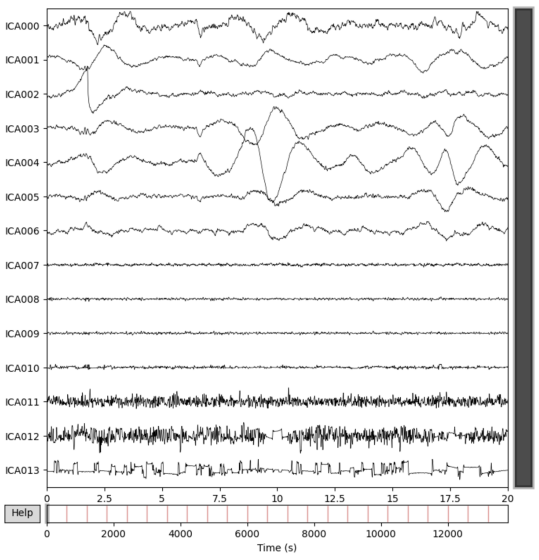
Phương pháp thực hiện:

Để tối ưu hiệu quả, ICA được fit trên toàn bộ dữ liệu thay vì từng trạng thái riêng lẻ. Cách tiếp cận này giúp ICA học tốt hơn các thành phần nhiễu chung giữa các trạng thái, nhờ số lượng mẫu lớn hơn, đảm bảo phân tách thành phần độc lập chính xác và nhất quán. Ma trận trộn và tách được áp dụng đồng nhất trên tất cả các trạng thái, giảm thiểu sự không đồng nhất và tiết kiệm tài nguyên tính toán.

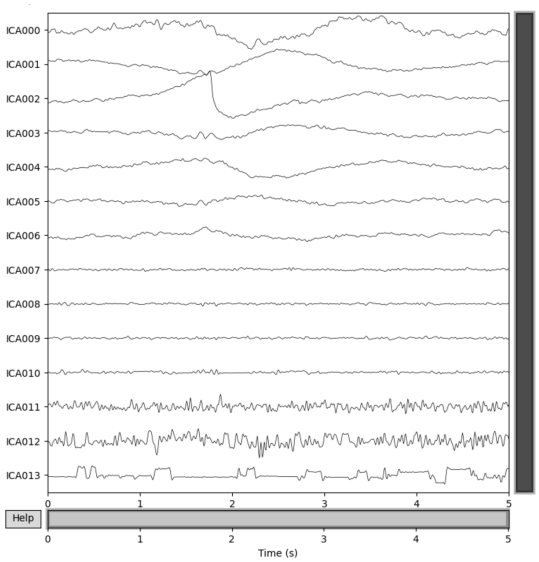
Chọn **14 components** khi áp dụng ICA là bởi vì mục tiêu ban đầu khi áp dụng là để lọc các tín hiệu nhiễu. Và có thể những tín hiệu này đến từ rất nhiều nguồn khác nhau. Nên sử dụng **14 components** là để đảm bảo tránh mất những thông tin quan trọng cũng như là để ICA có thể phân tách được nhiều nguồn nhiễu rõ ràng hơn

Sau khi fit ICA lên toàn bộ dữ liệu, ta sẽ trực quan các biểu đồ: topomap, segment image, ERP/ERF, phổ tần số, variance từ đó đánh giá và lọc ra các thành phần ICA từ nguồn nhiễu.

Dưới đây là các biểu đồ thể hiện dạng tín hiệu của các thành phần ICA theo các khoảng thời gian:

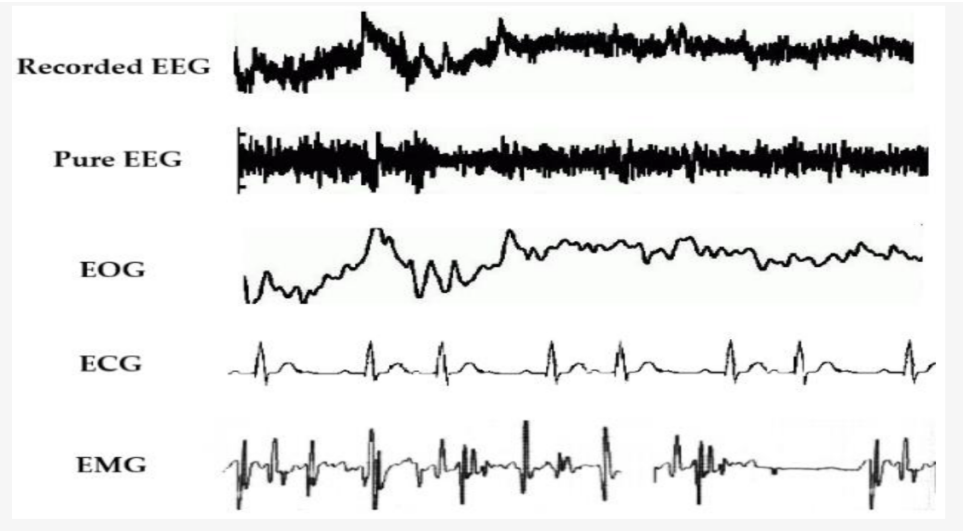


Hình 3: Dạng tín hiệu của các thành phần ICA



Hình 4: Dạng tín hiệu của các thành phần ICA trong khoảng thời gian ngắn (0 đến 5s)

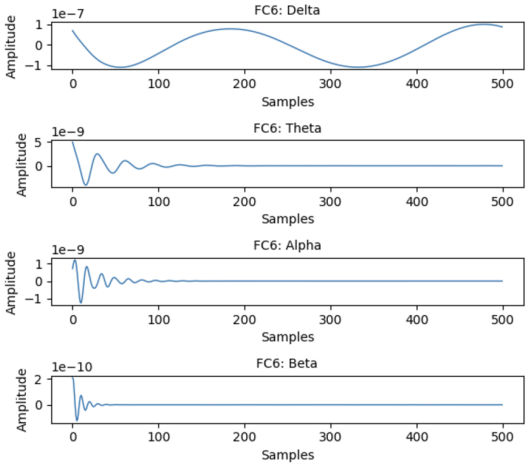
Ta thấy được dạng sóng của **ICA011, ICA012, ICA013** có nhiều gai, hình dạng giống như là nhịp tim, giống với nhiễu ECG (tín hiệu nhiễu điện từ tim). Ta có thể đối chiếu với các dạng sóng thường xuất hiện trong tín hiệu EEG:



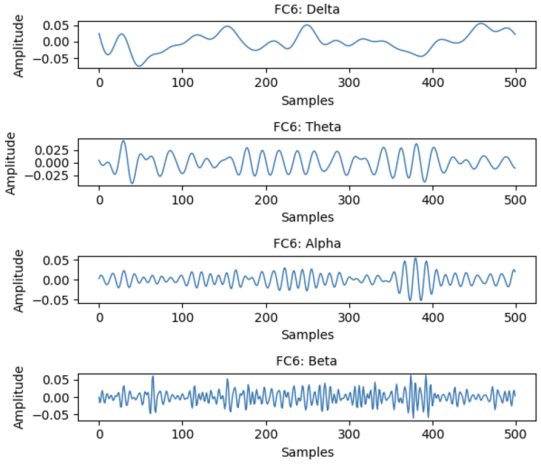
Hình 5: Dạng nhiễu sinh lý trong tín hiệu EEG

Các thành phần nhiễu được chọn để loại bỏ: ICA011, ICA012, ICA013

Kiểm tra các dải sóng sau khi loại bỏ thành phần nhiễu: Ta sẽ kiểm tra trên kênh FC6 của `'eeg_record7'` trong khoảng 4 giây đầu ở trạng thái tập trung vì dạng của các dải sóng trong đây trước khi thực hiện loại bỏ nhiễu rất kỳ lạ, không mang đúng đặc trưng của các dải sóng.



Hình 6: Dạng sóng của các dải sóng trước khi lọc các nguồn nhiễu



Hình 7: Dạng sóng của các dải sóng sau khi lọc các nguồn nhiễu

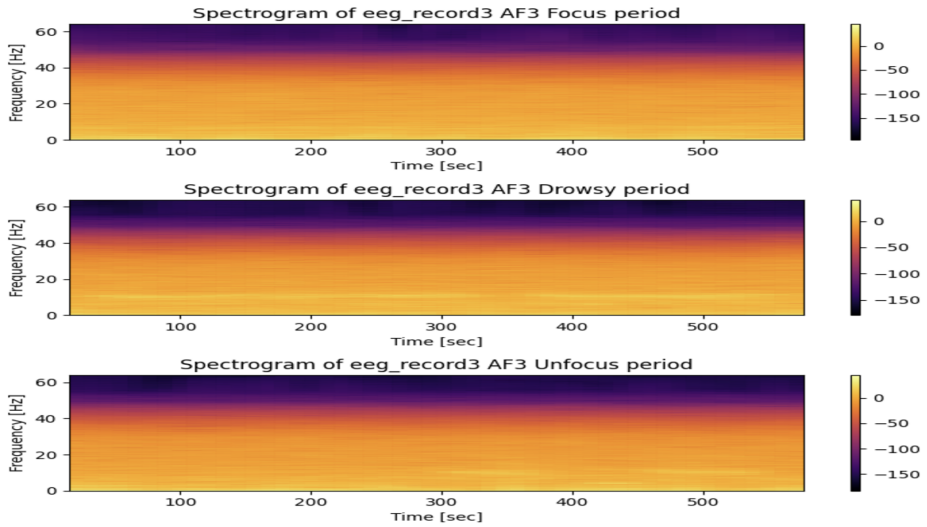
Ta dễ dàng nhận thấy được sự khác nhau của các dải sóng trong kênh FC6 trước và sau khi thực hiện lọc nhiễu bằng ICA. Các dải sóng sau khi lọc nhiễu bằng ICA đã trờ về đúng định dạng nên có của chúng.

3 Feature Engineering

3.1 Xử lý dữ liệu ban đầu

a. **Chia tín hiệu thành các window:** Bộ tín hiệu gốc được chia thành các window, mỗi window là tín hiệu trong một khoảng thời gian cố định. Tín hiệu trong window được tổ chức dưới dạng một ma trận gồm 14 cột (tương ứng với 14 channels).

b. **Biến đổi Fourier:** Fourier transform được sử dụng để chuyển đổi tín hiệu từ **miền thời gian** (time domain) sang **miền tần số** (frequency domain). Từ đó, ta có thể rút ra các đặc điểm về tần số của tín hiệu.



Hình 8: Hình Spectrogram được vẽ từ AF3 của bản ghi eeg_record3 theo từng giai đoạn

Nhận xét: Phân bố năng lượng phổ:
Trong cả ba giai đoạn, năng lượng phổ tập trung chủ yếu ở các dải tần sau:

(a) **Giai đoạn "Focus":**
Năng lượng phổ của các dải tần không có sự biến động đáng kể, cho thấy sự ổn định trong hoạt động não bộ.

(b) **Giai đoạn "Unfocus":**
Ở những thời điểm cuối của giai đoạn này, năng lượng của dải tần alpha có sự biến động nhẹ, cho thấy trạng thái mất tập trung và thay đổi trong hoạt động não bộ.

(c) **Giai đoạn "Drowsy":**
Năng lượng ở tần số xung quanh 10 Hz, thuộc dải tần alpha, có sự gia tăng rõ rệt. Điều này cho thấy trạng thái não bộ chuyển sang thư giãn mạnh mẽ hoặc buồn ngủ.

c. **Trích xuất đặc trưng:** Các đặc trưng được trích xuất từ từng dải tần bao gồm:

- **Power Band:** Năng lượng của dải tần (công suất phổ), biểu thị độ đóng góp của dải tần vào toàn bộ tín hiệu.
- **Mean Power:** Giá trị trung bình của công suất phổ, biểu thị sự phân bố năng lượng trong dải tần.
- **Max Power:** Công suất lớn nhất trong dải tần.
- **Peak Power:** Thành phần tần số có công suất lớn nhất trong dải tần.
- **Std Power:** Độ lệch chuẩn của công suất.
- **CV (Coefficient of Variation):** Hệ số biến thiên, biểu thị mức độ phân tán của năng lượng.
- **Ratio Power Band:** Tỷ lệ giữa công suất dải tần với công suất toàn bộ tín hiệu, chuẩn hóa mức độ đóng góp giữa các dải tần.

d. Đặc trưng tổng hợp: Mỗi window được chuyển thành một điểm dữ liệu gồm $14 \times 4 \times 7 = 392$ đặc trưng (chưa bao gồm state).

3.2 Lựa chọn đặc trưng (Feature Selection)

a. Vấn đề: Với 392 đặc trưng, việc phân tích và xử lý trở nên phức tạp. Do đó, cần lựa chọn các đặc trưng quan trọng nhất để giảm thiểu kích thước dữ liệu mà vẫn đảm bảo hiệu quả phân tích.

b. Phương pháp sử dụng: Sử dụng mô hình **Random Forest (RF)** để đánh giá tầm quan trọng của từng đặc trưng, nhờ thuộc tính `feature_importances_`. Quy trình như sau:

- Chia dữ liệu thành ba tập: train, validation và test.
- Dùng RF để huấn luyện trên tập train và tính toán độ quan trọng (*importance*) của từng đặc trưng.
- Đặt một ngưỡng (*threshold*) để chọn các đặc trưng có tầm quan trọng cao hơn ngưỡng này.
- Đánh giá tập đặc trưng mới bằng mô hình SVM trên tập validation.
- Nếu sau N lần giảm ngưỡng mà kết quả không cải thiện, chọn bộ đặc trưng cho kết quả tốt nhất làm bộ đặc trưng cuối cùng.

c. Lợi ích: Quy trình này đảm bảo rằng các đặc trưng quan trọng nhất được giữ lại, giảm kích thước dữ liệu và tăng hiệu quả của bài toán phân loại.

Kết quả so sánh hiệu suất các mô hình có thể được thấy trong Table 1.

4 Xây dựng mô hình

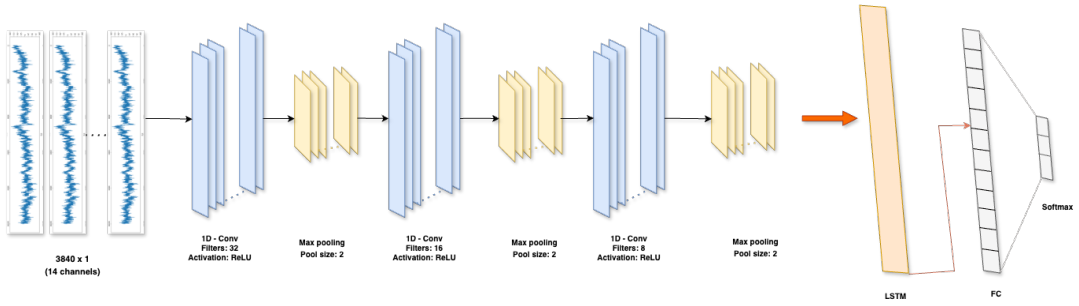
4.1 Tổng quan mô hình

Xây dựng hai hướng tiếp cận chính để phân loại trạng thái tinh thần (tập trung, không tập trung, buồn ngủ) từ dữ liệu EEG, tập trung vào cả miền thời gian và miền tần số.

Miền thời gian (Time Domain): Với dữ liệu ở miền thời gian, chúng ta sử dụng mô hình ConvLSTM 1D, một kiến trúc mạnh mẽ kết hợp giữa mạng tích chập (CNN) và mạng bộ nhớ dài- ngắn hạn (LSTM). CNN giúp trích xuất các đặc trưng không gian từ các kênh EEG, trong khi LSTM đảm nhận việc học các mối quan hệ thời gian trong chuỗi tín hiệu. Mô hình này được thiết kế để tận dụng tối đa tính tuần tự và không gian của dữ liệu EEG, giúp phân loại hiệu quả các trạng thái chú ý.

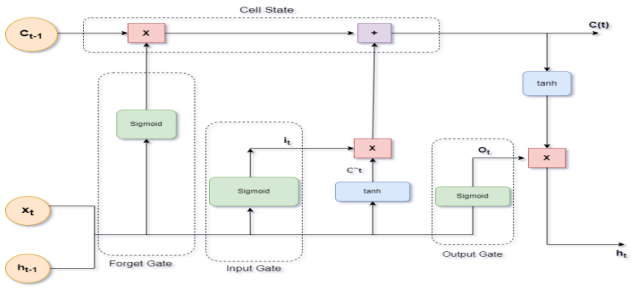
Miền tần số (Frequency Domain): Trong miền tần số, sử dụng các đặc trưng đã được trích xuất sau khi thực hiện Fourier transform, sau đó được dùng làm đầu vào cho các mô hình học máy truyền thống là SVM và Random Forest. Hai mô hình này được lựa chọn vì khả năng xử lý hiệu quả các đặc trưng tần số của tín hiệu EEG và tính linh hoạt trong việc phân loại dựa trên các đặc trưng phức tạp

4.2 ConvLSTM 1D



Hình 9: Kiến trúc mạng ConvLSTM 1D được áp dụng

Mạng CNN 1D được thiết kế để xử lý dữ liệu chuỗi như tín hiệu EEG, hoạt động trên đầu vào có cấu trúc **(số mẫu × chuỗi thời gian × số kênh)** với chuỗi thời gian là tín hiệu 30 giây và 14 kênh cảm biến EEG. Sau CNN, LSTM tiếp tục xử lý tín hiệu tại mỗi thời điểm, khai thác phụ thuộc thời gian giữa các mẫu và kênh EEG. Đầu vào của LSTM là đầu ra từ lớp CNN cuối cùng, có kích thước **128 kênh**, giúp lưu giữ mối quan hệ thời gian qua các trạng thái ẩn.



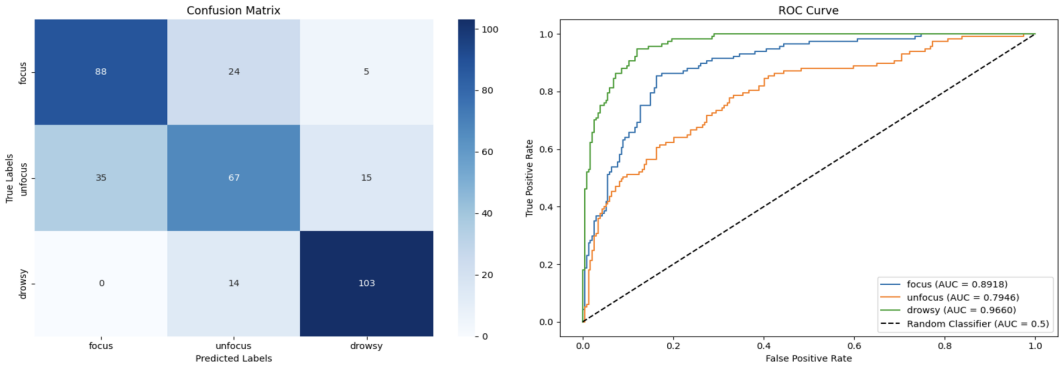
Hình 10: Kiến trúc LSTM (Long Short Term Memory)

Quá trình hoạt động:

Quá trình hoạt động của mô hình gồm các bước chính:

1. **Trích xuất đặc trưng không gian bằng CNN 1D**: Tín hiệu EEG trải qua các lớp CNN 1D để nhận diện các đặc trưng không gian quan trọng.
2. **Giảm kích thước và tăng tính trừu tượng bằng MaxPooling**: Các đặc trưng được giảm kích thước và tăng tính trừu tượng thông qua lớp MaxPooling.
3. **Xử lý phụ thuộc thời gian bằng LSTM**: LSTM học các mối quan hệ thời gian trong tín hiệu EEG.
4. **Sử dụng trạng thái ẩn cuối cùng**: Chỉ sử dụng trạng thái ẩn cuối cùng của LSTM để phân loại qua các lớp fully connected.

Đầu ra cuối cùng của mô hình: Đầu ra cuối cùng là một vector có kích thước bằng số lớp phân loại (`num_classes`). Sau khi thông qua các lớp fully connected (với dropout), kết quả được đưa qua hàm `log_softmax` để chuẩn bị cho việc tính toán hàm loss cross-entropy khi huấn luyện mô hình.



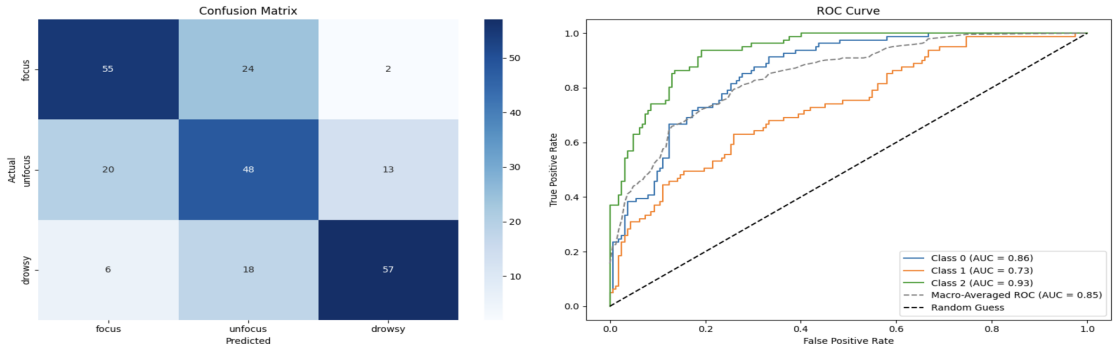
Hình 11: Confusion matrix and ROC of ConvLSTM

```
Classification Report:
              precision    recall  f1-score   support

       focus       0.72      0.75      0.73       117
     unfocus       0.64      0.57      0.60       117
     drowsy        0.84      0.88      0.86       117

    accuracy                           0.74       351
   macro avg       0.73      0.74      0.73       351
weighted avg       0.73      0.74      0.73       351
```

Hình 12: Classification Report of ConvLSTM

4.3 SVM

Dữ liệu được chuẩn hóa để cải thiện hiệu suất SVM, sử dụng kernel RBF, C=10 và gamma='scale' để tối ưu hóa. probability=True hỗ trợ đánh giá qua ROC và AUC. Chuẩn hóa giúp SVM cân bằng đặc trưng và tăng tốc hội tụ.



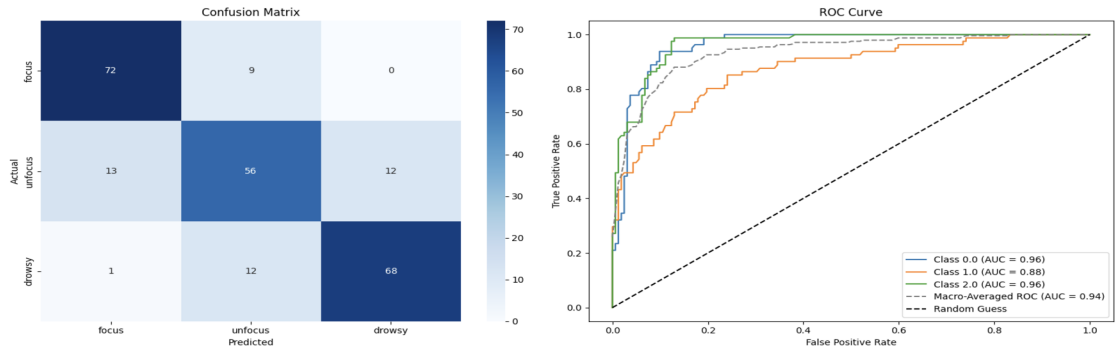
Hình 13: Confusion matrix and ROC of SVM

	precision	recall	f1-score	support
0.0	0.70	0.82	0.75	120
1.0	0.64	0.62	0.63	120
2.0	0.88	0.75	0.81	120
accuracy			0.73	360
macro avg	0.74	0.73	0.73	360
weighted avg	0.74	0.73	0.73	360

Hình 14: Classification Report of SVM

4.4 Random Forest

Dữ liệu được chuẩn hóa để phù hợp với SVM mặc dù việc này không ảnh hưởng đến Random Forest. Random Forest sử dụng ngưỡng chia dựa trên giá trị tương đối, nên không phụ thuộc vào đơn vị hay phạm vi đặc trưng.



Hình 15: Confusion matrix and ROC of RF

Classification Report:

	precision	recall	f1-score	support
0.0	0.84	0.89	0.86	81
1.0	0.73	0.69	0.71	81
2.0	0.85	0.84	0.84	81
accuracy			0.81	243
macro avg	0.80	0.81	0.81	243
weighted avg	0.80	0.81	0.81	243

Hình 16: Classification Report of RF

5 Kết luận

- **SVM (Support Vector Machine):** Hiệu quả được cải thiện rõ rệt khi áp dụng feature selection, giúp chọn bộ đặc trưng tối ưu, giảm overfitting và tăng độ chính xác.
- **RF (Random Forest):** Với khả năng chọn lọc đặc trưng tốt, RF mang lại hiệu suất đáng tin cậy và dễ triển khai. Mô hình này đã được sử dụng để lựa chọn các đặc trưng quan trọng, hỗ trợ cải thiện hiệu quả của các mô hình khác.
- **ConvLSTM:** ConvLSTM vượt trội hơn so với các mô hình truyền thống khi tự học được các đặc trưng của dữ liệu thô trên miền thời gian mà không cần trải qua các quá trình tiền xử lý dữ liệu phức tạp

So sánh kết quả

Hiệu suất của các mô hình được đánh giá dựa trên accuracy, precision,recall và F1. Kết quả cụ thể như sau:

Mô hình	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
RF	**0.81**	**0.81**	**0.81**	**0.81**
ConvLSTM	0.74	0.73	0.74	0.73
SVM (feature selected)	0.73	0.74	0.73	0.73
SVM	0.66	0.67	0.66	0.66

Bảng 1: So sánh hiệu suất các mô hình trên bộ dữ liệu EEG.

6 Thảo luận

Những thách thức khi tìm hiểu và tiếp cận bộ dữ liệu:

1. Để xử lý và phân tích dữ liệu EEG hiệu quả, cần có kiến thức chuyên môn sâu rộng trong các lĩnh vực như tín hiệu học, xử lý tín hiệu số, và các kỹ thuật học máy. Việc hiểu rõ các đặc tính của tín hiệu EEG, cách các sóng não khác nhau ảnh hưởng

đến kết quả, cùng với khả năng lựa chọn và xử lý các đặc trưng phù hợp, đòi hỏi sự am hiểu chuyên sâu

2. Một trong những thách thức lớn nhất trong xử lý dữ liệu EEG là việc đánh giá chất lượng dữ liệu. Dữ liệu có thể bị thiếu hoặc cần phải xác định và loại bỏ các đoạn dữ liệu không hợp lệ.

3. Dữ liệu EEG thường chứa nhiều nhiễu, bao gồm các tín hiệu không liên quan đến não bộ như nhiễu từ thiết bị (motion artifacts, electrical noise, etc.) và nhiễu từ các yếu tố bên ngoài như cử động cơ thể, tiếng ồn môi trường, hoặc các tín hiệu không mong muốn từ các nguồn khác.

4. Dữ liệu EEG là tín hiệu thời gian, thường có tính không gian và thời gian cao. Các tín hiệu này cần phải được phân tích và chuyển đổi để dễ dàng xử lý và phân loại. Điều này đòi hỏi sự kết hợp của nhiều kỹ thuật khác nhau.

5. Dữ liệu EEG có rất nhiều đặc trưng, và việc lựa chọn đặc trưng phù hợp cho các mô hình học máy là một thách thức lớn.

Những thách thức khi xây dựng mô hình:

1. Đối với SVM: Mặc dù SVM đạt hiệu quả tốt với bộ dữ liệu chuẩn hóa, nhưng mô hình vẫn gặp phải vấn đề khi đối diện với dữ liệu có kích thước lớn hoặc có nhiều biến thể, khiến cho việc tối ưu hóa trở nên khó khăn hơn.

2. Đối với RF: Mặc dù RF cho kết quả cao nhất, nhưng việc lựa chọn các đặc trưng quan trọng từ nhiều tính năng có thể gặp khó khăn, đặc biệt khi số lượng đặc trưng quá lớn. Điều này có thể dẫn đến việc mô hình bị quá tải với dữ liệu không quan trọng.

3. Đối với CNN: Một trong những thách thức lớn khi sử dụng CNN là yêu cầu về phần cứng cao, đặc biệt là trong quá trình huấn luyện với dữ liệu lớn. Hơn nữa, việc điều chỉnh các siêu tham số cho CNN cũng đòi hỏi kinh nghiệm và thời gian thử nghiệm lâu dài.