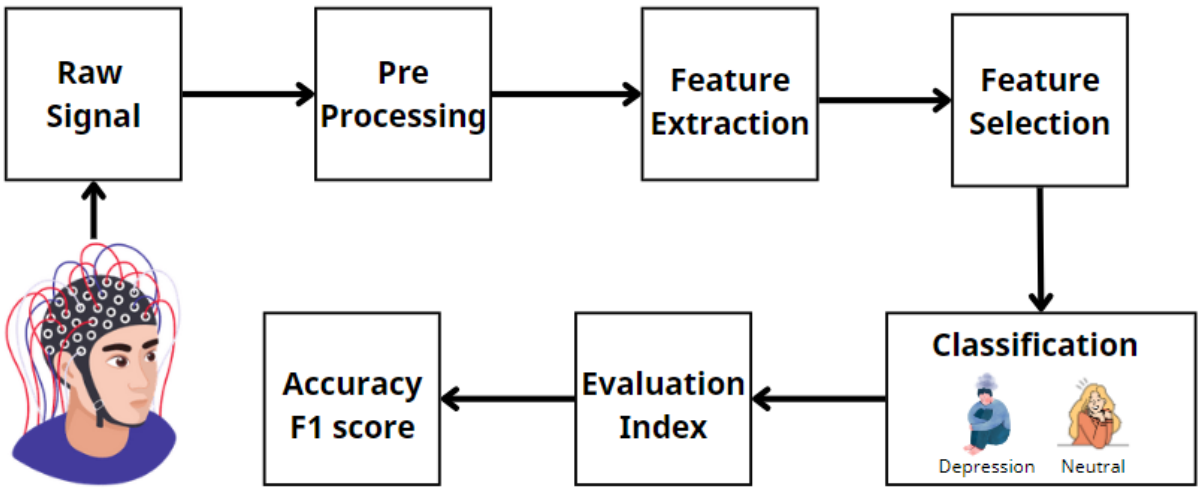
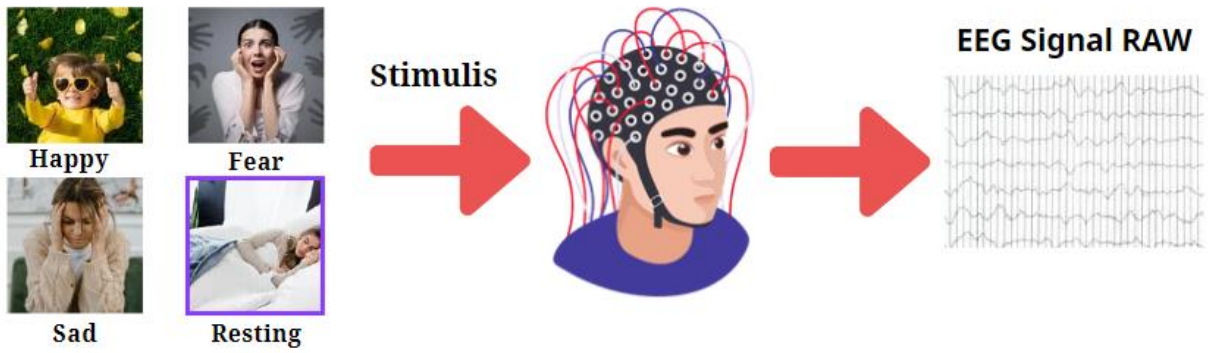
ELECTROENCEPHALOGRAPHY-BASED DEPRESSION DETECTION



Flow Chart prediction depression using EEG signal

1. Dataset:

Gồm có 2 dữ liệu dạng nghỉ và dạng hoạt động. Đối với dạng hoạt động thì gồm có 128 kênh EEG đầu vào và 21 kích thích **Stimulis** đại diện bởi 3 kích thích cảm xúc đầu vào Fear, Happy và Sad thông qua phim ảnh. (53 bệnh nhân)



Stage1: Data Collection (Data Set)

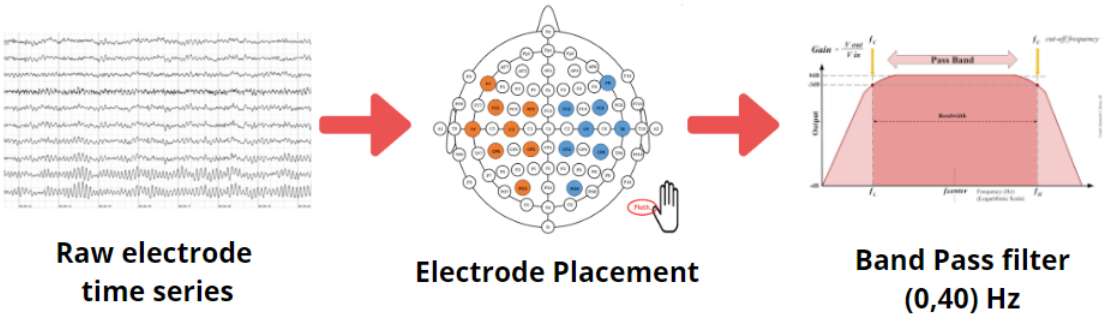
2. PreProcessing

Vì dữ liệu EEG đầu vào là gồm 128 kênh, và để đơn giản bài toán, ta sẽ rút gọn đi còn 16 kênh chủ đạo cho việc phân tích kết quả.

C3:E36	Left central	O1:E70	Back-left to center
C4:E104	Right central	O2:E83	Back-right to center
F3:E24	Front left	P3:E52	Back above
F4:E124	Front right	P4:E92	Back above
F7:E33	Front far left	T3-T7:E45	Side left
F8:E122	Front far right	T4-T8:E108	Side right
FP1:E22	Above left eye	T5-P7:E58	Left
FP2:E9	Above right eye	T6-P8:E96	Right

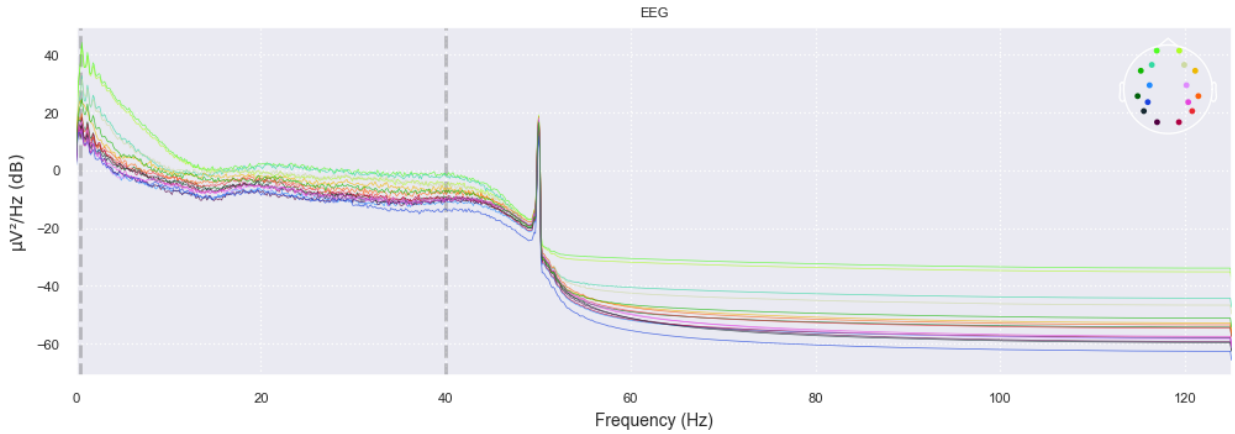
Bảng 16 điện cực chính

Sau khi nhận được tín hiệu thô đầu vào, ta sẽ đi xét 16 channels chính mà ta đã chọn ở bảng 1. Sau đó dữ liệu sẽ được đưa qua bộ lọc để tiếp tục quá trình xử lý.



Stage2: PreProcessing

Vì các thành phần hoạt động chính của tín hiệu EEG được xem xét từ tần số 0-40Hz, nên ta sẽ dùng lọc thông dải để loại bỏ các yếu tố phụ và chỉ giữ lại thành phần chính cần xét tới. Kết quả được thể hiện ở hình dưới:

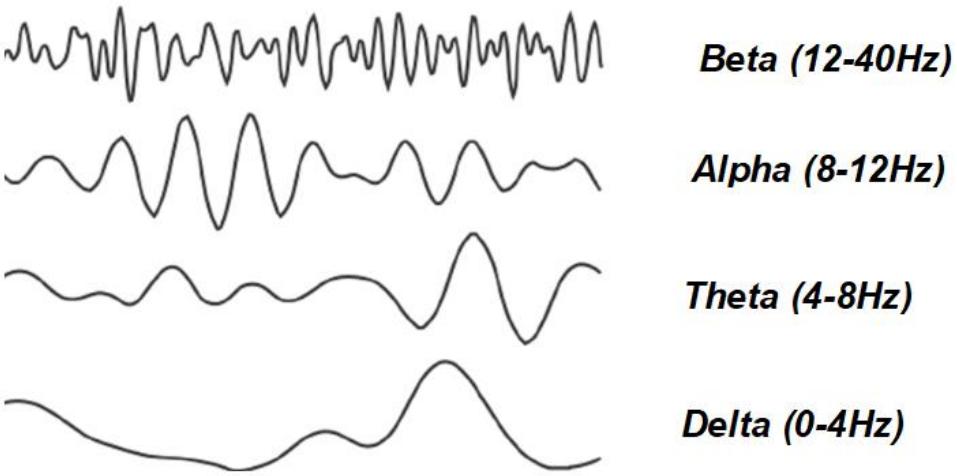


Áp dụng Band Pass filter lên dữ liệu

3. Feature Extraction

Sóng EEG được đặc trưng bởi 4 loại sóng cơ bản: Alpha, Beta, Delta, Theta, với mỗi sóng cơ bản sẽ được xuất hiện ở những tần số nhất định. Những sóng này sẽ cung cấp cho chúng ta thông tin về tình trạng sức khỏe của người bệnh thông qua những đặc điểm

được thể hiện ở từng loại sóng. Do đó, ta sẽ xem xét tất cả những đặc trưng mà nó thể hiện.



4 sóng cơ bản của EEG signal

Đặc trưng được chia ra làm 2 loại:

Linear Feature: là các đặc điểm tuyến tính bao gồm công suất ở 4 sóng cơ bản. Biên độ của tín hiệu công suất bao gồm giá trị trung bình, trung vị, cực đại và cực tiểu được sử dụng làm đặc điểm tuyến tính.

Với đặc điểm tuyến tính này, ta sẽ trích xuất **max, min, mean, median công suất** của cả đoạn (từ 0 – 40Hz) và giá trị trung bình biên độ của 4 sóng cơ bản.

	Alpha	Beta	Delta	Theta	Mean	Max	Min	Median
E9	0.0011332	0.004280	0.179939	0.041949	0.128638	2940.344	0.000024	0.004702
E22	0.017059	0.008127	0.206598	0.052404	0.064368	406.373	0.000026	0.008649
E24	0.014863	0.008314	0.081826	0.027243	0.096565	1676.277	0.000040	0.008894
E33	0.005416	0.003471	0.029618	0.009749	0.064770	1560.617	0.000013	0.003630

Ví dụ minh họa đặc trưng tuyến tính

Non-Linear Feature: Các đặc điểm phi tuyến được sử dụng trong phân tích là entropy phổ và entropy lắng đọng giá trị đơn (Singular – value, Spectral entropy, Permenone entropy…)

	svden	spec	permen
E1	0.472424	0.490312	0.783087
E2	0.460892	0.467293	0.778375
E3	0.574941	0.546001	0.803617
E4	0.576004	0.585663	0.826692

Ví dụ minh họa đặc trưng phi tuyến

Như đã nêu ra ở đầu, có 21 kênh kích thích đầu vào, đặc trưng cho từng cảm xúc, trong đó phân chia thành 3 kích thích cơ bản: **Fear, Happy, Sad**, với mỗi cảm xúc được theo dõi bởi chỉ số thời gian của các mẫu dữ liệu, nên ta có phép phân tích cụ thể các phản ứng của não đối với từng loại kích thích, và nó sẽ thể hiện đặc trưng trên từng sóng cơ bản. Do đó, ta sẽ trích xuất kích thích của **Fear, Happy, Sad** lên từng sóng cơ bản một xem những đặc trưng của nó được thể hiện như thế nào.

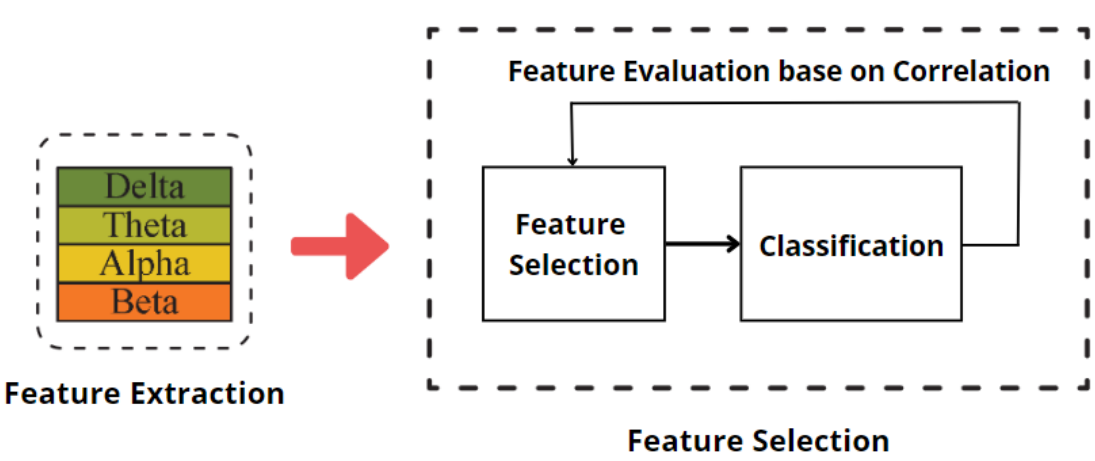
Với mỗi bệnh nhân, có tổng cộng 8 features được lựa chọn cho 16 vị trí điện cực. Tuy nhiên, có các chuỗi hình ảnh kích thích (Fear, Happy, Sad và trạng thái nghỉ ngơi) được tác động tính toán. Do đó không gian feature gồm 8 * 16 * 4 = 512 feature tuyến tính. 8 feature chính gồm Biên độ trung bình - mean (đỉnh đến đỉnh) của tín hiệu công suất, Biên độ trung vị - median (đỉnh đến đỉnh) của tín hiệu công suất, Biên độ tối đa - max (đỉnh đến đỉnh) của tín hiệu công suất, Biên độ tối thiểu - min (đỉnh đến đỉnh) của tín hiệu công suất và Công suất ở mức alpha Beta, Theta, Delta.

Tương tự với Non-linear ta sẽ có 16 channels, 4 kích thích đầu vào và 3 đặc điểm phi tuyến về công suất → **192 non-linear features**

Như vậy, tổng quan lại, 1 bệnh nhân sẽ có 704 feature để so sánh và đưa ra kết luận

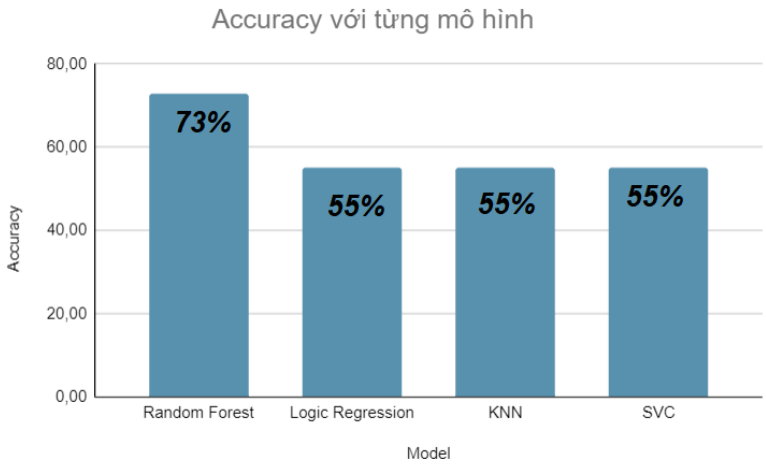
4. Feature Selection & Classification

Tổng quan quá trình Classification sẽ được thực hiện như hình ở dưới đây.



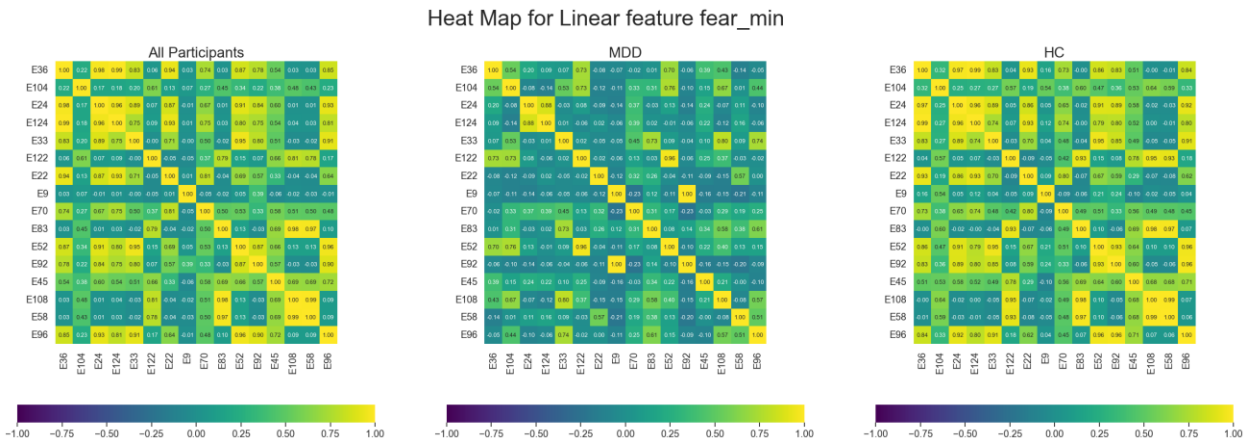
Satge3: Feature Extraction & Feature Slection

Với 704 features đã được tính toán và đưa ra ở bước trên, ta sẽ tiến hành đưa vào huấn huyện model.

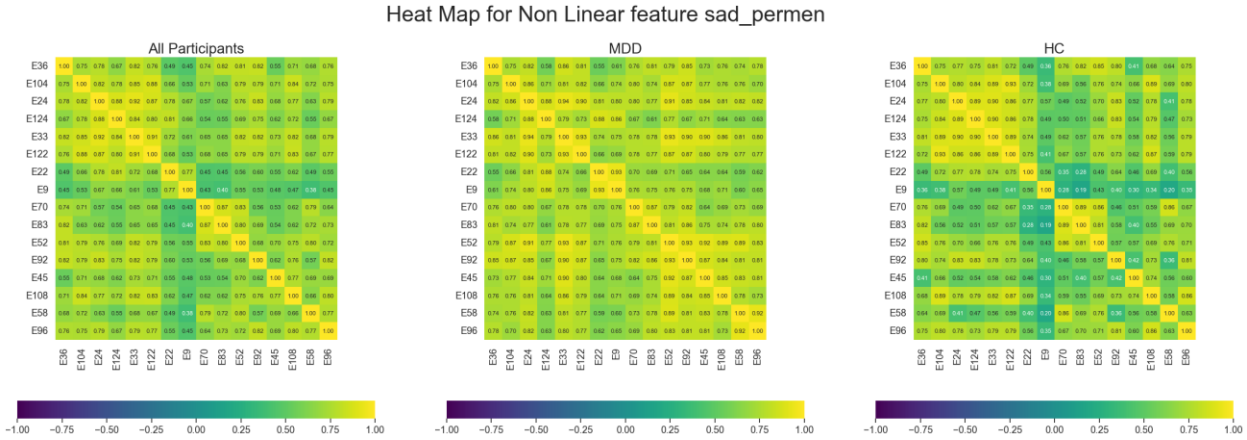


Bảng thể hiện độ chính xác khi chưa lọc feature

Vì hiệu suất đạt được chưa cao, nên ta sẽ phải chọn lọc các feature có độ ảnh hưởng nhiều nhất tới model. Ta sẽ xem xét các mối tương quan giữa các feature với nhau và loại bỏ đi những tương quan >0.9



Ví dụ tương quan 1 loại featrue tuyến tính của 16 kênh

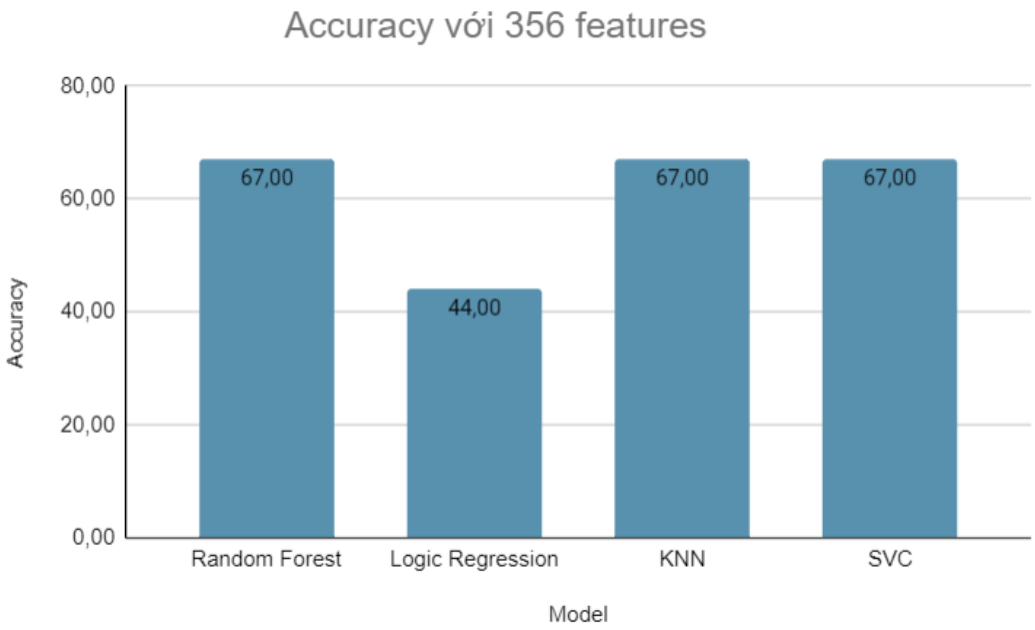


Tương quan 1 loại featrue phi tuyến tính

Mục đích của việc xét tương quan ở đây là: nếu có bất kỳ mối tương quan chặt chẽ nào giữa hai điện cực khác nhau, chúng ta sẽ loại bỏ chúng để nâng cao hiệu suất training model. Các feature sẽ được so sánh mức độ tương đồng với nhau, và nếu feature nào có độ tương đồng với các feature khác nhiều hơn thì sẽ bị loại bỏ.

Nhận xét: Ta có thể nhận thấy rằng, ở các điện cực E70, E52, E92 cho các giá trị tương quan ở mức xấp xỉ 1 (E70 đại diện cho phần cực sau của da đầu, E52, E92 đại diện cho tín hiệu từ phía sau da đầu) và nó thể hiện nhất quán ở hầu hết tất cả các đặc trưng ta xem xét. Các đặc điểm phi tuyến tính có mối tương quan cao khi xét về người bình thường (HC)

- Sau khi loại bỏ các kênh đó, ta giảm còn từ 704 → 356 features. Và kết quả khi training với 356 features đó là:



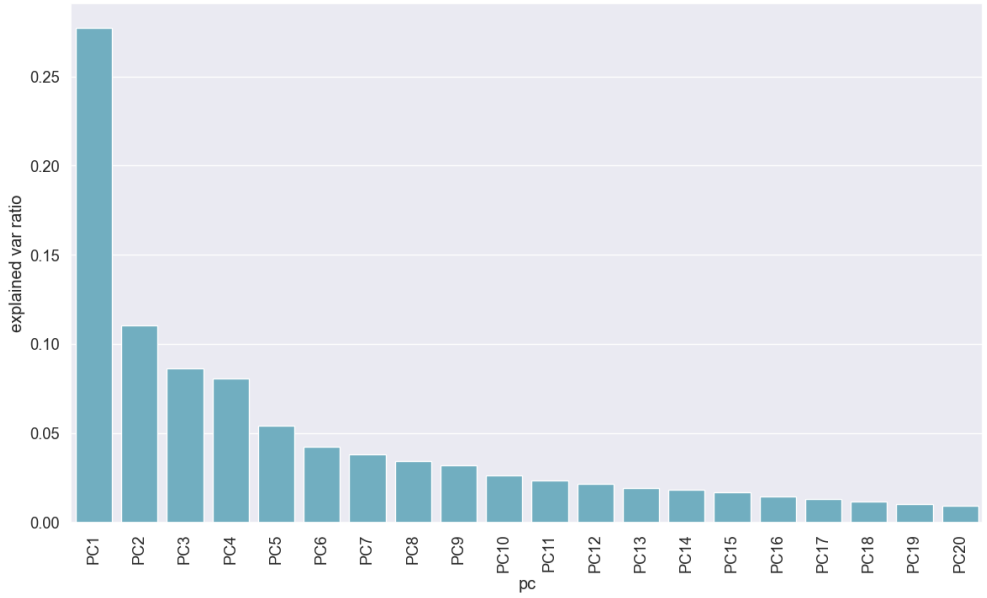
Bảng thể hiện độ chính xác sau khi loại bỏ feature thông qua đánh giá tương quan

PCA

Phân tích thành phần chính (PCA) là một phương pháp giảm chiều dữ liệu phổ biến, được sử dụng để biến đổi dữ liệu đa chiều thành một không gian có ít chiều hơn, trong khi vẫn cố gắng giữ lại phần lớn thông tin (biến thiên) từ dữ liệu gốc.

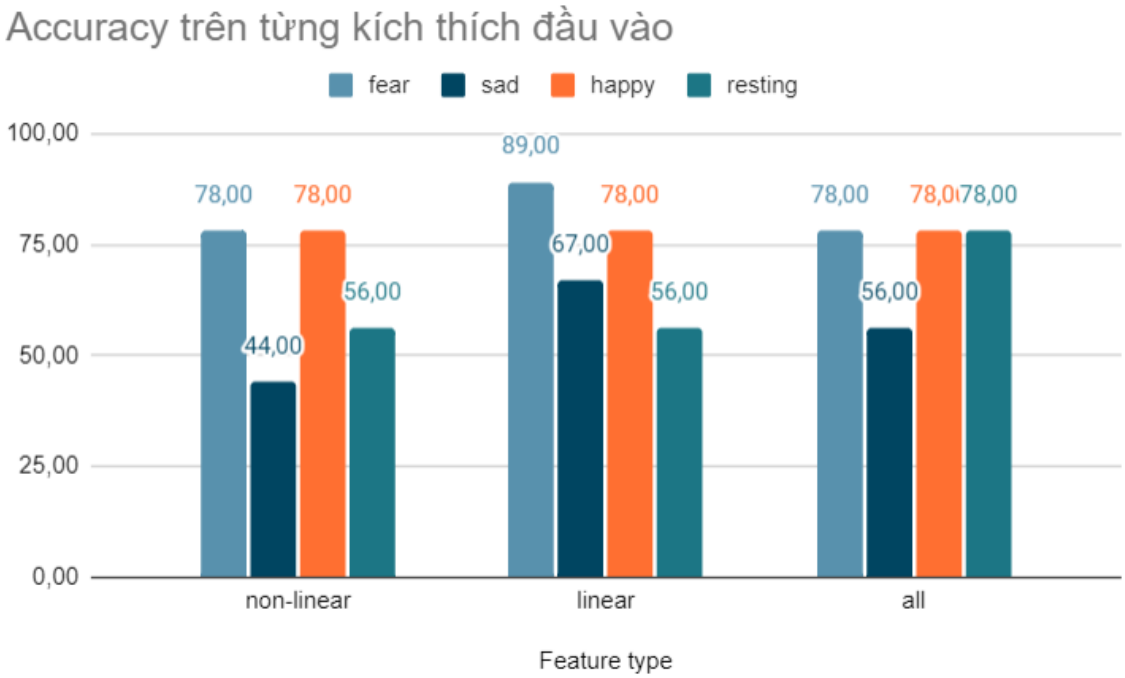
Bước đầu trong việc thử nghiệm và phân tích, ta sẽ áp dụng PCA thành 20 thành phần chính, và nhận thấy rằng thành phần đầu chiếm 27.7% tổng thông tin biến động, sau 20 thành phần thì giữ lại được 93.8%.

Việc lựa chọn 20 thành phần chính là để kiểm chứng xem PCA liệu có thực sự hoạt động tốt trên tập dữ liệu hay không. Ở các bài báo cáo sau, ta sẽ xem xét PCA với số lượng thành phần khác nhau và đưa ra so sánh.



Bảng phân bố tỉ lệ thông tin của từng thành phần chính

Vì mô hình Random Forest có độ chính xác cao nhất nên ta sẽ xét mức độ ảnh hưởng của từng kích thích lên mô hình. Kết quả sau khi phân tích thành 20 thành phần chính với các mô hình khác nhau được thể hiện ở đồ thị dưới:



Ta nhận thấy rằng Fear hoạt động tốt hơn hẳn so với các tính năng khác, cụ thể nếu chỉ xét các tính năng tuyến tính thì Fear giúp ta đạt được 89%. Do vậy, khi ta xét tình trạng bệnh nhân thông qua kích thích đầu vào, thì Fear sẽ là một sự lựa chọn tốt để dễ dàng nhận diện căn bệnh hơn.

5. Kết luận

Bài báo cáo trên thể hiện bước đầu trong việc phân tích dữ liệu EEG từ bộ dataset MODMA có sẵn. Các kết quả được thống kê trong bài báo cáo này chưa thực sự tối ưu và đạt kết quả mong muốn tốt nhất do mới được xét trên 1 phương diện (ví dụ 16 channels thay vì xét 32 channels; dataset MODMA 53 bệnh nhân thay vì các bộ dữ liệu khác). Do đó, ta cần có sự so sánh giữa các phương pháp được thực hiện trong nghiên cứu này để đưa ra được phương pháp đạt kết quả tốt nhất.