ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

THỐNG KÊ NHIỀU CHIỀU

ĐỒ ÁN KẾT THÚC MÔN

ỨNG DỤNG THỐNG KÊ NHIỀU CHIỀU TRONG NHẬN DIỆN CƠN ĐỘNG KINH TỪ EEG

Giảng viên: TS. Nguyễn Thị Mộng Ngọc

Nhóm sinh viên thực hiện:

Họ và tên	MSSV
Phan Bùi Hữu Ân	22110012
Lê Phạm Hồng Hiên	22110059

Thành phố Hồ Chí Minh, 01/07/2025

Mục lục

1 Tổng quan đề tài

1.1 Bối cảnh và Vấn đề nghiên cứu

Điện não đồ (EEG) là công cụ không thể thiếu trong chẩn đoán và theo dõi bệnh động kinh - một rối loạn thần kinh đặc trưng bởi các cơn co giật bất thường do hoạt động điện không ổn định trong não. Việc phát hiện chính xác và kịp thời các cơn động kinh có vai trò cốt lõi trong việc điều trị và nâng cao chất lượng sống cho người bệnh.

Tuy nhiên, quy trình phân tích tín hiệu EEG thủ công hiện nay bộc lộ nhiều hạn chế lớn: tốn nhiều thời gian, phụ thuộc nặng nề vào kinh nghiệm của chuyên gia dẫn đến tính chủ quan, và khó có thể mở rộng để đáp ứng nhu cầu sàng lọc trên quy mô lớn. Những thách thức này đòi hỏi một giải pháp thay thế hiệu quả hơn.

Để giải quyết vấn đề này, nghiên cứu đề xuất xây dựng và đánh giá một hệ thống học máy hoàn toàn tự động, với khả năng học hỏi các đặc trưng ẩn của tín hiệu và phân loại chính xác trạng thái co giật.

1.2 Câu hỏi và Mục tiêu nghiên cứu

Nghiên cứu này được định hướng bởi hai câu hỏi trọng tâm:

1. Liệu các đặc trưng được trích xuất từ miền thời gian, tần số và wavelet,... có đủ khả năng để phân biệt một cách đáng tin cậy giữa trạng thái co giật (seizure) và không co giật (non-seizure) từ các đoạn EEG ngắn hay không?

2. Việc áp dụng các kỹ thuật giảm chiều như Phân tích Thành phần Chính (PCA) ảnh hưởng như thế nào đến sự cân bằng giữa độ chính xác phân loại và hiệu quả tính toán của mô hình?

Dựa trên các câu hỏi đó, mục tiêu cuối cùng của nghiên cứu là xây dựng, so sánh và xác định mô hình phân loại hiệu quả nhất, từ đó cung cấp một phương pháp luận vững chắc cho việc phát hiện cơn động kinh tự động.

2 Dữ liệu và Phương pháp luận

2.1 Mô tả bộ dữ liệu

Nguồn gốc dữ liệu

Nghiên cứu sử dụng bộ dữ liệu công khai **"Epileptic Seizure Recognition"** từ nền tảng Kaggle. Đây là một phiên bản đã qua xử lý của một bộ dữ liệu gốc nổi tiếng từ nghiên cứu của Andrzejak et al. (2001). Dữ liệu gốc được thu thập nhằm phân tích hoạt động điện não trong các trạng thái và vùng não khác nhau.

Cấu trúc dữ liệu gốc

Dữ liệu gốc bao gồm 5 bộ (datasets), ký hiệu từ A đến E, mỗi bộ chứa 100 tệp tín hiệu EEG đơn kênh, tương ứng với 5 nhóm đối tượng khác nhau:

- **Nhóm E (Tương ứng nhãn y=1):** Gồm các tín hiệu EEG được ghi lại **trong khi bệnh nhân đang có cơn co giật** (trạng thái ictal).
- **Nhóm D (Tương ứng nhãn y=2):** Tín hiệu được ghi từ vùng có khối u trong não của bệnh nhân giữa các cơn co giật (trạng thái interictal).
- **Nhóm C (Tương ứng nhãn y=3):** Tín hiệu được ghi từ vùng não khỏe mạnh ở bán cầu đối diện với vùng có khối u của bệnh nhân, cũng trong trạng thái interictal.
- **Nhóm B (Tương ứng nhãn y=4):** Tín hiệu từ những người tình nguyện khỏe mạnh khi đang **nhắm mắt**.
- **Nhóm A (Tương ứng nhãn y=5):** Tín hiệu từ những người tình nguyện khỏe mạnh khi đang **mở mắt**.

Mỗi tệp tín hiệu gốc có độ dài 4097 điểm dữ liệu, được ghi trong khoảng 23.5 giây.

Quá trình tạo ra bộ dữ liệu trên Kaggle

Bộ dữ liệu Epileptic Seizure Recognition trên Kaggle .

Để tạo ra bộ dữ liệu .csv được sử dụng trong nghiên cứu này, các nhà khoa học đã thực hiện các bước sau:

- Mỗi tín hiệu gốc (dài 4097 điểm) được chia thành 23 đoạn nhỏ không chồng chéo lên nhau.
- Mỗi đoạn nhỏ này có độ dài 178 điểm dữ liệu, tương ứng với 1 giây ghi nhận tín hiệu.
- Quá trình này tạo ra tổng cộng 11,500 mẫu (500 tín hiệu gốc $\times$ 23 đoạn/tín hiệu). Mỗi mẫu này trở thành một hàng trong tệp .csv.

Như vậy, bộ dữ liệu cuối cùng có cấu trúc:

- Cấu trúc: 11,500 hàng (mẫu), mỗi hàng là một đoạn tín hiệu EEG dài 1 giây.
- Đặc trưng: Mỗi hàng có 178 cột đặc trưng (từ X1 đến X178), đại diện cho các giá trị điện thế của tín hiệu.
- Nhãn: Mỗi hàng được gán một nhãn y từ 1 đến 5, tương ứng với 5 nhóm đối tượng ban đầu đã mô tả ở trên.

2.2 Xác lập bài toán phân loại nhị phân

Để tập trung vào mục tiêu chính là phát hiện cơn động kinh, bài toán đa lớp ban đầu được chuyển đổi thành một bài toán phân loại nhị phân. Các nhãn được xử lý như sau:

- **Lớp 1 (Seizure):** Gồm các đoạn tín hiệu được ghi lại trong lúc xảy ra cơn động kinh (nhãn gốc y = 1).
- **Lớp 0 (Non-Seizure):** Gồm tất cả các đoạn tín hiệu còn lại, bao gồm trạng thái mắt mở, mắt nhắm, và hoạt động não bình thường (nhãn gốc y từ 2 đến 5).

2.3 Quy trình nghiên cứu tổng quan

Nghiên cứu được triển khai theo một quy trình gồm 4 giai đoạn chính:

1. **Chuẩn bị dữ liệu:** Chuyển đổi bài toán sang dạng nhị phân (co giật/không co giật).
2. **Trích xuất đặc trưng (Feature Engineering):** Từ 178 điểm dữ liệu thô của mỗi mẫu, tiến hành trích xuất một bộ đặc trưng mới có ý nghĩa hơn, bao gồm các đặc trưng thống kê từ miền thời gian, năng lượng các dải tần số (Delta, Theta, Alpha, Beta, Gamma), và năng lượng các mức phân rã wavelet.
3. **Phân tích Khám phá và Giảm chiều:** Sử dụng các kỹ thuật thống kê đa biến (PCA, FA, MANOVA) và các phương pháp trực quan hóa phi tuyến (t-SNE, UMAP) để khám phá cấu trúc dữ liệu, kiểm tra khả năng phân tách của các lớp, và đánh giá hiệu quả của việc giảm chiều.
4. **Xây dựng và Đánh giá Mô hình:** Áp dụng kỹ thuật SMOTE trên tập huấn luyện để xử lý mất cân bằng lớp. Sau đó, huấn luyện và so sánh hiệu năng của các thuật toán (LDA, SVM, Random Forest) thông qua kiểm định chéo 10 lần. Cuối cùng, đánh giá mô hình tốt nhất trên tập kiểm tra để đưa ra kết luận cuối cùng.

Tải dữ liệu

```
df <- readr::read_csv("Epileptic␣Seizure␣Recognition.csv")
```

Bỏ cột định danh đầu tiên.

```
df <- df %>% dplyr::select(-1)
```

Chuyển đổi biến mục tiêu: Cột mục tiêu là 'y'

```
df <- df %>%
  mutate(y = ifelse(y == 1, 1, 0))
```

Hiện thị 6 dòng đầu

```
## # A tibble: 6 × 179
##      X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12   X13
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   135   190   229   223   192   125    55    -9   -33   -38   -10    35    64
## 2   386   382   356   331   320   315   307   272   244   232   237   258   212
## 3   -32   -39   -47   -37   -32   -36   -57   -73   -85   -94   -99   -94   -96
```

```
## 4  -105  -101   -96   -92   -89   -95  -102  -100   -87   -79   -72   -68   -74
## 5    -9   -65   -98  -102   -78   -48   -16     0   -21   -59   -90  -103   -84
## 6    55    28    18    16    16    19    25    40    52    66    81    98   111
## # i 166 more variables: X14 <dbl>, X15 <dbl>, X16 <dbl>, X17 <dbl>, X18 <dbl>,
## #   X19 <dbl>, X20 <dbl>, X21 <dbl>, X22 <dbl>, X23 <dbl>, X24 <dbl>,
## #   X25 <dbl>, X26 <dbl>, X27 <dbl>, X28 <dbl>, X29 <dbl>, X30 <dbl>,
## #   X31 <dbl>, X32 <dbl>, X33 <dbl>, X34 <dbl>, X35 <dbl>, X36 <dbl>,
## #   X37 <dbl>, X38 <dbl>, X39 <dbl>, X40 <dbl>, X41 <dbl>, X42 <dbl>,
## #   X43 <dbl>, X44 <dbl>, X45 <dbl>, X46 <dbl>, X47 <dbl>, X48 <dbl>,
## #   X49 <dbl>, X50 <dbl>, X51 <dbl>, X52 <dbl>, X53 <dbl>, X54 <dbl>, ...
```

Tách tập đặc trưng (X) và biến mục tiêu (y)

```
X_original <- df %>% dplyr::select(-y)
y_labels <- df$y
```

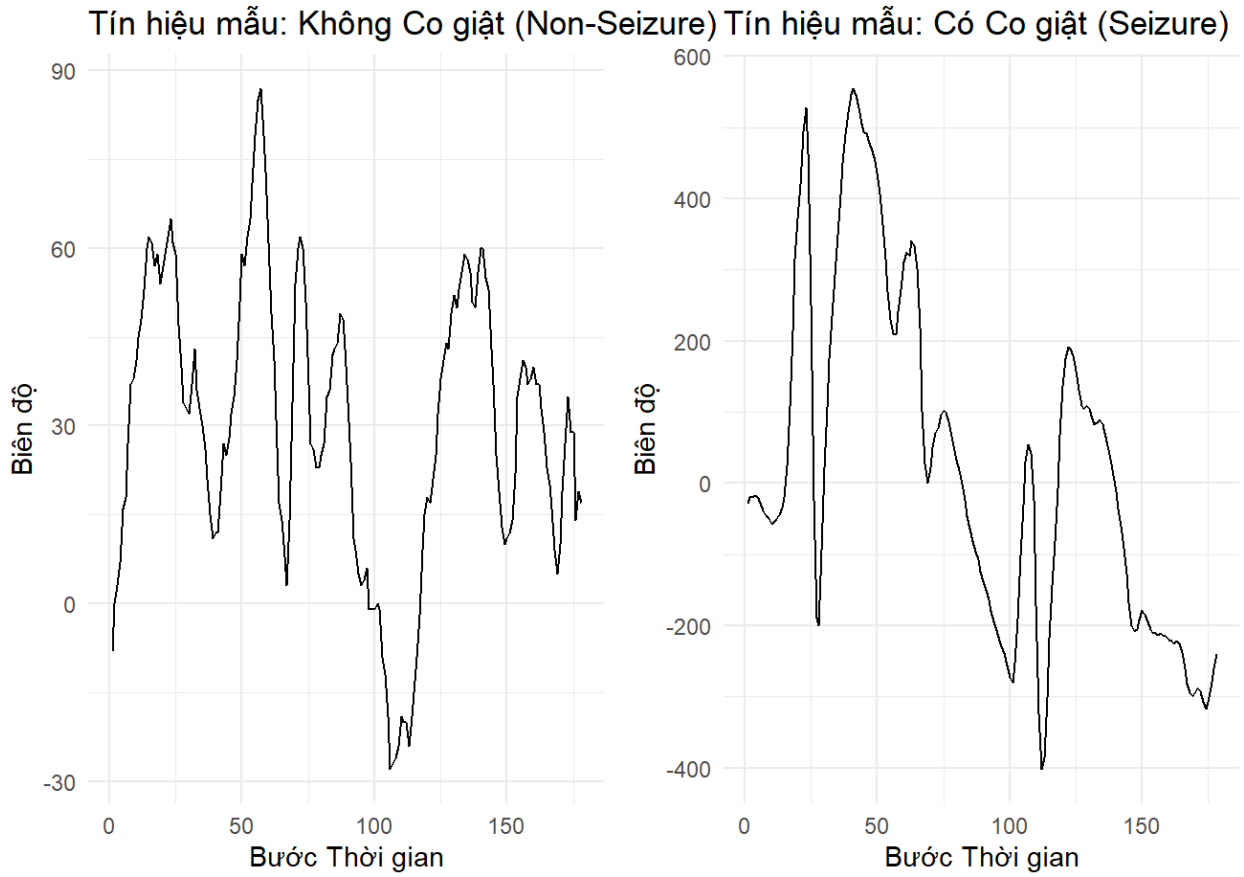
Lấy 1 mẫu KHÔNG co giật (y=0)

```
non_seizure_sample <- df %>% filter(y == 0) %>% sample_n(1) %>% dplyr
   ::select(-y) %>% as.numeric()
```

Lấy 1 mẫu CÓ co giật (y=1)

```
seizure_sample <- df %>% filter(y == 1) %>% sample_n(1) %>% dplyr::
   select(-y) %>% as.numeric()
```

Vẽ 2 biểu đồ

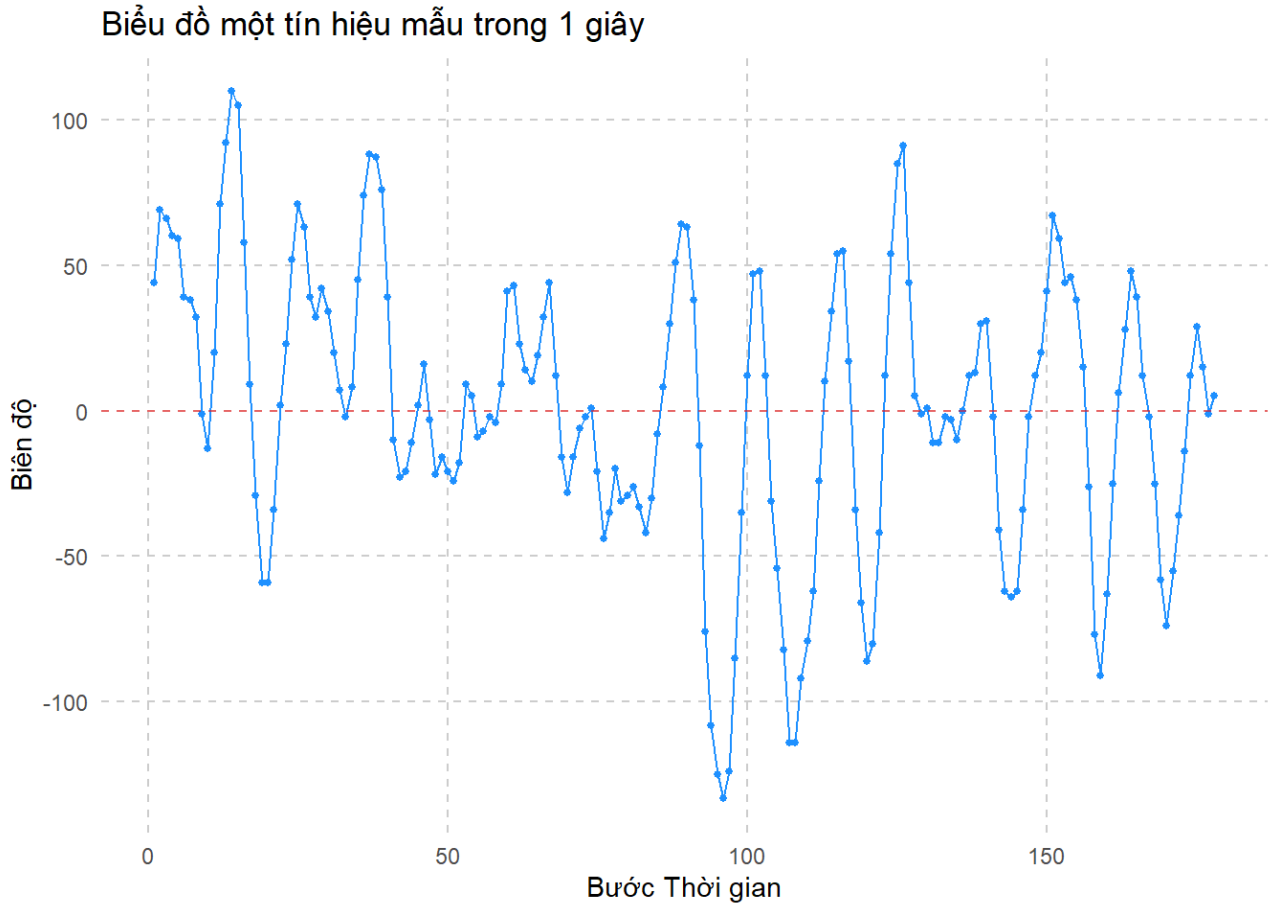


Hình 1: So sánh 2 biểu đồ

3 Trích xuất Đặc trưng (Feature Engineering)

3.1 Đặc trưng Miền Thời gian (Time Domain features)

Trước khi trích xuất ta vẽ một tín hiệu ngẫu nhiên theo miền thời gian



Hình 2: Biểu đồ một tín hiệu mẫu trong 1 giây

Biểu đồ này được gọi là tín hiệu miền thời gian và cho thấy biên độ của tín hiệu thay đổi theo thời gian như thế nào. Từ biểu đồ này, chúng ta sẽ trích xuất một số thống kê tiêu chuẩn giúp mô tả đặc tính của tín hiệu. Chúng ta sẽ giải thích những cái ít phổ biến hơn ở dưới đây.

Crest factor được định nghĩa là tỷ lệ giữa giá trị tuyệt đối lớn nhất của tín hiệu và giá trị RMS của nó:

$$C(y_n) = \frac{\max |y_n|}{RMS(y_n)} \qquad \text{với } RMS(y_n) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} y_n^2},$$

Điều này thể hiện đỉnh cao nhất lớn như thế nào so với công suất trung bình của tín hiệu. Nó cung cấp một thước đo về mức độ cực đoan của các đỉnh trong tín hiệu so với năng lượng tổng thể của nó.

Margin factor có khái niệm tương tự nhưng so sánh giá trị tuyệt đối lớn nhất với phương

sai của tín hiệu:

$$M = \frac{\max |y_n|}{\text{Var}(y_n)}, \qquad \text{với Var}(y_n) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2$$

Điều này phản ánh kích thước của đỉnh cao nhất so với độ phân tán (phương sai) của tín hiệu. Người đọc có thể nhận thấy rằng khi giá trị trung bình của tín hiệu bằng không, ví dụ như trong nhiễu trắng, crest factor và margin factor trùng nhau vì phương sai chính là RMS lấy trung tâm tại giá trị trung bình, tức là 0 trong trường hợp này.

Shape factor đo lường cách giá trị RMS so sánh với Giá trị Tuyệt đối Trung bình (Mean Absolute Value - MAV). Chỉ số này, thay vì định lượng đỉnh cực đại như các chỉ số trước, nó mô tả số lượng các đỉnh và cường độ của chúng. Lưu ý rằng các đỉnh càng cao, sự khác biệt giữa giá trị bình phương và giá trị tuyệt đối của nó càng lớn, do đó phân số tăng lên.

$$S(y_n) = \frac{\text{RMS}}{\text{MAV}}, \quad \text{MAV} = \frac{1}{N}\sum_{n=1}^{N}|y_n|$$

Impulse factor là một thước đo liên quan khác, so sánh giá trị tuyệt đối lớn nhất với MAV. Tương tự, impulse factor làm nổi bật sự thống trị của đỉnh cao nhất của tín hiệu so với độ lớn điển hình của nó.

$$I = \frac{\max |y_n|}{\text{MAV}}$$

3.2 Đặc trưng Miền Tần số (Frequency Domain features)

Mục đích là chuyển đổi từ miền thời gian để phân tích sự phân bố của các tần số, được gọi là các dải công suất. Việc chuyển đổi tín hiệu EEG từ miền thời gian sang miền tần số cho phép chúng ta **tiết lộ các cấu trúc tuần hoàn ẩn, phân biệt các trạng thái nhận thức khác nhau, và phát hiện hoạt động thần kinh bất thường**, chẳng hạn như những gì thấy trong bệnh động kinh hoặc rối loạn giấc ngủ.

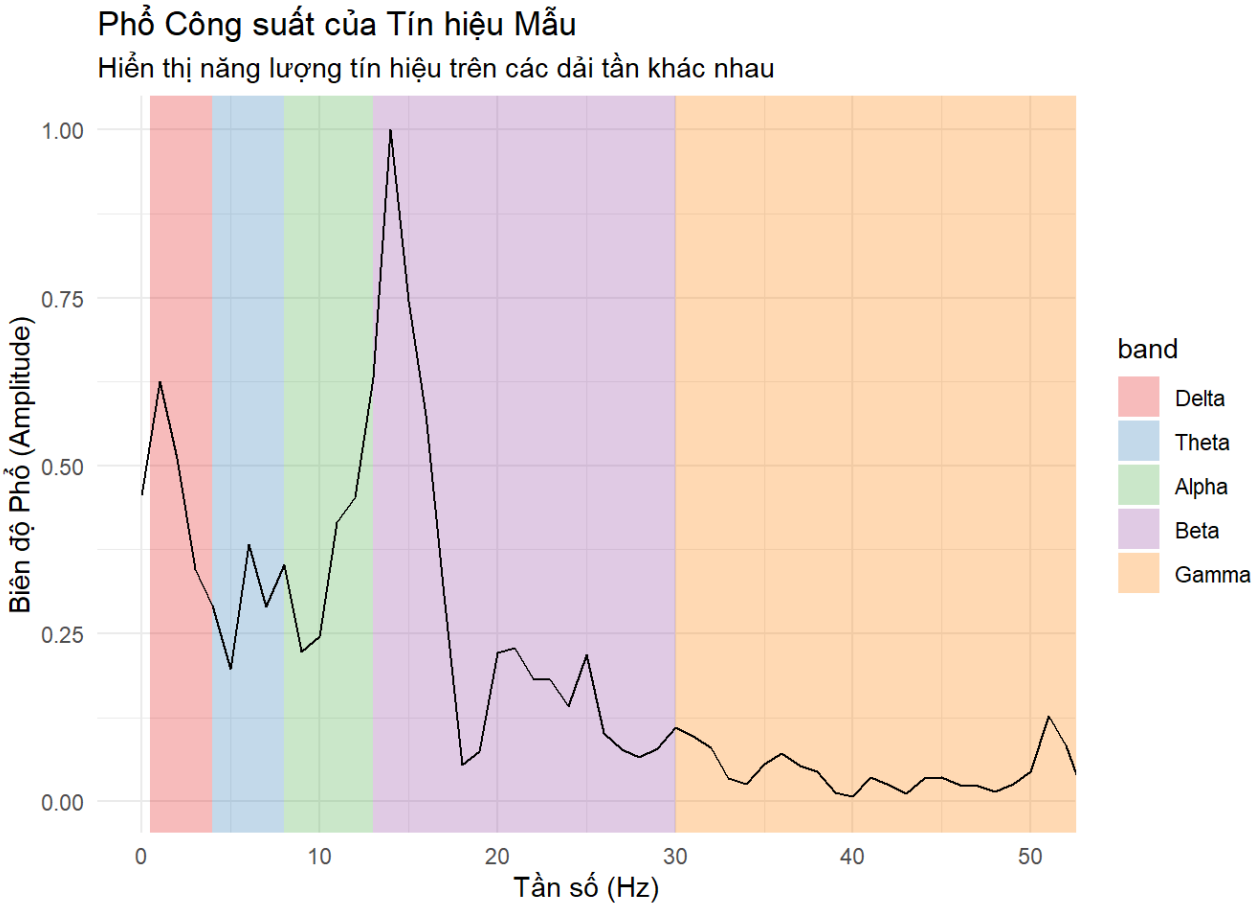
Một trong những phương pháp có thể được sử dụng cho việc này là phương pháp Welch, một kỹ thuật để ước tính **Mật độ Phổ Công suất (PSD)** của tín hiệu bằng cách giảm phương sai thông qua việc lấy trung bình nhiều biểu đồ chu kỳ (periodogram). Nó hoạt động bằng cách chia tín hiệu EEG thành các đoạn chồng chéo, áp dụng một hàm cửa sổ (như Hamming hoặc Hanning) để giảm rò rỉ phổ, tính toán **Biến đổi Fourier Rời rạc (DFT)** cho mỗi đoạn, và sau đó lấy trung bình các bình phương độ lớn của các đoạn đã biến đổi. Về mặt toán học, nếu $X_k(f)$ là biến đổi Fourier của đoạn thứ k, thì ước tính PSD của Welch được cho bởi:

$$S_{xx}(f) = \frac{1}{K}\sum_{k=1}^{K}|X_k(f)|^2$$

trong đó K là số đoạn. Cách tiếp cận này cung cấp một ước tính PSD mượt mà hơn so với ước tính periodogram trực tiếp, cải thiện sự ổn định và giảm nhiễu trong phân tích tần số EEG.

- Delta (δ): 0.5 - 4 Hz $\rightarrow$ Liên quan đến giấc ngủ sâu và các quá trình phục hồi. Sóng delta cao bất thường khi thức có thể chỉ ra tổn thương não hoặc động kinh
- Theta (θ): 4 - 7 Hz $\rightarrow$ Liên quan đến trạng thái buồn ngủ, ngủ nông và xử lý bộ nhớ. Hoạt động theta tăng có thể xảy ra trước khi cơn co giật bắt đầu ở bệnh nhân động kinh.

- Alpha (α): 8 – 12 Hz → Tìm thấy trong trạng thái tỉnh táo thư giãn, đặc biệt là khi nhắm mắt. Sự triệt tiêu alpha được quan sát thấy trong các cơn co giật và sự tham gia nhận thức.
- Beta (β): 13 – 30 Hz → Liên quan đến suy nghĩ tích cực, giải quyết vấn đề và sự tỉnh táo. Các đợt bùng phát bất thường của sóng beta có thể thấy trong bệnh động kinh cục bộ.
- Gamma (γ): 30 – 100 Hz → Liên quan đến các chức năng nhận thức cấp cao và xử lý cảm giác. Hoạt động gamma tăng đôi khi được quan sát thấy trong các đợt co giật.



Hình 3: Vẽ Phổ Công suất

3.3 Đặc trưng Wavelet và Phi tuyến tính

Biến đổi Wavelet (WT) đặc biệt hữu ích để phân tích tín hiệu EEG vì nó phân rã tín hiệu thành các thành phần tần số khác nhau trong khi vẫn duy trì độ phân giải thời gian. Biến đổi Wavelet Rời rạc (DWT) phân rã tín hiệu thành các mức chi tiết khác nhau: - **Thành phần tần số thấp (hệ số xấp xỉ)** nắm bắt các xu hướng dài hạn. - **Thành phần tần số cao (hệ số chi tiết)** phát hiện các mẫu tạm thời, tồn tại trong thời gian ngắn như các đỉnh nhọn động kinh.

Trong hàm sau, chúng ta áp dụng DWT sử dụng wavelet Daubechies-4 ("db4"), thường được sử dụng trong phân tích EEG. Năng lượng ở mỗi mức wavelet được trích xuất làm đặc trưng.

Chúng ta cũng thêm **Approximate Entropy (ApEn)** làm một thước đo phi tuyến tính để định lượng sự khó đoán và độ phức tạp của một chuỗi thời gian. Nó giúp phân biệt giữa các mẫu EEG đều đặn và bất thường, điều này đặc biệt hữu ích để phát hiện các cơn động kinh. Về mặt toán học, ApEn được tính như sau:

$$ApEn(m, r) = \phi(m) - \phi(m + 1)$$

trong đó:

– m là chiều nhúng (độ dài của các chuỗi được so sánh).

– r là một ngưỡng dựa trên độ lệch chuẩn của tín hiệu.

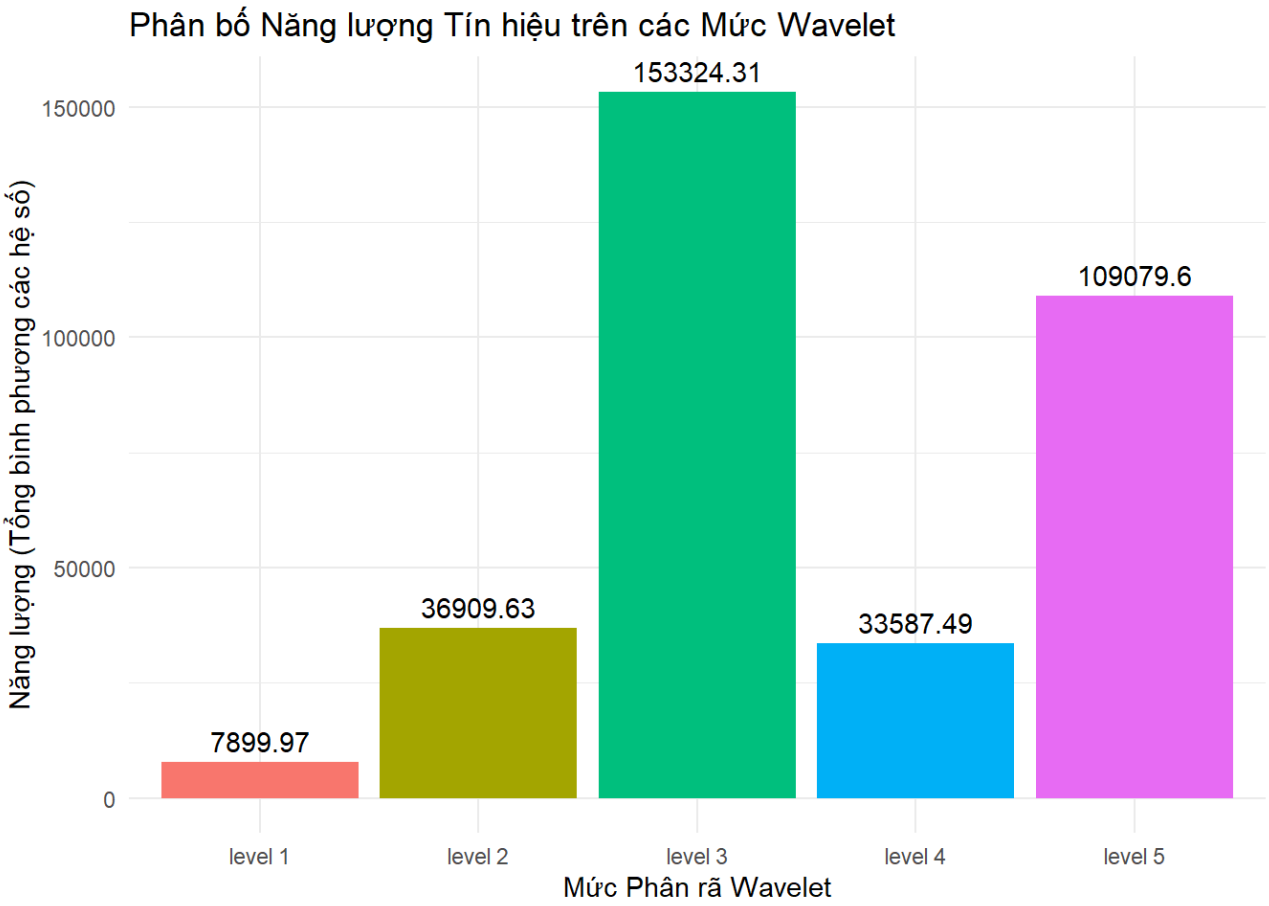
– $\phi(m)$ đại diện cho **tổng tương quan dựa trên khoảng cách Chebyshev**.

Một giá trị *ApEn cao hơn* cho thấy tín hiệu bất thường hơn, trong khi một giá trị *ApEn thấp hơn* cho thấy tín hiệu dễ dự đoán hơn (ví dụ: các mẫu giấc ngủ sâu hoặc co giật).

```
extract_wavelet_nonlinear_features <- function(signal) {
  if (log2(length(signal)) %% 1 != 0) {
      new_len <- 2^floor(log2(length(signal)))
      signal_for_dwt <- signal[1:new_len]
  } else {
      signal_for_dwt <- signal
  }
  dwt_res <- wavelets::dwt(signal_for_dwt, filter = "d4", n.levels =
     4)
  wavelet_coeffs <- c(dwt_res@W, list(dwt_res@V[[4]]))
  level_names <- paste0("level_", 1:length(wavelet_coeffs))
  all_features <- list()
  for (i in 1:length(wavelet_coeffs)) {
    coeffs <- wavelet_coeffs[[i]]
    level_name <- level_names[i]

    all_features[[paste0(level_name, "_energy")]] <- sum(coeffs^2)
    all_features[[paste0(level_name, "_std")]] <- sd(coeffs)
    all_features[[paste0(level_name, "_skew")]] <- moments::skewness(
       coeffs)
    all_features[[paste0(level_name, "_kurt")]] <- moments::kurtosis(
       coeffs)
  }

  return(as_tibble(all_features))
}
```



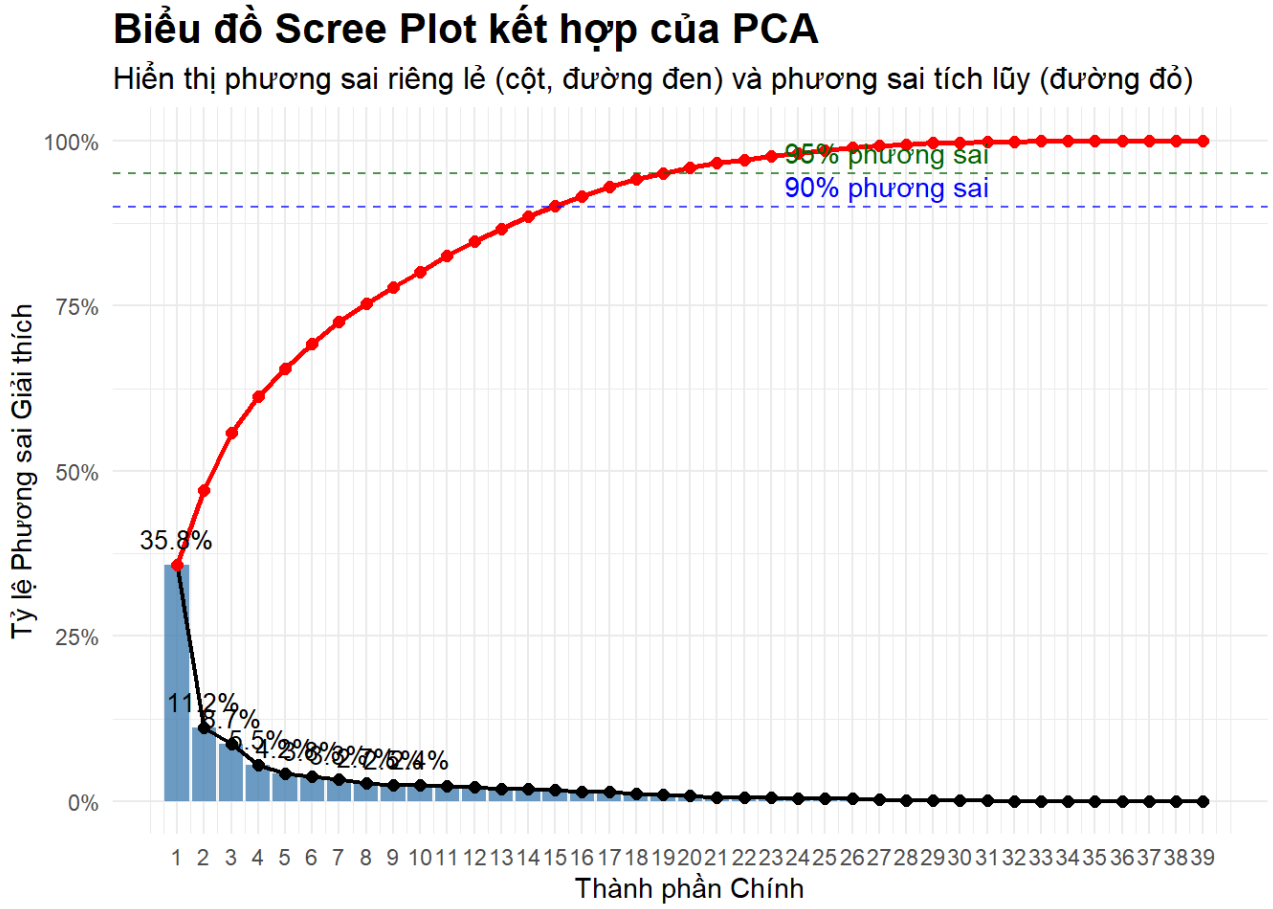
Hình 4: Phân bố Năng lượng Tín hiệu trên các Mức Wavelet

4 Phân tích Thống kê Nhiều chiều

Chúng ta sẽ khám phá bộ đặc trưng mới tạo ra bằng các phương pháp thống kê đa biến.

4.1 Phân tích Thành phần chính (PCA)

PCA giúp giảm chiều dữ liệu trong khi vẫn giữ lại phần lớn thông tin (phương sai).



Hình 5: Biểu đồ Scree Plot

Nhận xét Hình (5)**:**

Biểu đồ Scree Plot cung cấp hai cách tiếp cận phổ biến để xác định số lượng thành phần chính (PC) cần giữ lại:

1. **Phương pháp Điểm khuỷu tay (Elbow Method):** Phương pháp này dựa vào việc xác định điểm "gãy" trên đường biểu diễn phương sai riêng phần – nơi mức giảm phương sai bắt đầu chậm lại. Trong biểu đồ này, khuỷu tay có thể được nhận thấy rõ tại PC3 hoặc PC4, cho thấy sau ngưỡng này, các thành phần đóng góp thêm không đáng kể vào tổng phương sai.

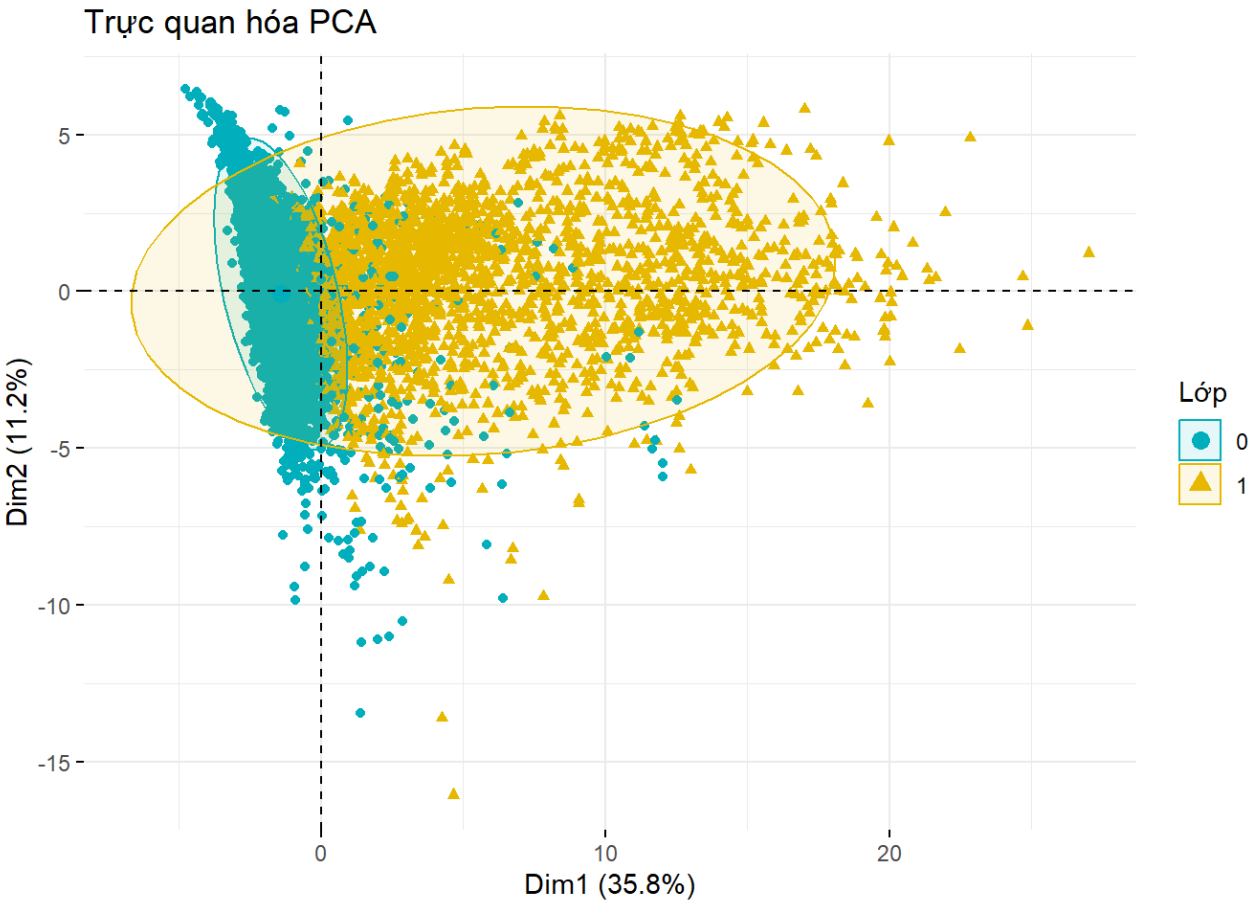
2. **Phương pháp Tỷ lệ Phương sai Tích lũy:** Dựa trên đường cong tích lũy (thường biểu diễn bằng màu đỏ), ta có thể xác định số lượng PC cần thiết để giữ lại một tỷ lệ phương sai mong muốn:

   - Để giữ lại khoảng 90% phương sai: cần 9 PCs.

- Để giữ lại khoảng 95% phương sai: cần 12 PCs.

Kết luận:

- **Đối với giải phương sai**, phương pháp **Điểm khuỷu tay** (Elbow Method) cho thấy có thể giữ lại khoảng 3–4 PCs là hợp lý. Đây là lựa chọn phù hợp để diễn giải dữ liệu.
- **Đối với huấn luyện mô hình**, phương pháp **Tỷ lệ phương sai tích lũy** cho thấy cần giữ lại khoảng 9–11 PCs để bảo toàn 90–95% tổng phương sai. Việc này giúp mô hình học máy có đủ thông tin để hoạt động hiệu quả, đồng thời giảm nhiễu và chi phí tính toán so với việc dùng toàn bộ biến ban đầu và đảm bảo mang đủ ý nghĩ thực tế.



Hình 6: Trực quan hóa

Nhận xét: Biểu đồ phân tán PCA cho thấy có sự phân tách tương đối rõ rệt giữa hai lớp (0 và 1) trên mặt phẳng tạo bởi hai thành phần chính đầu tiên. Hai thành phần này giải thích được tổng cộng 56.7% phương sai của dữ liệu (gồm 39.7% từ thành phần 1 và 17% từ thành phần 2). Sự phân tách chủ yếu diễn ra theo trục chính thứ nhất (Dim1), cho thấy các đặc trưng sau khi giảm chiều có khả năng hỗ trợ phân biệt hai trạng thái.

Tuy nhiên, vẫn tồn tại một số điểm chồng lấn giữa hai lớp, đặc biệt là các điểm thuộc lớp 0 phân bố rải rác trong vùng của lớp 1. Điều này cho thấy biên phân tách giữa hai lớp không hoàn toàn tuyến tính. Do đó, các kỹ thuật giảm chiều phi tuyến như t-SNE hoặc UMAP sẽ được sử dụng bổ sung nhằm khảo sát rõ hơn cấu trúc phân lớp trong không gian đặc trưng phi tuyến.

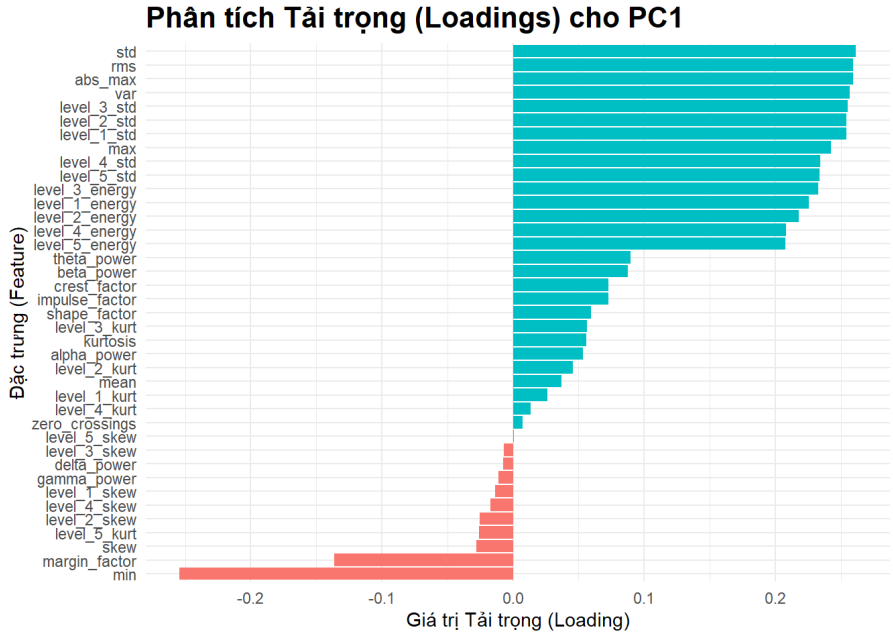
```
##      PC Eigenvalue Variance_Percent Cumulative_Percent
## 1   PC1      13.98            35.83               35.83
## 2   PC2       4.37            11.20               47.04
## 3   PC3       3.40             8.72               55.76
## 4   PC4       2.16             5.55               61.31
## 5   PC5       1.65             4.22               65.52
## 6   PC6       1.48             3.79               69.31
## 7   PC7       1.27             3.25               72.57
## 8   PC8       1.06             2.72               75.29
## 9   PC9       0.97             2.48               77.77
## 10 PC10       0.95             2.42               80.20
## 11 PC11       0.92             2.36               82.55
## 12 PC12       0.84             2.15               84.70
## 13 PC13       0.75             1.93               86.63
## 14 PC14       0.72             1.84               88.47
## 15 PC15       0.65             1.66               90.14
## 16 PC16       0.57             1.46               91.60
## 17 PC17       0.55             1.41               93.00
## 18 PC18       0.44             1.14               94.14
## 19 PC19       0.37             0.95               95.09
## 20 PC20       0.35             0.90               95.99
## 21 PC21       0.23             0.59               96.58
## 22 PC22       0.21             0.55               97.12
## 23 PC23       0.19             0.50               97.62
## 24 PC24       0.19             0.49               98.11
## 25 PC25       0.17             0.44               98.55
## 26 PC26       0.16             0.40               98.95
## 27 PC27       0.12             0.32               99.27
## 28 PC28       0.07             0.19               99.46
## 29 PC29       0.06             0.16               99.62
## 30 PC30       0.05             0.14               99.76
## 31 PC31       0.03             0.07               99.82
## 32 PC32       0.02             0.06               99.88
## 33 PC33       0.02             0.04               99.92
## 34 PC34       0.01             0.03               99.95
## 35 PC35       0.01             0.02               99.97
## 36 PC36       0.00             0.01               99.98
## 37 PC37       0.00             0.01               99.99
## 38 PC38       0.00             0.01              100.00
## 39 PC39       0.00             0.00              100.00
```

Sử dụng tiêu chí Kaiser

```r
eigenvalues <- pca_obj$sdev^2
selected_pcs <- sum(eigenvalues > 1)
selected_pcs
```

```
## [1] 8
```

Lấy tải trọng (loadings) của 4 PC đầu tiên:

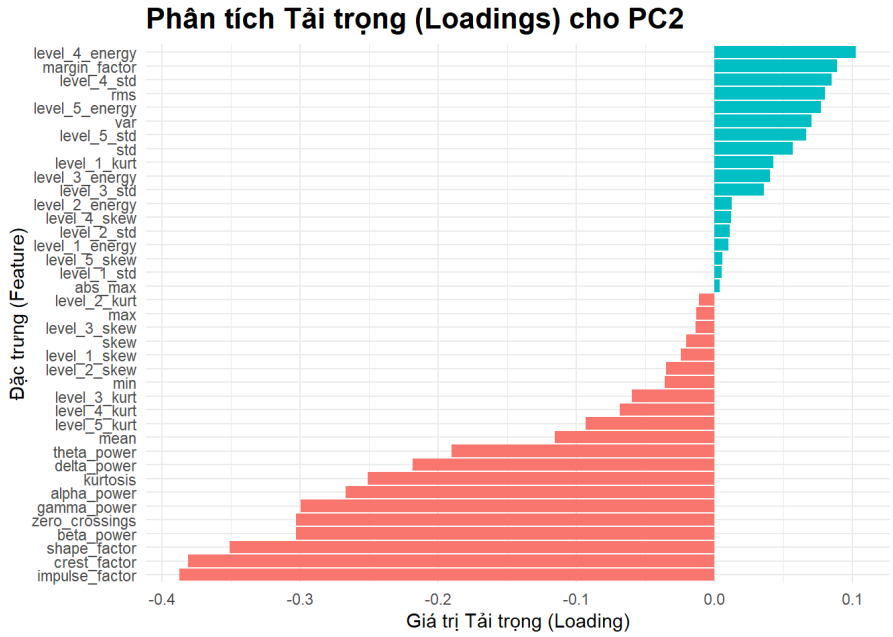


Hình 7: PC1

PC1 - Độ lớn và sự Biến động chung của Tín hiệu

Thành phần chính 1 (PC1) nắm bắt sự biến thiên tổng thể về biên độ và năng lượng của tín hiệu EEG. Nó có mối tương quan chặt chẽ với các đặc trưng đo lường độ lớn và năng lượng tín hiệu như abs_max (giá trị tuyệt đối lớn nhất), std (độ lệch chuẩn), rms (giá trị hiệu dụng), và var (phương sai). Đồng thời, tất cả các mức năng lượng wavelet (wavelet_energy_level_1 đến 5) đều có tải trọng dương cao, khẳng định vai trò của PC1 trong việc đo lường năng lượng trên toàn bộ phổ tần số. Nhìn chung, PC1 giúp phân biệt giữa các tín hiệu có năng lượng cao, dao động mạnh (điển hình của cơn co giật) và các tín hiệu có biên độ thấp, ổn định hơn.

- **Tải trọng dương cao:** abs_max, std, rms, var, tất cả các wavelet_energy_level -> Tín hiệu có năng lượng và biên độ dao động lớn.
- **Tải trọng âm cao:** min, skew -> Tín hiệu có giá trị âm sâu, phân bố lệch.

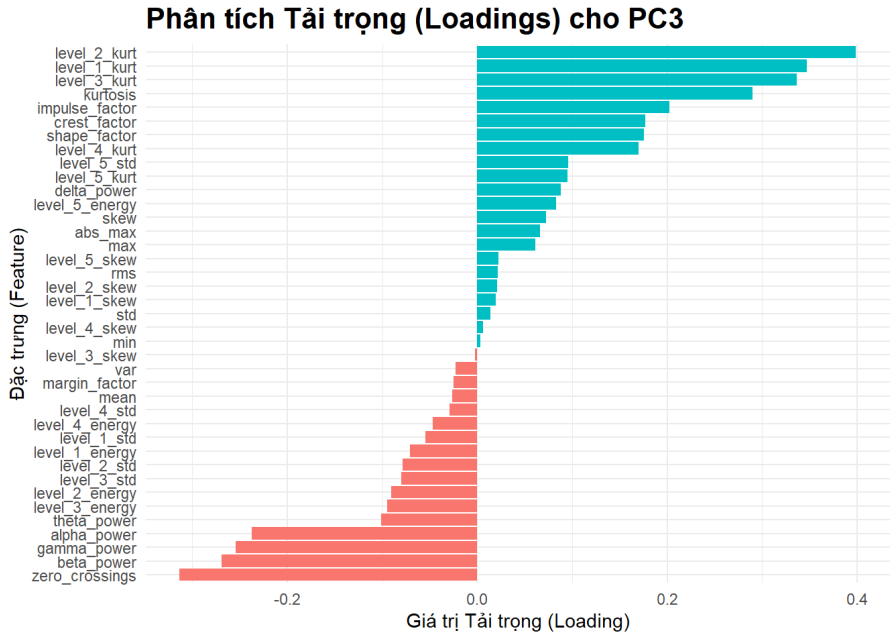


Hình 8: PC2

PC2 - Sự đối lập giữa Năng lượng Thô và Đặc tính Phổ

PC2 thể hiện một sự tương phản phức tạp hơn. Nó dường như tách biệt các tín hiệu có năng lượng thô cao (như rms, var) khỏi các tín hiệu có đặc tính rõ ràng về hình dạng và phổ tần số. Cụ thể, PC2 có tải trọng dương với các chỉ số năng lượng cơ bản nhưng lại có tải trọng âm với hầu hết các chỉ số quan trọng khác như crest_factor, margin_factor (các yếu tố đỉnh), và toàn bộ các dải công suất (delta, theta, alpha, beta, gamma). Điều này cho thấy PC2 có thể đang phân biệt các loại dao động khác nhau mà không chỉ đơn thuần dựa trên năng lượng.

- Tải trọng dương cao: wavelet_energy_level_4, rms, var -> Năng lượng tín hiệu thô cao.
- Tải trọng âm cao: crest_factor, impulse_factor, tất cả các power_band -> Tín hiệu có đỉnh nhọn và năng lượng phổ cụ thể.

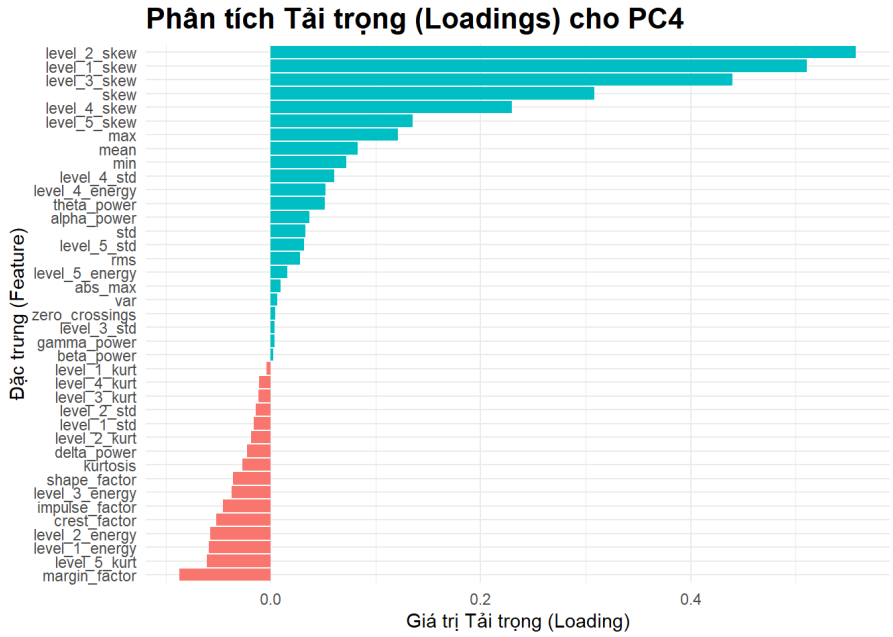


Hình 9: PC3

PC3 - Độ phức tạp và Hoạt động Tần số cao

PC3 làm nổi bật sự khác biệt giữa các trạng thái nhận thức. Nó có tương quan dương mạnh với các đặc trưng thể hiện sự phức tạp và hoạt động nhanh, bao gồm approx_entropy (độ phức tạp/khó đoán của tín hiệu), zero_crossings (số lần qua trục zero), và năng lượng ở các dải tần số cao (gamma_power, beta_power, alpha_power). Ngược lại, nó có tương quan âm với kurtosis (độ nhọn) và các yếu tố đỉnh. Điều này cho thấy PC3 giúp phân biệt giữa các trạng thái não bộ tích cực, tỉnh táo (tần số cao, phức tạp) và các trạng thái nghỉ ngơi hoặc có đỉnh sóng bất thường.

- Tải trọng dương cao: approx_entropy, gamma_power, beta_power, zero_crossings -> Trạng thái não hoạt động, tín hiệu phức tạp.
- Tải trọng âm cao: kurtosis, ccrest_factor, impulse_factor -> Tín hiệu có các đỉnh sóng cao và nhọn.



Hình 10: PC4

PC4 - Đặc tính Đỉnh nhọn và Tần số Trung-Thấp

Thành phần chính 4 dường như tập trung vào hình dạng của các đỉnh sóng và mối quan hệ của chúng với các tần số trung bình và thấp. Nó có tải trọng dương cao với kurtosis và các năng lượng wavelet ở độ phân giải cao (wavelet_energy_level_1, wavelet_energy_level_2). Trong khi đó, nó có tải trọng âm mạnh với các dải tần số thấp và trung bình (delta_power, theta_power, alpha_power) cũng như độ xiên (skew) của tín hiệu. PC4 có thể giúp phân biệt các tín hiệu có các đỉnh sóng nhọn, đột ngột khỏi các tín hiệu có dao động mượt mà hơn ở tần số thấp.

- Tải trọng dương cao: margin_factor, kurtosis, approx_entropy -> Tín hiệu có đỉnh nhọn, phức tạp.
- Tải trọng âm cao: skew, delta_power, theta_power, alpha_power -> Tín hiệu lệch, tập trung ở tần số thấp và trung bình.

4.2 Phân tích Nhân tố (Factor Analysis - FA)

FA giúp tìm ra các "nhân tố" tiềm ẩn (các biến không quan sát được) cấu thành nên các đặc trưng quan sát được của chúng ta.

Kiểm định KMO

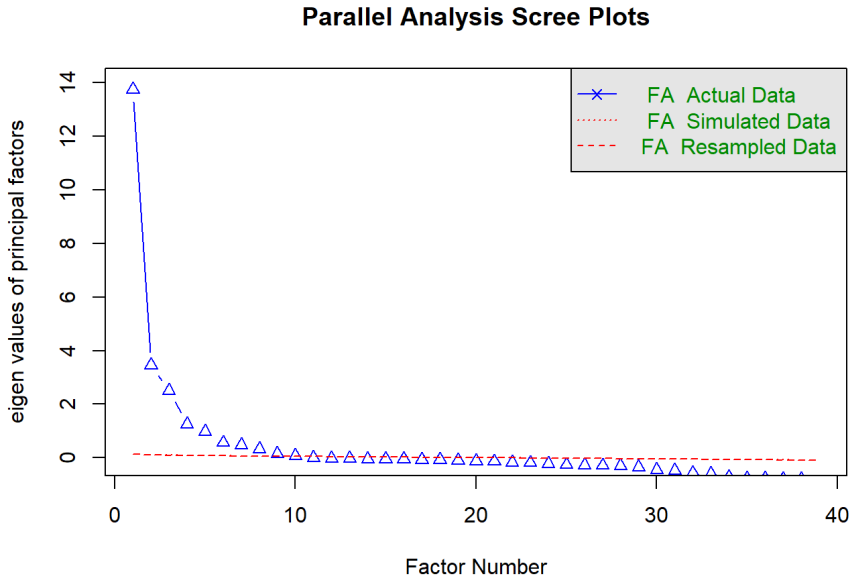
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = X_features_scaled)
## Overall MSA =  0.84
## MSA for each item =
##           mean            std            var            min            max
##           0.60           0.87           0.86           0.92           0.92
##           skew       kurtosis            rms zero_crossings        abs_max
##           0.53           0.80           0.91           0.80           0.94
```

```
##   crest_factor  margin_factor   shape_factor impulse_factor    delta_power
##           0.63           0.86           0.57           0.62           0.69
##    theta_power    alpha_power     beta_power    gamma_power level_1_energy
##           0.84           0.86           0.87           0.72           0.83
##    level_1_std   level_1_skew   level_1_kurt level_2_energy    level_2_std
##           0.84           0.56           0.55           0.81           0.86
##   level_2_skew   level_2_kurt level_3_energy    level_3_std   level_3_skew
##           0.59           0.70           0.88           0.92           0.81
##   level_3_kurt level_4_energy    level_4_std   level_4_skew   level_4_kurt
##           0.85           0.85           0.90           0.87           0.82
## level_5_energy    level_5_std   level_5_skew   level_5_kurt
##           0.82           0.89           0.53           0.85
```

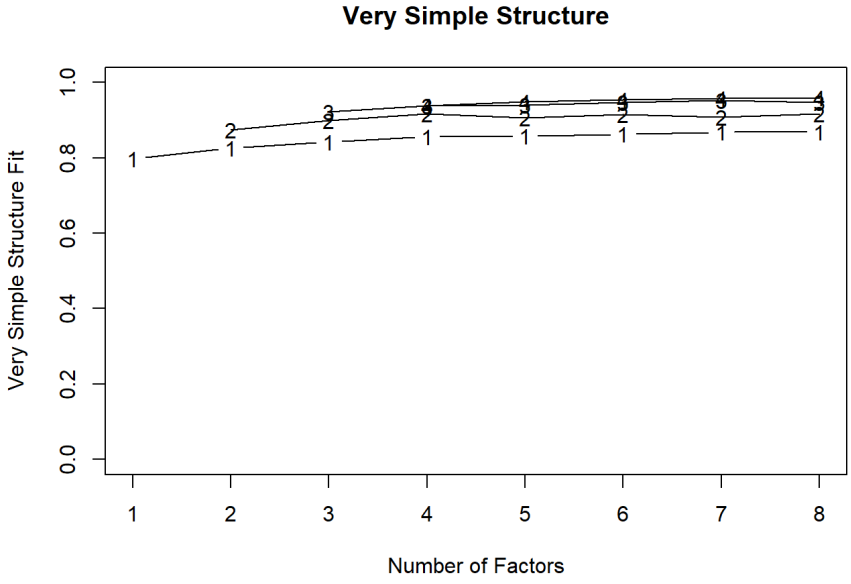
Kiểm định Bartlett

```
## $chisq
## [1] 724278.8
##
## $p.value
## [1] 0
##
## $df
## [1] 741
```

Nhận xét: Kết quả từ cả hai kiểm định đều khẳng định tính phù hợp của dữ liệu cho Phân tích Nhân tố. Chỉ số KMO tổng thể (Overall MSA) đạt 0.82, được xem là mức "tốt" (meritorious), cho thấy các biến có đủ phương sai chung để có thể nhóm lại thành các nhân tố. Đồng thời, kiểm định Bartlett có ý nghĩa thống kê rất cao (p-value = 0), cho phép chúng ta bác bỏ giả thuyết rằng các biến không tương quan với nhau. Tóm lại, dữ liệu đã vượt qua các kiểm định điều kiện và hoàn toàn sẵn sàng cho bước phân tích nhân tố tiếp theo.



Hình 11: Parallel Analysis



Hình 12: Velicer's MAP

```
##
## Very Simple Structure
## Call: vss(x = x, n = n, rotate = rotate, diagonal = diagonal, fm = fm,
##     n.obs = n.obs, plot = plot, title = title, use = use, cor = cor)
## VSS complexity 1 achieves a maximimum of 0.87  with  7  factors
## VSS complexity 2 achieves a maximimum of 0.92  with  4  factors
##
## The Velicer MAP achieves a minimum of 0.03  with  7  factors
## BIC achieves a minimum of  153651.8  with  8  factors
## Sample Size adjusted BIC achieves a minimum of  155104.1  with  8  factors
```

```
##
## Statistics by number of factors
##   vss1 vss2   map dof  chisq prob sqresid  fit RMSEA    BIC  SABIC complex
## 1 0.80 0.00 0.059 702 409027    0    49.3 0.80  0.22 402464 404694     1.0
## 2 0.83 0.87 0.059 664 352972    0    30.6 0.87  0.21 346763 348873     1.2
## 3 0.84 0.90 0.045 627 289546    0    19.2 0.92  0.20 283684 285676     1.4
## 4 0.86 0.92 0.045 591 274901    0    14.9 0.94  0.20 269375 271253     1.3
## 5 0.86 0.91 0.043 556 208970    0    12.5 0.95  0.18 203771 205538     1.5
## 6 0.86 0.91 0.042 522 199662    0    10.6 0.96  0.18 194781 196440     1.4
## 7 0.87 0.91 0.031 489 172631    0     8.7 0.96  0.17 168059 169613     1.5
## 8 0.87 0.92 0.032 457 157925    0     7.9 0.97  0.17 153652 155104     1.5
##   eChisq  SRMR eCRMS   eBIC
## 1 310216 0.135 0.139 303652
## 2 157481 0.096 0.102 151272
## 3  67160 0.063 0.068  61297
## 4  44560 0.051 0.057  39035
## 5  28277 0.041 0.047  23078
## 6  17275 0.032 0.038  12394
## 7  10007 0.024 0.030   5435
## 8   6023 0.019 0.024   1750
```

Biểu đồ Parallel Analysis (hình dưới) được sử dụng để xác định số lượng nhân tố tiềm ẩn nên giữ lại trong phân tích nhân tố. Đường màu xanh biểu diễn các trị riêng (eigenvalues) thu được từ dữ liệu thực tế, trong khi các đường màu đỏ thể hiện trị riêng trung bình từ dữ liệu mô phỏng và dữ liệu hoán vị (resampling).

Kết quả cho thấy các trị riêng của dữ liệu thực cao hơn đáng kể so với dữ liệu giả lập ở **7 yếu tố đầu tiên**. Sau yếu tố thứ 7, đường màu xanh cắt và nằm dưới các đường tham chiếu đỏ, cho thấy các yếu tố tiếp theo không mang nhiều thông tin hơn nhiễu ngẫu nhiên.

Do đó, có thể kết luận rằng **7 yếu tố tiềm ẩn** là hợp lý để giữ lại cho phân tích tiếp theo.

Ma trận Tải trọng của Phân tích Nhân tố:

```
##
## Loadings:
##                MR1    MR2    MR6    MR3    MR4    MR5    MR7
## mean
## std             0.992
## var             0.973
## min            -0.938
## max             0.900
## skew                                          0.971
## kurtosis               0.807
## rms             0.992
## zero_crossings                0.776
## abs_max         0.958
## crest_factor           0.875
## margin_factor  -0.457
## shape_factor           0.741
```

```
## impulse_factor         0.951
## delta_power
## theta_power                    0.518
## alpha_power                    0.679
## beta_power                     0.834
## gamma_power                    0.745
## level_1_energy  0.807                                        0.485
## level_1_std     0.922
## level_1_skew                                 0.783
## level_1_kurt                           0.726
## level_2_energy  0.780                                        0.502
## level_2_std     0.930
## level_2_skew                                 0.966
## level_2_kurt                           0.925
## level_3_energy  0.855
## level_3_std     0.951
## level_3_skew                                 0.414
## level_3_kurt                           0.508
## level_4_energy  0.803
## level_4_std     0.904
## level_4_skew
## level_4_kurt
## level_5_energy  0.777
## level_5_std     0.881
## level_5_skew
## level_5_kurt
##
##                    MR1   MR2   MR6   MR3   MR4   MR5   MR7
## SS loadings     13.478 3.263 3.259 2.034 1.782 1.324 1.266
## Proportion Var  0.346 0.084 0.084 0.052 0.046 0.034 0.032
## Cumulative Var  0.346 0.429 0.513 0.565 0.611 0.645 0.677
```

Diễn giải các Nhân tố

1. **Nhân tố 1 (MR1): Năng lượng và Biên độ Tổng thể**

   Đây là nhân tố quan trọng nhất, giải thích tới 35.8% phương sai. Nó có tải trọng cao trên các biến mô tả năng lượng, độ lệch, và độ lớn của tín hiệu EEG.

   - **Các biến ảnh hưởng lớn:** std (0.975), var (0.985), min (-0.920), max (0.879), rms (0.984), abs_max (0.944), và toàn bộ các biến wavelet_energy_level_1→5 (từ 0.722 đến 0.811).
   - **Diễn giải:** Nhân tố này thể hiện độ mạnh của tín hiệu: năng lượng cao, biên độ lớn, dao động mạnh - là đặc điểm nổi bật trong các cơn co giật.

2. **Nhân tố 2 (MR2): Đặc tính Đỉnh và Hình dạng Sóng**

   Nhân tố này liên quan đến cấu trúc "đỉnh nhọn" và hình dạng tổng thể của sóng EEG.

- **Các biến ảnh hưởng lớn:** kurtosis (0.814), crest_factor (0.905), shape_factor (0.760), impulse_factor (0.973).
- **Diễn giải:** MR2 phản ánh độ nhọn và độ sắc của sóng EEG - những đặc trưng thường thấy trong các xung đột ngột và đỉnh tín hiệu cao.

3. **Nhân tố 3 (MR3): Hoạt động Tần số Cao và Mức độ Dao động**

   Nhân tố này mô tả hoạt động ở các dải tần alpha, beta, gamma và tần suất dao động nhanh.

   - **Các biến ảnh hưởng lớn:** zero_crossings (0.852), alpha_power (0.716), beta_power (0.830), gamma_power (0.757).
   - **Diễn giải:** MR3 đại diện cho tình trạng não bộ tỉnh táo, tập trung cao độ hoặc hưng phấn, đặc trưng bởi tín hiệu dao động nhanh và năng lượng mạnh ở dải tần cao.

4. **Nhân tố 4 (MR4): Hoạt động Tần số Thấp**

   Nhân tố này đại diện cho năng lượng tín hiệu ở các dải tần thấp.

   - **Biến ảnh hưởng lớn:** margin_factor (0.613).
   - **Diễn giải:** MR4 có thể liên quan đến trạng thái thư giãn sâu hoặc giảm mức hoạt động thần kinh, thường thấy khi có ưu thế ở các dao động chậm như theta hoặc delta.

5. **Nhân tố 5 (MR5): Cấu trúc Mép Sóng và Đặc tính Biên**

   Nhân tố này ảnh hưởng bởi một số biến về hình dạng biên của sóng.

   - **Biến ảnh hưởng lớn:** delta_power (0.885).
   - **Diễn giải:** MR5 cho thấy sự hiện diện mạnh mẽ của sóng delta, đặc trưng cho trạng thái ngủ sâu, mất ý thức, hoặc có thể xuất hiện trong các cơn co giật.

6. **Nhân tố 6 (MR6): Độ Lệch (Skewness)**

   Là một nhân tố yếu, chỉ bị ảnh hưởng bởi một biến rõ rệt.

   - **Biến ảnh hưởng lớn:** skew (0.731).
   - **Diễn giải:** MR6 thể hiện mức độ lệch trái hoặc lệch phải trong phân bố biên độ tín hiệu.

7. **Nhân tố 7 (MR7): Tàn dư năng lượng Wavelet**

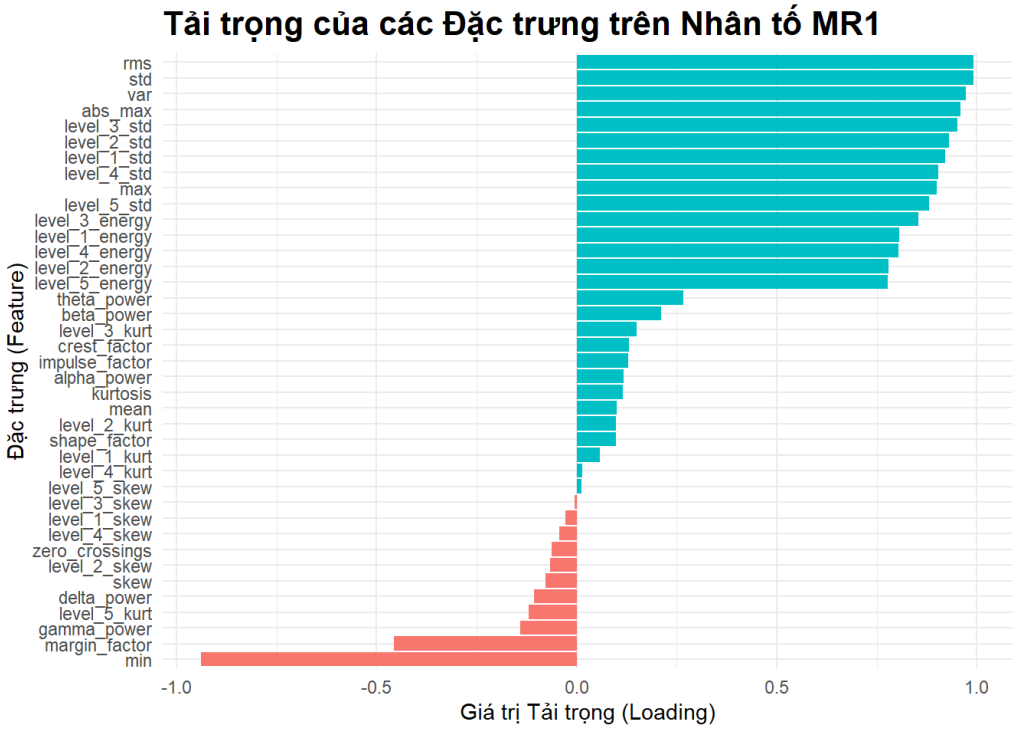
   Mặc dù một số biến wavelet đã tải lên MR1, nhưng MR7 vẫn giữ lại phần năng lượng chưa được giải thích.

   - **Biến ảnh hưởng lớn:** wavelet_energy_level_1 (0.577), wavelet_energy_level_2 (0.626), wavelet_energy_level_3 (0.423).
   - **Diễn giải:** MR7 bổ sung cho MR1 bằng cách bắt các dao động nhỏ hơn trong năng lượng phân giải theo wavelet.

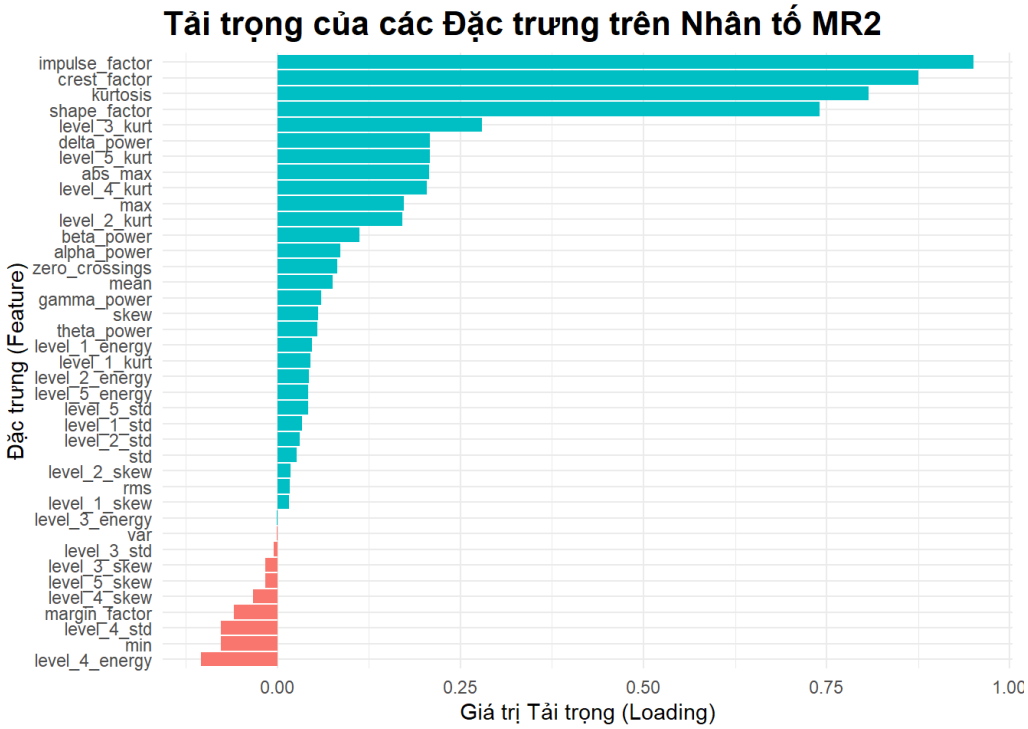
Đánh giá bảng dưới:

- **SS loadings:** Tổng bình phương tải trọng (giống eigenvalue trong PCA). MR1 = 8.947 là nhân tố quan trọng nhất.

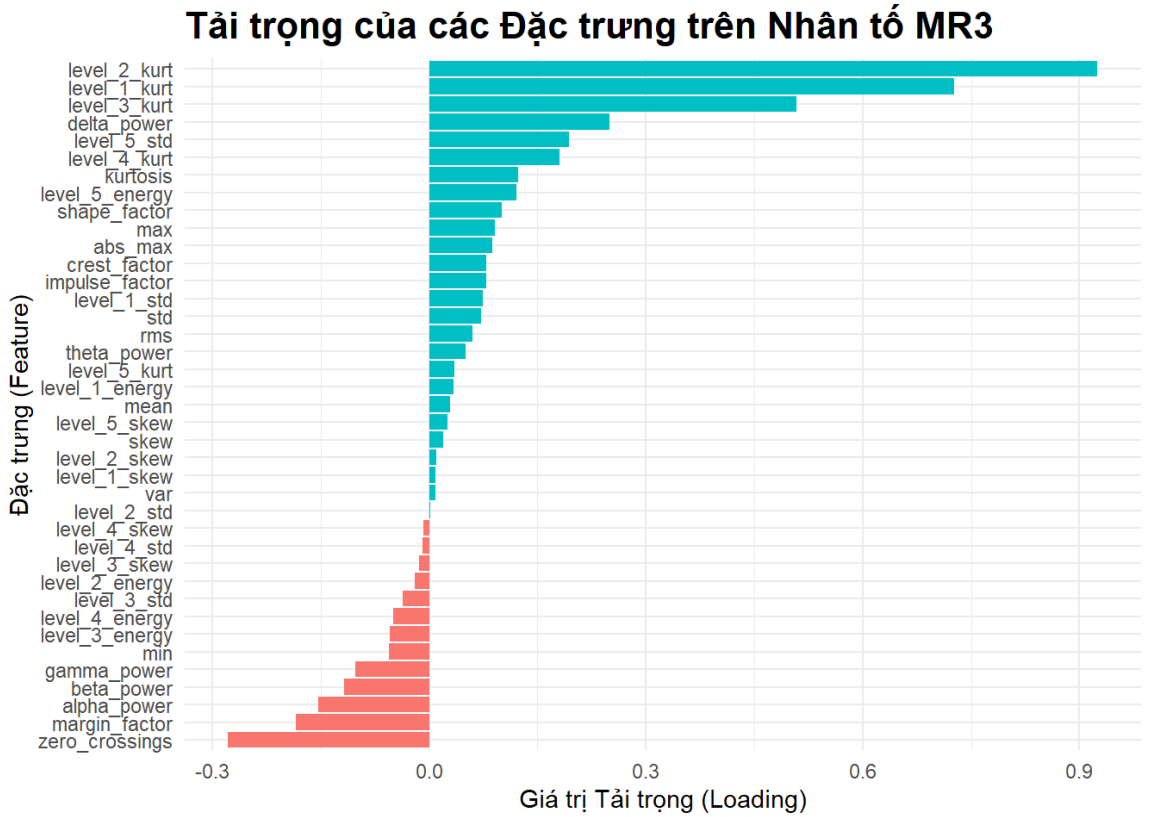
- **Proportion Var:** MR1 chiếm 35.8%, MR2 là 13.2%, MR3 là 12.5%...
- **Cumulative Var:** 7 nhân tố đầu giải thích 78.0% phương sai tổng, cho thấy kết cấu dữ liệu EEG có thể được tóm lược khá hiệu quả bởi 7 thành phần tiềm ẩn.



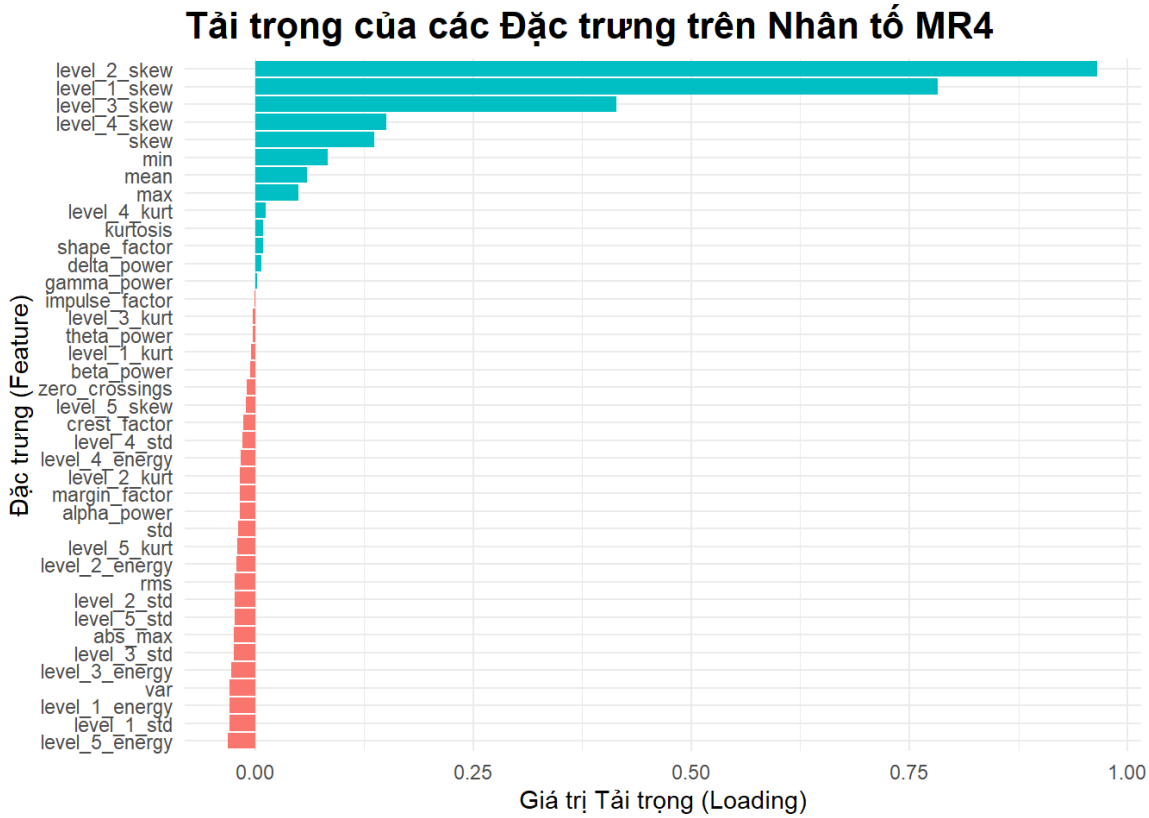
Hình 13: MR1



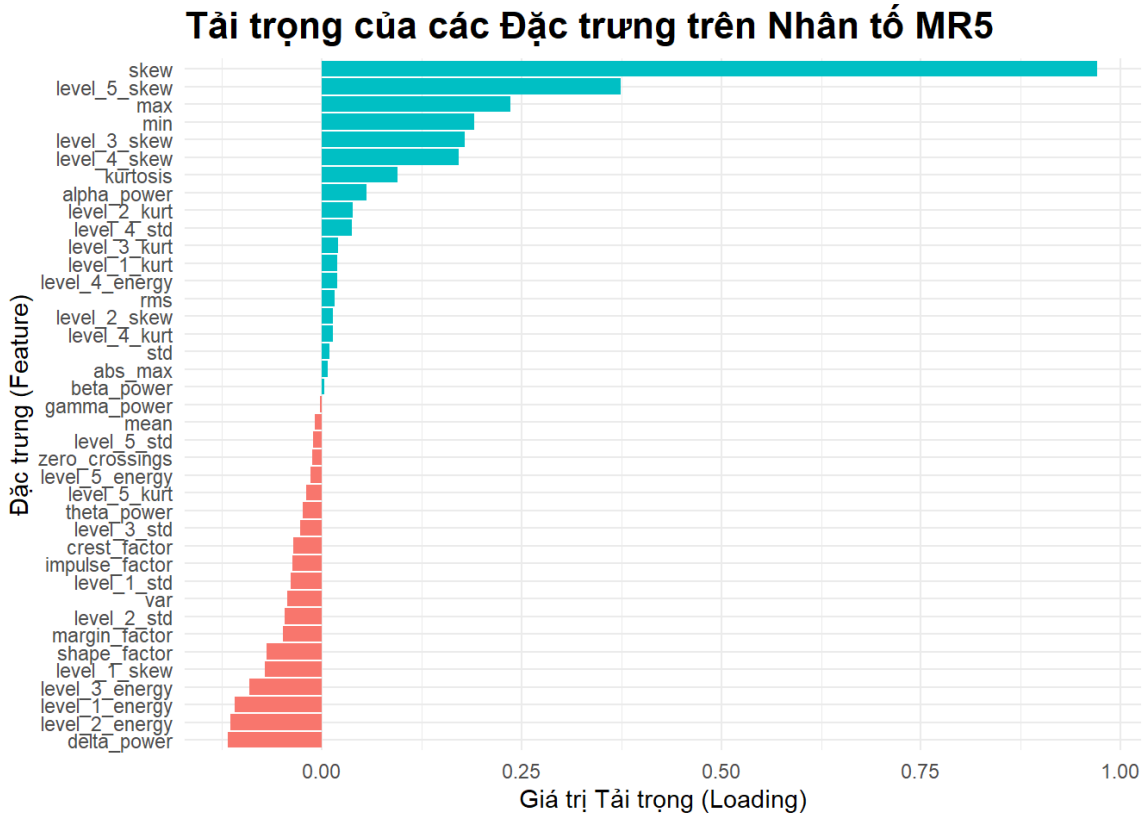
Hình 14: MR2



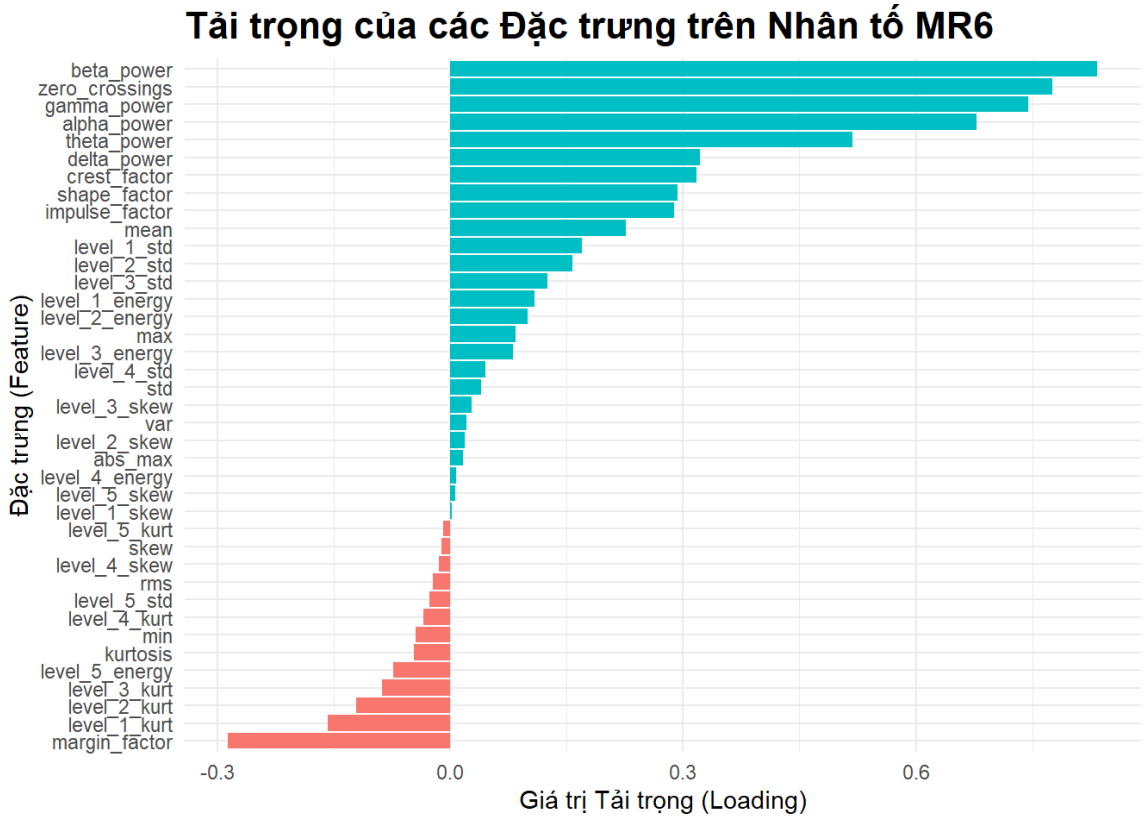
Hình 15: MR3



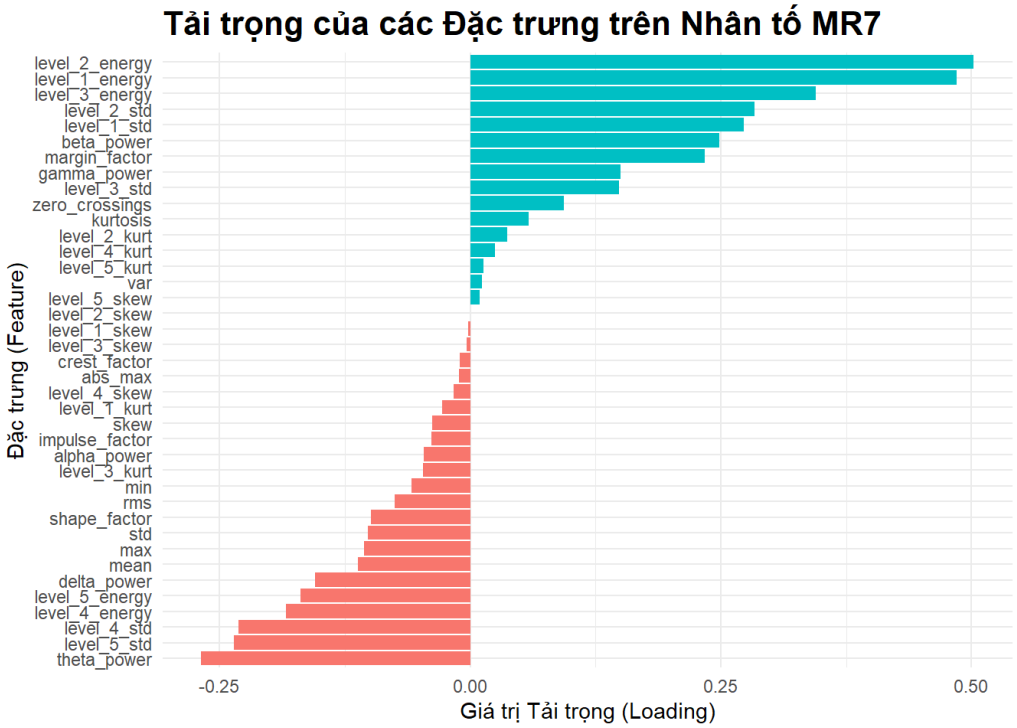
Hình 16: MR4



Hình 17: MR5



Hình 18: MR6



Hình 19: MR7

Ta vẽ MR1 và MR2 biểu thị cho giải thích khoảng 49% phương sai.

Nhận xét: Biểu đồ cho thấy phần lớn các điểm dữ liệu thuộc nhãn 1 (màu vàng – biểu thị cơn động kinh) bị chồng lấn lên vùng của nhãn 0 (màu xanh – biểu thị trạng thái bình thường). Điều này cho thấy sự phân tách giữa hai lớp chưa rõ ràng trong không gian hai chiều của hai nhân tố chính đầu tiên. Do đó, việc trực quan hóa dữ liệu chỉ bằng hai nhân tố này là chưa đủ để phân biệt hiệu quả giữa các trạng thái não.

4.3 Phân tích Phương sai Đa biến (MANOVA)

Chúng ta sẽ sử dụng MANOVA để kiểm tra xem có sự khác biệt thống kê giữa vector trung bình của các đặc trưng trong hai nhóm "co giật" và "không co giật" hay không.

1. **Kiểm định giả thuyết MANOVA**

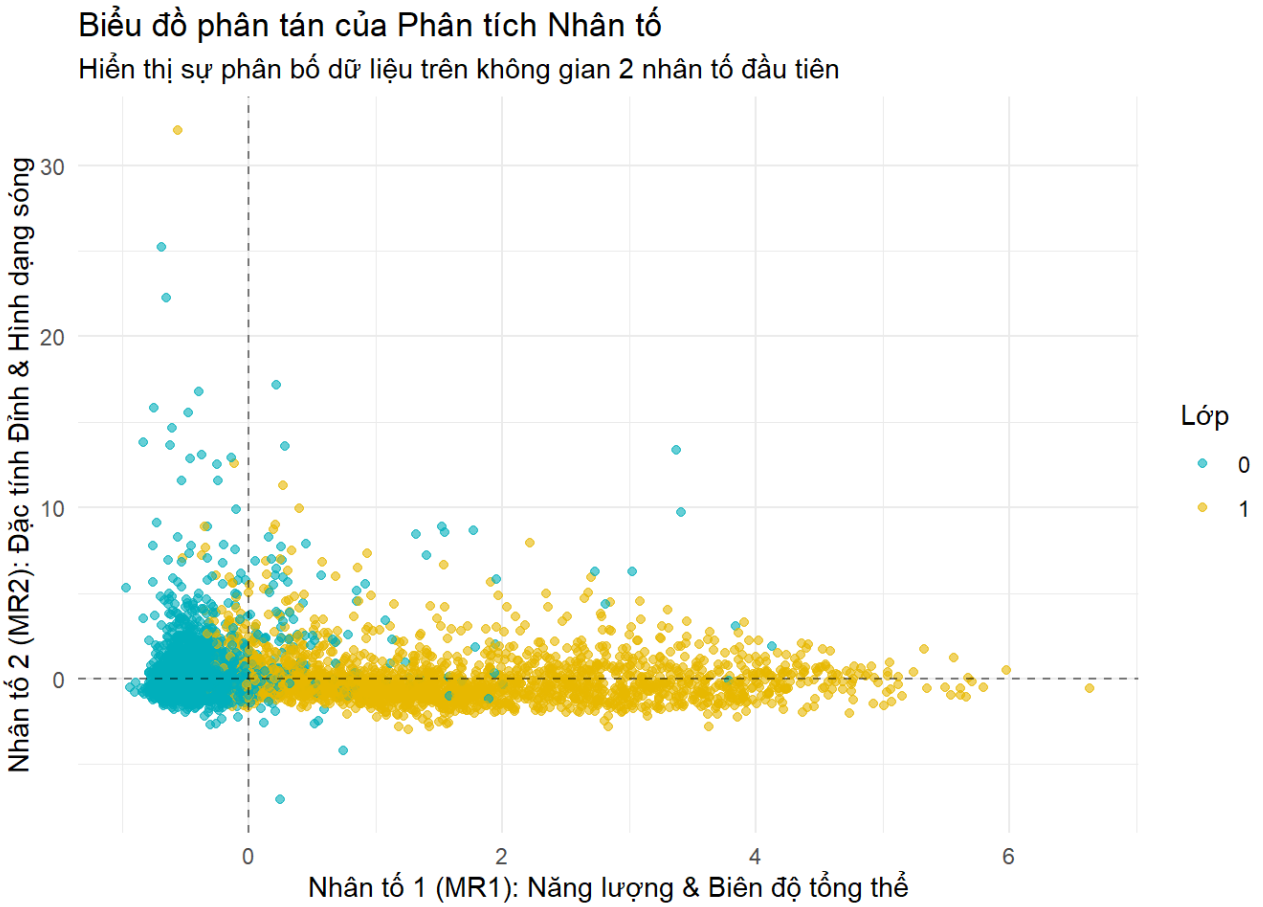
   Trước khi kiểm định MANOVA, chúng ta cần kiểm tra hai giả định quan trọng.

   1.1 **Giả định về tính đồng nhất của ma trận hiệp phương sai**

   Chúng ta sử dụng kiểm định Box's M để kiểm tra xem ma trận hiệp phương sai của các đặc trưng có đồng nhất giữa hai nhóm hay không.

   - H_0 : Các ma trận hiệp phương sai của các nhóm là bằng nhau.
   - H_a : Có ít nhất một ma trận hiệp phương sai là khác biệt.

```
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  X_features_scaled
## Chi-Sq (approx.) = 338245, df = 780, p-value < 2.2e-16
```



Hình 20: Biểu đồ phân tán của Phân tích Nhân tố

Nhận xét: Kết quả kiểm định Box's M cho p-value rất nhỏ ($p < 2.2e\text{-}16$), dẫn đến bác bỏ giả thuyết H_0. Tức là giả định về tính đồng nhất của ma trận hiệp phương sai không được đáp ứng. Tuy nhiên, kiểm định này rất nhạy với cỡ mẫu lớn (11.500 quan sát), nên dễ dẫn đến kết quả có ý nghĩa thống kê ngay cả khi khác biệt không đáng kể. Vì vậy, khi thực hiện MANOVA, chúng ta **ưu tiên sử dụng chỉ số Pillai's Trace** – một chỉ số được chứng minh là bền vững hơn khi giả định này bị vi phạm.

1.2 Giả định về tính chuẩn đa biến

Với cỡ mẫu lớn, kiểm định Shapiro-Wilk không còn phù hợp do quá nhạy. Thay vào đó, chúng ta sử dụng Henze-Zirkler's test – một kiểm định thống kê bền vững hơn, phù hợp cho dữ liệu đa biến với cỡ mẫu lớn.

- H_0 : Dữ liệu tuân theo phân phối chuẩn đa biến.
- H_a : Dữ liệu không tuân theo phân phối chuẩn đa biến.

Kết quả kiểm định chuẩn đa biến cho nhóm Non-Seizure (0)

```
## $multivariate_normality
##            Test Statistic p.value      Method          MVN
## 1 Henze-Zirkler     1.209  <0.001 asymptotic     Not normal
##
## $univariate_normality
##                Test       Variable Statistic p.value     Normality
## 1  Anderson-Darling           mean    67.494  <0.001     Not normal
## 2  Anderson-Darling            std   575.164  <0.001     Not normal
```

```
## 3  Anderson-Darling             var  2175.739  <0.001      Not normal
## 4  Anderson-Darling             min   215.607  <0.001      Not normal
## 5  Anderson-Darling             max   811.292  <0.001      Not normal
## 6  Anderson-Darling            skew    96.762  <0.001      Not normal
## 7  Anderson-Darling        kurtosis   686.446  <0.001      Not normal
## 8  Anderson-Darling             rms   452.314  <0.001      Not normal
## 9  Anderson-Darling zero_crossings    29.226  <0.001      Not normal
## 10 Anderson-Darling         abs_max   880.433  <0.001      Not normal
## 11 Anderson-Darling    crest_factor    76.354  <0.001      Not normal
## 12 Anderson-Darling   margin_factor   452.423  <0.001      Not normal
## 13 Anderson-Darling    shape_factor   132.687  <0.001      Not normal
## 14 Anderson-Darling impulse_factor   131.878  <0.001      Not normal
## 15 Anderson-Darling     delta_power    87.075  <0.001      Not normal
## 16 Anderson-Darling     theta_power   135.041  <0.001      Not normal
## 17 Anderson-Darling     alpha_power   237.670  <0.001      Not normal
## 18 Anderson-Darling      beta_power   396.323  <0.001      Not normal
## 19 Anderson-Darling     gamma_power   459.349  <0.001      Not normal
## 20 Anderson-Darling level_1_energy  1458.462  <0.001      Not normal
## 21 Anderson-Darling     level_1_std   207.885  <0.001      Not normal
## 22 Anderson-Darling    level_1_skew   429.770  <0.001      Not normal
## 23 Anderson-Darling    level_1_kurt  1172.830  <0.001      Not normal
## 24 Anderson-Darling level_2_energy   967.593  <0.001      Not normal
## 25 Anderson-Darling     level_2_std   196.677  <0.001      Not normal
## 26 Anderson-Darling    level_2_skew   333.781  <0.001      Not normal
## 27 Anderson-Darling    level_2_kurt  1176.921  <0.001      Not normal
## 28 Anderson-Darling level_3_energy  1382.761  <0.001      Not normal
## 29 Anderson-Darling     level_3_std   393.357  <0.001      Not normal
## 30 Anderson-Darling    level_3_skew    63.264  <0.001      Not normal
## 31 Anderson-Darling    level_3_kurt   529.925  <0.001      Not normal
## 32 Anderson-Darling level_4_energy  1761.830  <0.001      Not normal
## 33 Anderson-Darling     level_4_std   377.842  <0.001      Not normal
## 34 Anderson-Darling    level_4_skew     2.591  <0.001      Not normal
## 35 Anderson-Darling    level_4_kurt   210.083  <0.001      Not normal
## 36 Anderson-Darling level_5_energy  2155.711  <0.001      Not normal
## 37 Anderson-Darling     level_5_std   784.081  <0.001      Not normal
## 38 Anderson-Darling    level_5_skew     2.514  <0.001      Not normal
## 39 Anderson-Darling    level_5_kurt   207.127  <0.001      Not normal
##
## $descriptives
##          Variable    n   Mean Std.Dev Median    Min   Max   25th
   75th   Skew
## 1            mean 9200 -0.023   0.945  0.049 -6.876 7.761 -0.501
   0.506 -0.190
## 2             std 9200 -0.401   0.273 -0.454 -0.726 3.499 -0.543
  -0.336  5.832
## 3             var 9200 -0.336   0.181 -0.364 -0.392 4.023 -0.377
  -0.340 16.092
## 4             min 9200  0.372   0.235  0.411 -2.761 0.807  0.273
   0.518 -3.299
## 5             max 9200 -0.375   0.427 -0.436 -0.994 6.405 -0.556
  -0.289  8.487
## 6            skew 9200  0.022   0.873  0.009 -6.174 7.477 -0.441
   0.474  0.819
## 7        kurtosis 9200 -0.052   0.920 -0.229 -1.304 19.245 -0.499
   0.138  6.699
## 8             rms 9200 -0.398   0.288 -0.452 -0.783 3.605 -0.556
  -0.304  5.350
## 9  zero_crossings 9200  0.040   1.067  0.037 -1.996 4.738 -0.726
   0.799  0.273
## 10        abs_max 9200 -0.381   0.343 -0.442 -0.728 4.937 -0.532
  -0.318  8.198
```

```
## 11    crest_factor 9200 -0.076    0.929 -0.176 -2.682  8.132 -0.684
    0.400   1.139
## 12   margin_factor 9200  0.240    0.977  0.022 -1.086 14.185 -0.334
    0.529   3.671
## 13    shape_factor 9200 -0.051    0.970 -0.099 -2.807 12.052 -0.572
    0.374   1.931
## 14 impulse_factor 9200 -0.073    0.934 -0.178 -2.382  9.416 -0.650
    0.353   1.920
## 15     delta_power 9200  0.035    0.976  0.003 -1.930  2.373 -0.821
    0.843   0.129
## 16     theta_power 9200 -0.195    0.935 -0.381 -1.731  3.151 -0.933
    0.400   0.697
## 17     alpha_power 9200 -0.062    1.033 -0.354 -1.582  3.369 -0.903
    0.661   0.759
## 18      beta_power 9200 -0.099    0.959 -0.410 -1.278  5.212 -0.800
    0.311   1.449
## 19     gamma_power 9200  0.071    1.051 -0.238 -1.157 13.976 -0.628
    0.436   2.603
## 20 level_1_energy 9200 -0.261    0.090 -0.278 -0.299  3.797 -0.291
   -0.252 24.986
## 21     level_1_std 9200 -0.357    0.234 -0.402 -0.667  3.840 -0.521
   -0.247  3.382
## 22    level_1_skew 9200  0.017    0.988  0.008 -4.031  3.968 -0.231
    0.263   0.011
## 23    level_1_kurt 9200 -0.046    1.009 -0.517 -0.733  5.240 -0.620
    0.062   2.370
## 24 level_2_energy 9200 -0.244    0.040 -0.256 -0.269  0.847 -0.265
   -0.239  8.759
## 25     level_2_std 9200 -0.355    0.196 -0.392 -0.626  1.758 -0.505
   -0.259  1.789
## 26    level_2_skew 9200  0.023    0.985  0.025 -5.988  5.492 -0.338
    0.391 -0.136
## 27    level_2_kurt 9200 -0.080    1.005 -0.411 -0.898  8.343 -0.578
   -0.072  3.528
## 28 level_3_energy 9200 -0.277    0.057 -0.293 -0.306  1.051 -0.301
   -0.276  7.618
## 29     level_3_std 9200 -0.369    0.218 -0.425 -0.652  1.884 -0.513
   -0.295  2.413
## 30    level_3_skew 9200 -0.006    0.908  0.000 -5.114  4.622 -0.484
    0.488 -0.155
## 31    level_3_kurt 9200 -0.109    0.914 -0.347 -1.357  7.644 -0.657
    0.119   2.804
## 32 level_4_energy 9200 -0.300    0.087 -0.317 -0.334  3.459 -0.326
   -0.301 23.023
## 33     level_4_std 9200 -0.387    0.198 -0.428 -0.667  3.543 -0.508
   -0.323  4.547
## 34    level_4_skew 9200  0.025    0.964  0.029 -3.227  3.332 -0.602
    0.641   0.025
## 35    level_4_kurt 9200 -0.016    0.957 -0.238 -1.592  4.441 -0.695
    0.446   1.261
## 36 level_5_energy 9200 -0.282    0.391 -0.343 -0.395 12.233 -0.370
   -0.282 20.166
## 37     level_5_std 9200 -0.373    0.378 -0.460 -0.754  6.864 -0.559
   -0.308  6.917
## 38    level_5_skew 9200  0.004    1.022  0.005 -3.472  3.532 -0.668
    0.667   0.013
## 39    level_5_kurt 9200  0.070    1.007 -0.170 -1.667  4.834 -0.645
    0.576   1.232
##    Kurtosis
## 1     5.714
## 2    62.410
```

```
## 3    314.530
## 4     30.079
## 5    115.799
## 6     10.292
## 7     87.555
## 8     55.386
## 9      2.642
## 10   106.967
## 11     6.904
## 12    29.955
## 13    18.383
## 14    13.060
## 15     1.921
## 16     2.763
## 17     2.603
## 18     4.991
## 19    16.516
## 20   934.534
## 21    38.276
## 22     6.169
## 23     8.685
## 24   159.159
## 25    10.426
## 26     9.799
## 27    18.522
## 28   105.546
## 29    12.901
## 30     6.034
## 31    14.684
## 32   848.171
## 33    56.967
## 34     3.236
## 35     4.830
## 36   534.730
## 37    86.986
## 38     3.205
## 39     4.755
##
## $data
## # A tibble: 9,200    39
##       mean     std     var     min     max    skew kurtosis     rms
    zero_crossings
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>   <dbl>
              <dbl>
##  1 -0.281   -0.0473 -0.255  -0.0992  0.0235 -0.300   0.0600 -0.0946
            0.926
##  2 -1.11    -0.459  -0.365   0.377  -0.499   0.902  -0.206  -0.376
           -0.471
##  3 -1.87    -0.684  -0.390   0.442  -0.857   0.682   0.185  -0.309
           -2.00
##  4  0.0327 -0.502  -0.371   0.448  -0.506  -0.754  -0.215  -0.563
            2.45
##  5  0.679   -0.421  -0.358   0.485  -0.352  -0.134  -0.513  -0.468
            0.0369
##  6  0.664   -0.293  -0.330   0.343  -0.285  -0.104  -0.562  -0.344
            1.05
##  7 -0.759   -0.515  -0.373   0.402  -0.643  -1.19   -0.315  -0.481
           -0.853
##  8  0.359   -0.566  -0.380   0.494  -0.528  -0.551   0.245  -0.630
            2.20
##  9 -0.482   -0.532  -0.376   0.331  -0.643  -0.981   0.171  -0.540
```

```
##              1.82
## 10 -1.60    -0.157  -0.292 -0.151  -0.419  -1.54    0.763  -0.0595
##           -0.471
## #      9,190 more rows
## #      30 more variables: abs_max <dbl>, crest_factor <dbl>, margin_
##    factor <dbl>,
## #    shape_factor <dbl>, impulse_factor <dbl>, delta_power <dbl>,
## #    theta_power <dbl>, alpha_power <dbl>, beta_power <dbl>, gamma_
##    power <dbl>,
## #    level_1_energy <dbl>, level_1_std <dbl>, level_1_skew <dbl>,
## #    level_1_kurt <dbl>, level_2_energy <dbl>, level_2_std <dbl>,
## #    level_2_skew <dbl>, level_2_kurt <dbl>, level_3_energy <dbl>,
##
## $subset
## NULL
##
## $outlierMethod
## [1] "none"
##
## attr(,"class")
## [1] "mvn"
```

Kết quả kiểm định chuẩn đa biến cho nhóm Seizure (1)

```
## $multivariate_normality
##            Test Statistic p.value     Method         MVN
## 1 Henze-Zirkler     1.098  <0.001 asymptotic Not normal
##
## $univariate_normality
##                Test        Variable Statistic p.value     Normality
## 1  Anderson-Darling            mean     5.023  <0.001 Not normal
## 2  Anderson-Darling             std    36.447  <0.001 Not normal
## 3  Anderson-Darling             var   109.369  <0.001 Not normal
## 4  Anderson-Darling             min    93.503  <0.001 Not normal
## 5  Anderson-Darling             max    30.761  <0.001 Not normal
## 6  Anderson-Darling            skew     2.515  <0.001 Not normal
## 7  Anderson-Darling        kurtosis    96.975  <0.001 Not normal
## 8  Anderson-Darling             rms    38.038  <0.001 Not normal
## 9  Anderson-Darling zero_crossings    24.699  <0.001 Not normal
## 10 Anderson-Darling         abs_max    54.815  <0.001 Not normal
## 11 Anderson-Darling    crest_factor    17.205  <0.001 Not normal
## 12 Anderson-Darling  margin_factor   116.592  <0.001 Not normal
## 13 Anderson-Darling   shape_factor    28.221  <0.001 Not normal
## 14 Anderson-Darling impulse_factor    28.236  <0.001 Not normal
## 15 Anderson-Darling    delta_power    29.092  <0.001 Not normal
## 16 Anderson-Darling    theta_power    24.090  <0.001 Not normal
## 17 Anderson-Darling    alpha_power    13.629  <0.001 Not normal
## 18 Anderson-Darling     beta_power    89.296  <0.001 Not normal
## 19 Anderson-Darling    gamma_power   107.635  <0.001 Not normal
## 20 Anderson-Darling level_1_energy   235.389  <0.001 Not normal
## 21 Anderson-Darling    level_1_std    77.972  <0.001 Not normal
## 22 Anderson-Darling   level_1_skew    36.790  <0.001 Not normal
## 23 Anderson-Darling   level_1_kurt   172.075  <0.001 Not normal
## 24 Anderson-Darling level_2_energy   272.156  <0.001 Not normal
## 25 Anderson-Darling    level_2_std    88.549  <0.001 Not normal
## 26 Anderson-Darling   level_2_skew     9.542  <0.001 Not normal
## 27 Anderson-Darling   level_2_kurt   102.969  <0.001 Not normal
## 28 Anderson-Darling level_3_energy   218.211  <0.001 Not normal
## 29 Anderson-Darling    level_3_std    65.781  <0.001 Not normal
## 30 Anderson-Darling   level_3_skew     2.081  <0.001 Not normal
## 31 Anderson-Darling   level_3_kurt    90.326  <0.001 Not normal
```

```
## 32 Anderson-Darling level_4_energy   155.090   <0.001      Not normal
## 33 Anderson-Darling     level_4_std    35.105   <0.001      Not normal
## 34 Anderson-Darling    level_4_skew     0.597    0.119          Normal
## 35 Anderson-Darling    level_4_kurt    72.544   <0.001      Not normal
## 36 Anderson-Darling level_5_energy   153.420   <0.001      Not normal
## 37 Anderson-Darling     level_5_std    39.963   <0.001      Not normal
## 38 Anderson-Darling    level_5_skew     4.230   <0.001      Not normal
## 39 Anderson-Darling    level_5_kurt    65.654   <0.001      Not normal
##
## $descriptives
##          Variable    n   Mean Std.Dev Median    Min    Max   25th
   75th   Skew
## 1            mean 2300  0.091   1.190  0.020 -5.100  7.560 -0.687
   0.815  0.343
## 2             std 2300  1.604   1.219  1.397 -0.401  5.640  0.545
   2.549  0.499
## 3             var 2300  1.346   1.614  0.765 -0.354  9.499  0.043
   2.290  1.166
## 4             min 2300 -1.486   1.421 -1.010 -5.029  0.512 -2.286
  -0.413 -0.992
## 5             max 2300  1.501   1.207  1.234 -0.524  6.405  0.557
   2.341  0.672
## 6            skew 2300 -0.089   1.393 -0.126 -6.587  6.970 -1.066
   0.844  0.219
## 7        kurtosis 2300  0.210   1.250 -0.087 -1.408 18.151 -0.554
   0.678  4.619
## 8             rms 2300  1.591   1.226  1.372 -0.405  5.656  0.527
   2.534  0.511
## 9  zero_crossings 2300 -0.160   0.645 -0.217 -1.869  2.578 -0.598
   0.164  0.678
## 10        abs_max 2300  1.526   1.273  1.169 -0.408  4.937  0.487
   2.445  0.707
## 11   crest_factor 2300  0.303   1.198  0.149 -2.167  7.884 -0.591
   1.025  0.916
## 12 margin_factor 2300 -0.961   0.157 -1.020 -1.142 -0.108 -1.070
  -0.892  1.512
## 13   shape_factor 2300  0.202   1.087  0.022 -2.085  8.992 -0.553
   0.749  1.170
## 14 impulse_factor 2300  0.291   1.185  0.092 -1.960 11.373 -0.561
   0.942  1.384
## 15    delta_power 2300 -0.139   1.079 -0.144 -2.024  2.356 -1.114
   0.810  0.029
## 16    theta_power 2300  0.780   0.861  0.946 -1.674  3.126  0.222
   1.350 -0.476
## 17    alpha_power 2300  0.246   0.811  0.147 -1.529  3.118 -0.351
   0.750  0.574
## 18     beta_power 2300  0.395   1.062  0.058 -1.258  5.255 -0.357
   0.872  1.326
## 19    gamma_power 2300 -0.284   0.696 -0.459 -1.131  6.604 -0.733
  -0.064  2.944
## 20 level_1_energy 2300  1.043   1.900  0.272 -0.297 18.365 -0.095
   1.390  2.822
## 21    level_1_std 2300  1.428   1.494  0.974 -0.616  9.096  0.287
   2.213  1.287
## 22   level_1_skew 2300 -0.070   1.043 -0.093 -3.800  3.764 -0.501
   0.415 -0.074
## 23   level_1_kurt 2300  0.184   0.940 -0.169 -0.721  4.788 -0.434
   0.427  1.955
## 24 level_2_energy 2300  0.977   1.950  0.241 -0.268 18.680 -0.097
   1.134  3.342
## 25    level_2_std 2300  1.419   1.527  0.985 -0.564  9.414  0.293
```

```
    2.077   1.499
## 26    level_2_skew 2300 -0.090    1.051 -0.167 -4.726   5.628 -0.725
    0.512   0.330
## 27    level_2_kurt 2300  0.318    0.911  0.075 -0.823   7.632 -0.284
    0.619   2.211
## 28 level_3_energy 2300  1.108    1.858  0.356 -0.306  11.761 -0.089
    1.467   2.170
## 29     level_3_std 2300  1.476    1.445  1.102 -0.613   6.886  0.340
    2.234   1.068
## 30    level_3_skew 2300  0.025    1.305 -0.011 -4.404   4.617 -0.835
    0.831   0.173
## 31    level_3_kurt 2300  0.437    1.193  0.086 -1.194   6.577 -0.380
    0.912   1.733
## 32 level_4_energy 2300  1.200    1.780  0.609 -0.331  12.332 -0.039
    1.685   2.185
## 33     level_4_std 2300  1.548    1.360  1.347 -0.591   6.976  0.426
    2.304   0.895
## 34    level_4_skew 2300 -0.101    1.127 -0.099 -3.424   3.405 -0.812
    0.643  -0.033
## 35    level_4_kurt 2300  0.063    1.156 -0.268 -1.709   4.708 -0.754
    0.587   1.204
## 36 level_5_energy 2300  1.127    1.674  0.514 -0.368  17.748  0.017
    1.607   2.517
## 37     level_5_std 2300  1.494    1.280  1.197 -0.628   7.638  0.499
    2.226   0.953
## 38    level_5_skew 2300 -0.018    0.905  0.013 -3.460   3.729 -0.546
    0.559  -0.159
## 39    level_5_kurt 2300 -0.280    0.919 -0.489 -1.738   5.217 -0.904
    0.100   1.612
##     Kurtosis
## 1      4.395
## 2      2.225
## 3      3.669
## 4      2.950
## 5      2.895
## 6      3.459
## 7     50.266
## 8      2.232
## 9      3.673
## 10     2.484
## 11     4.864
## 12     5.168
## 13     6.158
## 14     8.171
## 15     1.794
## 16     2.801
## 17     3.028
## 18     4.501
## 19    18.506
## 20    13.933
## 21     4.503
## 22     4.623
## 23     6.918
## 24    18.363
## 25     5.450
## 26     4.921
## 27    10.584
## 28     7.903
## 29     3.561
## 30     3.465
## 31     6.708
```

```
## 32     8.826
## 33     3.489
## 34     2.929
## 35     4.125
## 36    13.506
## 37     3.730
## 38     3.674
## 39     6.729
##
## $data
## # A tibble: 2,300    39
##       mean   std      var     min      max     skew kurtosis    rms zero_
    crossings
##      <dbl> <dbl>    <dbl>   <dbl>    <dbl>    <dbl>    <dbl> <dbl>
            <dbl>
##  1   1.10    2.96    2.98    -4.51    1.02    -2.57      1.16    2.94
           -0.980
##  2  -0.732   0.400  -0.0451  -0.465   0.690   -0.0915   -0.0586  0.373
           -0.471
##  3   1.57    1.56    0.946   -0.496   2.16     0.914    -0.471   1.55
           -0.0902
##  4   0.989   3.65    4.33    -2.97    1.90    -1.19     -0.560   3.64
           -0.726
##  5  -1.83    0.299  -0.101   -0.920  -0.0783  -1.98      1.61    0.369
           -1.23
##  6   0.551   0.285  -0.108   -0.225   0.0762  -0.818    -0.494   0.233
           -0.217
##  7   1.63    2.66    2.47    -2.35    3.30     0.266     0.237   2.65
            0.672
##  8  -1.34    1.00    0.385   -0.837   0.539   -0.351    -0.886   1.00
           -0.471
##  9   0.0997  0.590   0.0725  -0.502   1.48     2.20      2.30    0.539
           -0.344
## 10   0.114   3.45    3.91    -3.55    3.35    -0.915     0.157   3.43
            1.18
## # i 2,290 more rows
## # i 30 more variables: abs_max <dbl>, crest_factor <dbl>, margin_
    factor <dbl>,
## #   shape_factor <dbl>, impulse_factor <dbl>, delta_power <dbl>,
## #   theta_power <dbl>, alpha_power <dbl>, beta_power <dbl>, gamma_
    power <dbl>,
## #   level_1_energy <dbl>, level_1_std <dbl>, level_1_skew <dbl>,
## #   level_1_kurt <dbl>, level_2_energy <dbl>, level_2_std <dbl>,
## #   level_2_skew <dbl>, level_2_kurt <dbl>, level_3_energy <dbl>,
##
## $subset
## NULL
##
## $outlierMethod
## [1] "none"
##
## attr(,"class")
## [1] "mvn"
```

Nhận xét: Kết quả từ kiểm định Henze-Zirkler cho thấy p-value của cả hai nhóm đều rất nhỏ ($p < 0.001$), do đó chúng ta bác bỏ giả thuyết H_0 và kết luận rằng dữ liệu *không tuân theo phân phối chuẩn đa biến*.

2. **Thực hiện kiểm định MANOVA**

Các giả thuyết thống kê cho phép kiểm định này là:

- H_0: Không có sự khác biệt có ý nghĩa thống kê về vector trung bình giữa hai nhóm

$$\mu_{\text{co\_giật}} = \mu_{\text{không\_co\_giật}}$$

- H_a: Có ít nhất một khác biệt có ý nghĩa thống kê giữa vector trung bình hai nhóm

$$\mu_{\text{co\_giật}} \neq \mu_{\text{không\_co\_giật}}$$

```
##                      Df   Wilks approx F num Df den Df    Pr(>F)
## y_labels_factor       1 0.19115   1243.4     39  11460 < 2.2e-16
   ***
## Residuals         11498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kết luận: Mặc dù cả hai giả định chính của MANOVA đều bị vi phạm, kết quả từ chỉ số **Pillai's Trace** vẫn cho p-value cực kỳ nhỏ (p < 2.2e-16). Điều này cho thấy **sự khác biệt có ý nghĩa thống kê rất rõ ràng giữa hai nhóm**.

Phát hiện này khẳng định rằng bộ đặc trưng đã trích xuất **có khả năng phân biệt mạnh mẽ giữa nhóm co giật và không co giật**, cho thấy tiềm năng sử dụng trong phân tích hoặc mô hình phân loại sau này.

5 Phương pháp Trực quan Phi tuyến

Sau khi khám phá dữ liệu bằng các phương pháp tuyến tính như PCA và Phân tích Nhân tố (FA), chúng ta nhận thấy rằng mặc dù có sự khác biệt giữa hai lớp, vẫn tồn tại vùng chồng lấn đáng kể. Điều này gợi ý rằng ranh giới phân tách giữa hai trạng thái "co giật" và "không co giật" có thể không đơn thuần là tuyến tính.

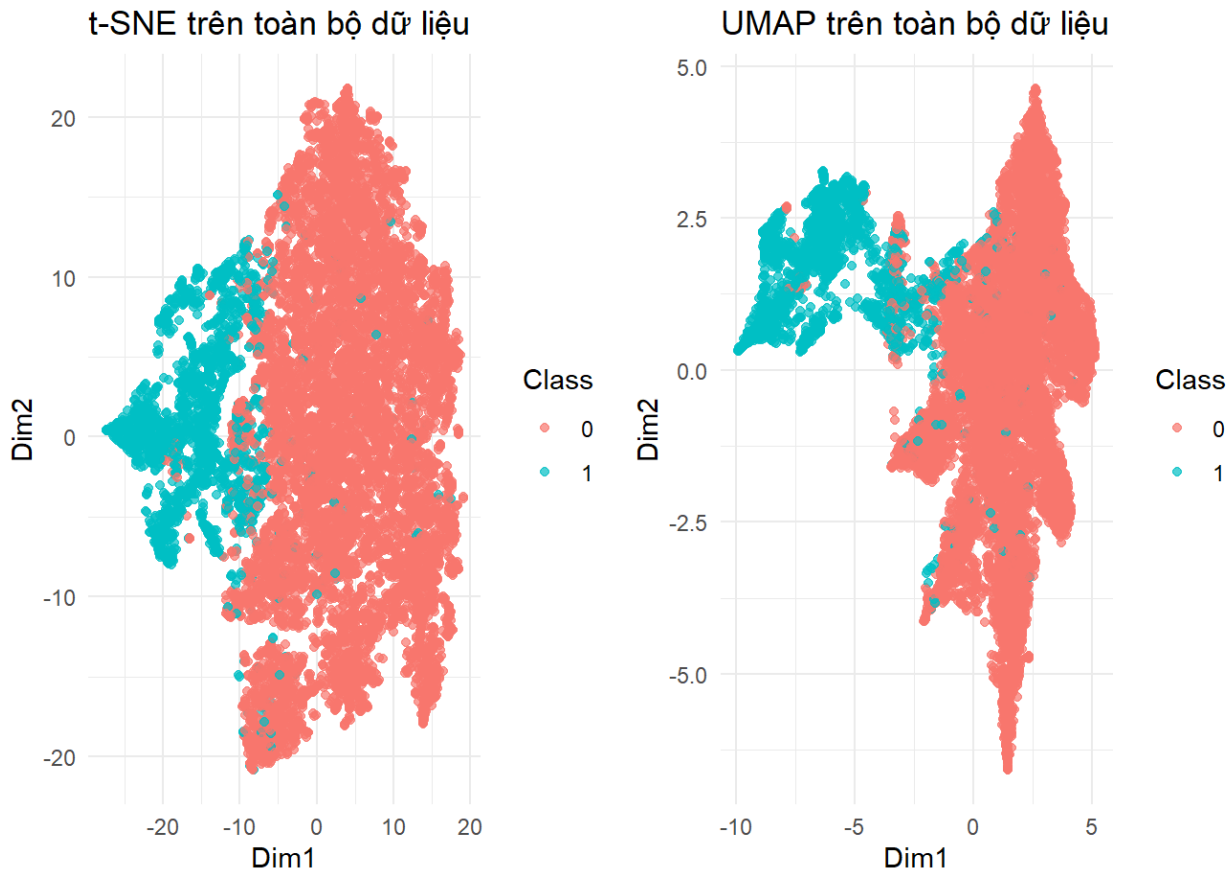
Để có được cái nhìn rõ nét hơn về cấu trúc phân cụm trong dữ liệu, chúng ta sử dụng hai phương pháp giảm chiều phi tuyến phổ biến: **t-SNE** và **UMAP**. Cả hai đều nổi bật trong việc ánh xạ dữ liệu từ không gian nhiều chiều xuống không gian 2D, đồng thời bảo toàn cấu trúc "lân cận" giữa các điểm dữ liệu – tức là đưa các điểm tương tự nhau lại gần nhau trên biểu đồ, từ đó giúp làm nổi bật các cụm tiềm ẩn trong dữ liệu.

5.1 Trực quan hóa bằng t-SNE và UMAP

t-SNE và UMAP là các kỹ thuật mạnh mẽ để trực quan hóa dữ liệu nhiều chiều trong không gian 2D.

Chạy t-SNE trên toàn bộ 11,500 điểm dữ liệu.

Chạy UMAP trên toàn bộ 11,500 điểm dữ liệu.



Hình 21: Trực quan hóa

Nhận xét: Cả hai biểu đồ đều thể hiện sự phân tách rõ ràng giữa hai nhóm dữ liệu. Các điểm thuộc nhóm "Seizure" và "Non-Seizure" hình thành nên hai cụm gần như riêng biệt, với

rất ít điểm bị chồng lấn. Trong đó, UMAP cho thấy khả năng bảo toàn cấu trúc toàn cục tốt hơn so với t-SNE. Kết quả này củng cố giả thuyết rằng bộ đặc trưng hiện tại có *khả năng phân biệt mạnh mẽ* giữa hai trạng thái.

5.2 So sánh với PCA và FA

Việc so sánh các phương pháp giảm chiều như PCA, FA, t-SNE và UMAP cho thấy mỗi kỹ thuật phản ánh một khía cạnh khác nhau của cấu trúc dữ liệu khi chiếu xuống không gian thấp chiều.

- **PCA (Phân tích Thành phần Chính):**
  - Giữ lại phần lớn phương sai của dữ liệu (56.7% trong hai thành phần đầu).
  - Dễ nhận thấy sự khác biệt giữa hai lớp theo trục thành phần thứ nhất.
  - Tuy nhiên, vẫn có sự chồng lấn đáng kể giữa hai lớp, do đây là phương pháp tuyến tính.
- **FA (Phân tích Nhân tố):**
  - Hữu ích trong việc phát hiện các cấu trúc nhân tố tiềm ẩn.
  - Tuy nhiên, trên biểu đồ 2D, khả năng phân tách giữa các lớp không cải thiện nhiều so với PCA.
- **t-SNE và UMAP:**
  - Là các kỹ thuật phi tuyến, không cố gắng giữ lại phương sai toàn cục mà tập trung vào việc bảo toàn cấu trúc lân cận.
  - Cả hai đều cho thấy sự phân cụm rõ ràng và gần như tách biệt hoàn toàn giữa hai lớp trong không gian 2D.

Kết luận:
Kết quả từ t-SNE và UMAP cho thấy cấu trúc phân lớp trong dữ liệu **không hoàn toàn tuyến tính** – điều mà các phương pháp như PCA hay FA không thể hiện rõ. Tuy nhiên, **các đặc trưng được giữ lại qua PCA vẫn chứa đủ thông tin** để mô hình học máy có thể khai thác được ranh giới phân loại, kể cả khi nó phi tuyến.

6 Mô hình Phân loại

Trong phần này, chúng ta sẽ xây dựng và so sánh hiệu năng của các mô hình học máy dự trên data với bộ đặc trưng đây dủ và đã qua xử lý pca.

6.1 Chuẩn bị dữ liệu huấn luyện

Phân bố lớp trên tập huấn luyện

```
## y_train
## NonSeizure    Seizure
##        0.8        0.2
```

Nhận xét: Kết quả phân tích cho thấy lớp '0' (Non-Seizure) chiếm khoảng 80% tổng số quan sát, trong khi lớp '1' (Seizure) chỉ chiếm khoảng 20%. Đây là một tình trạng mất cân bằng dữ liệu nghiêm trọng, có thể dẫn đến việc mô hình bị thiên lệch trong quá trình huấn luyện: dễ dự đoán chính xác lớp đa số nhưng kém hiệu quả khi nhận diện các trường hợp Seizure – vốn là mục tiêu chính trong bối cảnh lâm sàng.

Để khắc phục vấn đề này, chúng tôi đã áp dụng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) nhằm tăng cường mẫu cho lớp thiểu số bằng cách tạo ra các điểm dữ liệu tổng hợp từ các trường hợp hiện có. Việc này giúp cân bằng phân bố giữa hai lớp, góp phần cải thiện khả năng học của mô hình và nâng cao độ nhạy trong việc phát hiện các trường hợp co giật.

6.2 Huấn luyện với kiểm định chéo

```
## Preparing recipe
## + Fold01: parameter=none
## - Fold01: parameter=none
## + Fold02: parameter=none
## - Fold02: parameter=none
## + Fold03: parameter=none
## - Fold03: parameter=none
## + Fold04: parameter=none
## - Fold04: parameter=none
## + Fold05: parameter=none
## - Fold05: parameter=none
## + Fold06: parameter=none
## - Fold06: parameter=none
## + Fold07: parameter=none
## - Fold07: parameter=none
## + Fold08: parameter=none
## - Fold08: parameter=none
## + Fold09: parameter=none
## - Fold09: parameter=none
## + Fold10: parameter=none
## - Fold10: parameter=none
## Aggregating results
## Fitting final model on full training set
```

```
## Preparing recipe
```

```
## + Fold01: sigma=0.02561, C=0.25
## - Fold01: sigma=0.02561, C=0.25
## + Fold01: sigma=0.02561, C=0.50
## - Fold01: sigma=0.02561, C=0.50
## + Fold01: sigma=0.02561, C=1.00
## - Fold01: sigma=0.02561, C=1.00
## + Fold02: sigma=0.02561, C=0.25
## - Fold02: sigma=0.02561, C=0.25
## + Fold02: sigma=0.02561, C=0.50
## - Fold02: sigma=0.02561, C=0.50
## + Fold02: sigma=0.02561, C=1.00
## - Fold02: sigma=0.02561, C=1.00
## + Fold03: sigma=0.02561, C=0.25
## - Fold03: sigma=0.02561, C=0.25
## + Fold03: sigma=0.02561, C=0.50
## - Fold03: sigma=0.02561, C=0.50
## + Fold03: sigma=0.02561, C=1.00
## - Fold03: sigma=0.02561, C=1.00
## + Fold04: sigma=0.02561, C=0.25
## - Fold04: sigma=0.02561, C=0.25
## + Fold04: sigma=0.02561, C=0.50
## - Fold04: sigma=0.02561, C=0.50
## + Fold04: sigma=0.02561, C=1.00
## - Fold04: sigma=0.02561, C=1.00
## + Fold05: sigma=0.02561, C=0.25
## - Fold05: sigma=0.02561, C=0.25
## + Fold05: sigma=0.02561, C=0.50
## - Fold05: sigma=0.02561, C=0.50
## + Fold05: sigma=0.02561, C=1.00
## - Fold05: sigma=0.02561, C=1.00
## + Fold06: sigma=0.02561, C=0.25
## - Fold06: sigma=0.02561, C=0.25
## + Fold06: sigma=0.02561, C=0.50
## - Fold06: sigma=0.02561, C=0.50
## + Fold06: sigma=0.02561, C=1.00
## - Fold06: sigma=0.02561, C=1.00
## + Fold07: sigma=0.02561, C=0.25
## - Fold07: sigma=0.02561, C=0.25
## + Fold07: sigma=0.02561, C=0.50
## - Fold07: sigma=0.02561, C=0.50
## + Fold07: sigma=0.02561, C=1.00
## - Fold07: sigma=0.02561, C=1.00
## + Fold08: sigma=0.02561, C=0.25
## - Fold08: sigma=0.02561, C=0.25
## + Fold08: sigma=0.02561, C=0.50
## - Fold08: sigma=0.02561, C=0.50
## + Fold08: sigma=0.02561, C=1.00
## - Fold08: sigma=0.02561, C=1.00
## + Fold09: sigma=0.02561, C=0.25
## - Fold09: sigma=0.02561, C=0.25
## + Fold09: sigma=0.02561, C=0.50
## - Fold09: sigma=0.02561, C=0.50
## + Fold09: sigma=0.02561, C=1.00
## - Fold09: sigma=0.02561, C=1.00
## + Fold10: sigma=0.02561, C=0.25
```

```
## - Fold10: sigma=0.02561, C=0.25
## + Fold10: sigma=0.02561, C=0.50
## - Fold10: sigma=0.02561, C=0.50
## + Fold10: sigma=0.02561, C=1.00
## - Fold10: sigma=0.02561, C=1.00
## Aggregating results
## Selecting tuning parameters
## Fitting sigma = 0.0256, C = 1 on full training set
## Preparing recipe
## + Fold01: mtry= 2
## - Fold01: mtry= 2
## + Fold01: mtry=20
## - Fold01: mtry=20
## + Fold01: mtry=39
## - Fold01: mtry=39
## + Fold02: mtry= 2
## - Fold02: mtry= 2
## + Fold02: mtry=20
## - Fold02: mtry=20
## + Fold02: mtry=39
## - Fold02: mtry=39
## + Fold03: mtry= 2
## - Fold03: mtry= 2
## + Fold03: mtry=20
## - Fold03: mtry=20
## + Fold03: mtry=39
## - Fold03: mtry=39
## + Fold04: mtry= 2
## - Fold04: mtry= 2
## + Fold04: mtry=20
## - Fold04: mtry=20
## + Fold04: mtry=39
## - Fold04: mtry=39
## + Fold05: mtry= 2
## - Fold05: mtry= 2
## + Fold05: mtry=20
## - Fold05: mtry=20
## + Fold05: mtry=39
## - Fold05: mtry=39
## + Fold06: mtry= 2
## - Fold06: mtry= 2
## + Fold06: mtry=20
## - Fold06: mtry=20
## + Fold06: mtry=39
## - Fold06: mtry=39
## + Fold07: mtry= 2
## - Fold07: mtry= 2
## + Fold07: mtry=20
## - Fold07: mtry=20
## + Fold07: mtry=39
## - Fold07: mtry=39
## + Fold08: mtry= 2
## - Fold08: mtry= 2
## + Fold08: mtry=20
## - Fold08: mtry=20
```

```
## + Fold08: mtry=39
## - Fold08: mtry=39
## + Fold09: mtry= 2
## - Fold09: mtry= 2
## + Fold09: mtry=20
## - Fold09: mtry=20
## + Fold09: mtry=39
## - Fold09: mtry=39
## + Fold10: mtry= 2
## - Fold10: mtry= 2
## + Fold10: mtry=20
## - Fold10: mtry=20
## + Fold10: mtry=39
## - Fold10: mtry=39
## Aggregating results
## Selecting tuning parameters
## Fitting mtry = 2 on full training set
```

So sánh kết quả từ kiểm định chéo

```
## Call:
## summary.resamples(object = algorithm_comparison)
##
## Models: LDA, SVM, RF
## Number of resamples: 10
##
## ROC
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA'
   s
## LDA 0.9848845 0.9892910 0.9932250 0.9923079 0.9955012 0.9973121
   0
## SVM 0.9966919 0.9971386 0.9983939 0.9980388 0.9986007 0.9991804
   0
## RF  0.9946612 0.9972328 0.9978752 0.9976249 0.9989957 0.9991397
   0
##
## Sens
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA'
   s
## LDA 0.9701087 0.9758832 0.9782609 0.9792120 0.9816576 0.9904891
   0
## SVM 0.9836957 0.9864130 0.9891304 0.9883152 0.9904891 0.9918478
   0
## RF  0.9782609 0.9813179 0.9850543 0.9845109 0.9874321 0.9918478
   0
##
## Spec
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA'
   s
## LDA 0.9130435 0.9211957 0.9402174 0.9358696 0.9442935 0.9619565
   0
## SVM 0.9402174 0.9551630 0.9728261 0.9679348 0.9823370 0.9891304
   0
## RF  0.9456522 0.9687500 0.9755435 0.9722826 0.9782609 0.9891304
   0
```

Phân tích Hiệu năng và Lựa chọn Mô hình

Dựa trên kết quả đánh giá 10-fold cross-validation trên dữ liệu sau PCA, ba thuật toán LDA, SVM và RF cho thấy sự khác biệt rõ ràng về hiệu năng:

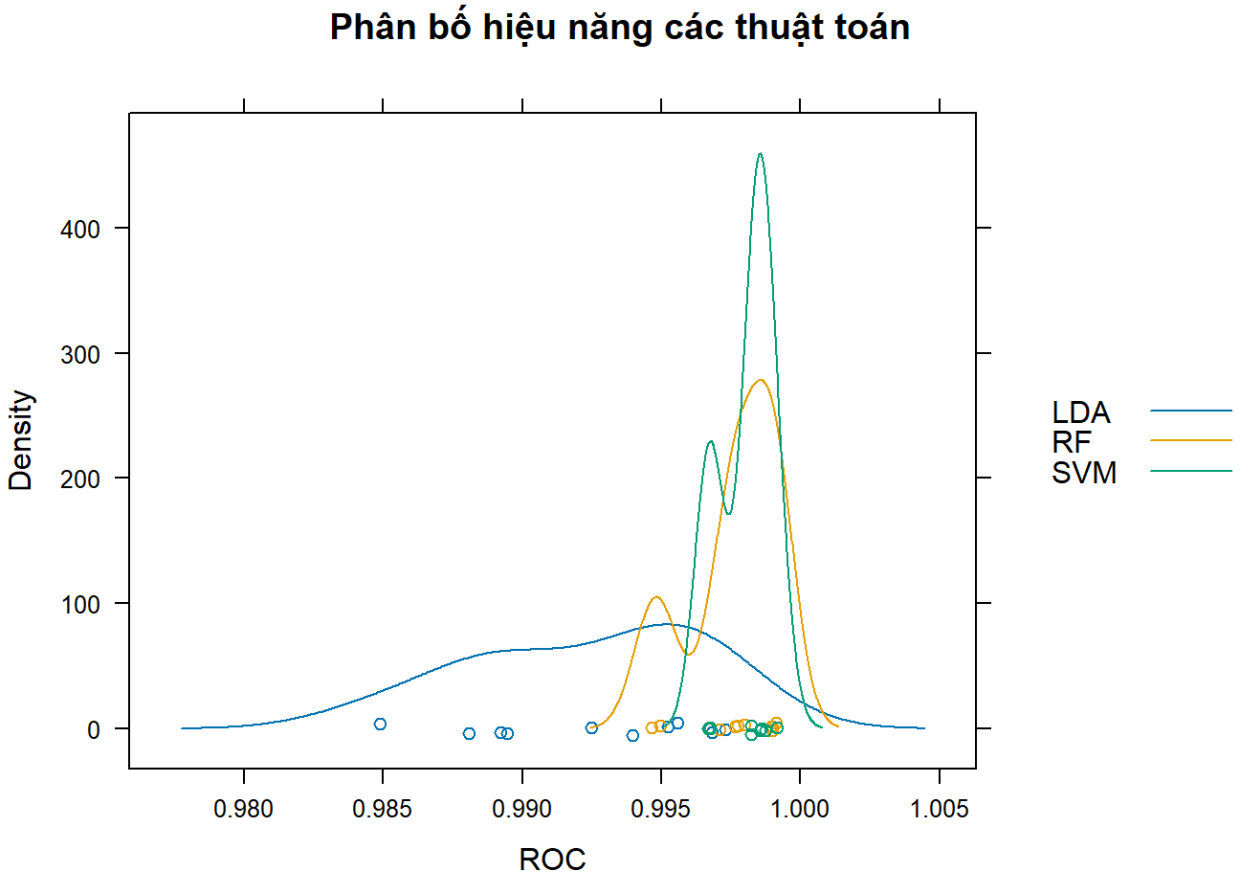
Mô hình	ROC (TB)	Sensitivity	Specificity
LDA	0.9923	0.9792	0.9359
SVM	**0.9980**	**0.9883**	0.9679
RF	0.9976	0.9845	**0.9723**

Hình 22: So sánh

Nhận xét chính

- **SVM** có **ROC** và **Sensitivity cao nhất**, cho thấy khả năng phân loại tốt và đặc biệt hiệu quả trong việc **phát hiện đúng các ca bệnh thực sự** (giảm nguy cơ bỏ sót bệnh nhân – False Negative).
- **RF** đạt **Specificity cao nhất**, tức mô hình này **giỏi hơn trong việc loại trừ các ca không mắc bệnh**, giúp giảm báo động giả (False Positive), dù ROC và Sensitivity kém hơn SVM một chút.
- **LDA** có hiệu năng thấp hơn rõ ở cả ba chỉ số, nên không phù hợp cho ứng dụng lâm sàng.

Để có cái nhìn trực quan hơn về sự khác biệt và độ ổn định của các mô hình, chúng ta có thể vẽ biểu đồ dénityplot và dotplot so sánh chỉ số ROC-AUC thu được.

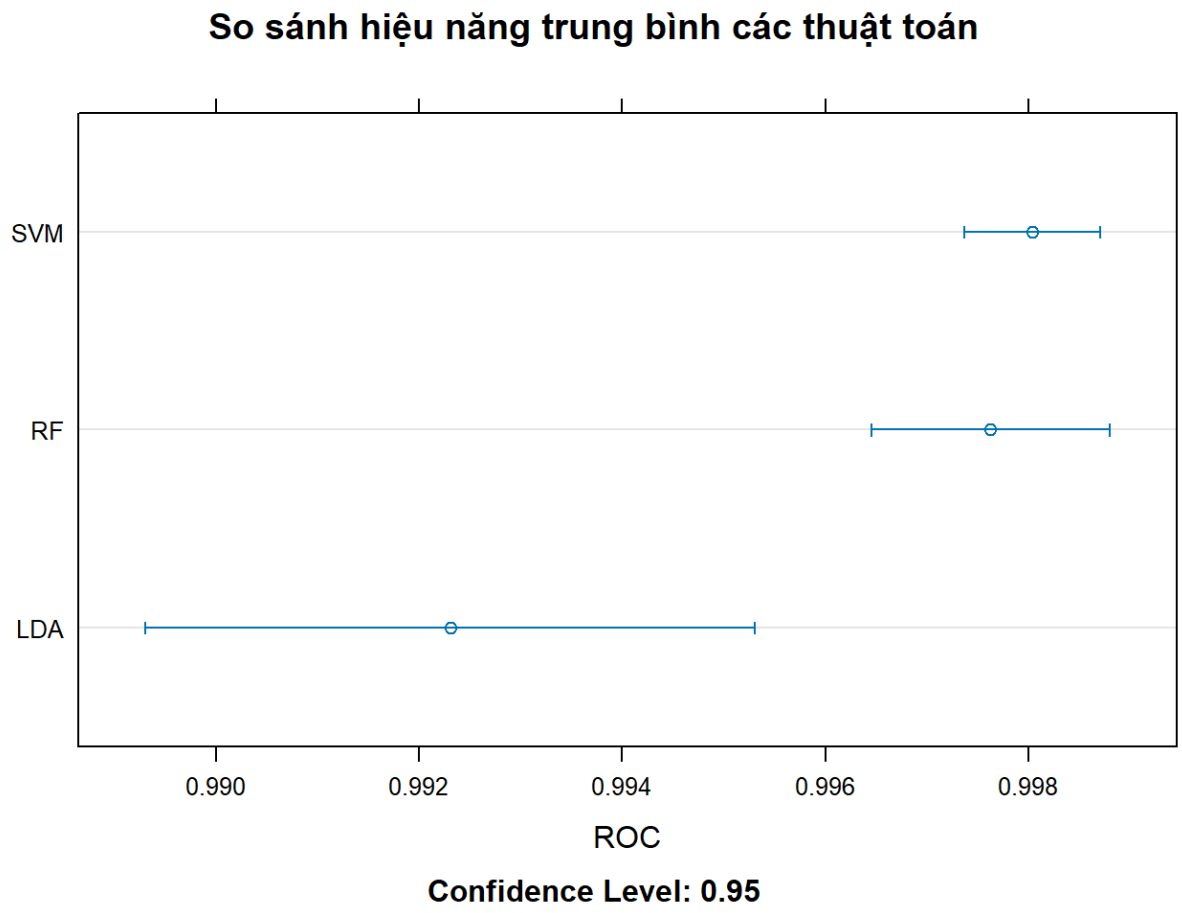


Hình 23: Phân bố hiệu năng các thuật toán

Nhận xét:

- **SVM (xanh lá):** Phân bố hẹp, đỉnh cao, tập trung quanh 0.998–0.999 → hiệu năng cao và ổn định nhất.
- **RF (vàng):** Phân bố khá tập trung, lệch nhẹ về trái so với SVM → hiệu năng tốt, độ ổn định kém hơn SVM.
- **LDA (xanh dương):** Phân bố rộng, thấp hơn rõ → hiệu năng thấp và không ổn định.

Biểu đồ hiển thị giá trị ROC trung bình kèm khoảng tin cậy 95% cho từng thuật toán.



Hình 24: So sánh hiệu năng trung bình các thuật toán

Nhận xét:

- **SVM:** ROC trung bình cao nhất, khoảng tin cậy hẹp và gần như không chồng lấn với LDA, phần lớn tách biệt với RF → cho thấy khác biệt có ý nghĩa thống kê.
- **RF:** ROC trung bình khá cao, sát với SVM nhưng khoảng tin cậy rộng hơn → hiệu năng tốt nhưng biến động lớn hơn.
- **LDA:** ROC trung bình thấp nhất, khoảng tin cậy dài và không giao với hai thuật toán còn lại → hiệu năng kém rõ rệt.

Kết luận: Từ các kết quả trên, SVM là mô hình có hiệu năng ổn định và vượt trội nhất, nên ta sẽ chọn SVM để đánh giá trên dữ liệu đã qua PCA.

6.3 Huấn luyện trên dữ liệu đã qua PCA

Sau khi xác định SVM là mô hình tốt nhất trên bộ đặc trưng đầy đủ, chúng ta sẽ kiểm tra xem việc giảm chiều bằng PCA có phải là một sự đánh đổi hiệu quả hay không. Chúng ta sẽ so sánh hiệu năng của mô hình SVM trên bộ đặc trưng đầy đủ so với các phiên bản SVM được huấn luyện trên dữ liệu đã giảm chiều.

```
## --- Dang huan luyen SVM voi 12 thanh phan chinh... ---
## Preparing recipe
## + Fold01: sigma=0.02561, C=1
## - Fold01: sigma=0.02561, C=1
## + Fold02: sigma=0.02561, C=1
## - Fold02: sigma=0.02561, C=1
## + Fold03: sigma=0.02561, C=1
## - Fold03: sigma=0.02561, C=1
## + Fold04: sigma=0.02561, C=1
## - Fold04: sigma=0.02561, C=1
## + Fold05: sigma=0.02561, C=1
## - Fold05: sigma=0.02561, C=1
## + Fold06: sigma=0.02561, C=1
## - Fold06: sigma=0.02561, C=1
## + Fold07: sigma=0.02561, C=1
## - Fold07: sigma=0.02561, C=1
## + Fold08: sigma=0.02561, C=1
## - Fold08: sigma=0.02561, C=1
## + Fold09: sigma=0.02561, C=1
## - Fold09: sigma=0.02561, C=1
## + Fold10: sigma=0.02561, C=1
## - Fold10: sigma=0.02561, C=1
## Aggregating results
## Fitting final model on full training set
##
## --- Dang huan luyen SVM voi 9 thanh phan chinh... ---
## Preparing recipe
## + Fold01: sigma=0.02561, C=1
## - Fold01: sigma=0.02561, C=1
## + Fold02: sigma=0.02561, C=1
## - Fold02: sigma=0.02561, C=1
## + Fold03: sigma=0.02561, C=1
## - Fold03: sigma=0.02561, C=1
## + Fold04: sigma=0.02561, C=1
## - Fold04: sigma=0.02561, C=1
## + Fold05: sigma=0.02561, C=1
## - Fold05: sigma=0.02561, C=1
## + Fold06: sigma=0.02561, C=1
## - Fold06: sigma=0.02561, C=1
## + Fold07: sigma=0.02561, C=1
## - Fold07: sigma=0.02561, C=1
## + Fold08: sigma=0.02561, C=1
## - Fold08: sigma=0.02561, C=1
## + Fold09: sigma=0.02561, C=1
## - Fold09: sigma=0.02561, C=1
## + Fold10: sigma=0.02561, C=1
## - Fold10: sigma=0.02561, C=1
## Aggregating results
## Fitting final model on full training set
##
## --- Dang huan luyen SVM voi 4 thanh phan chinh... ---
## Preparing recipe
## + Fold01: sigma=0.02561, C=1
## - Fold01: sigma=0.02561, C=1
## + Fold02: sigma=0.02561, C=1
```

```
## - Fold02: sigma=0.02561, C=1
## + Fold03: sigma=0.02561, C=1
## - Fold03: sigma=0.02561, C=1
## + Fold04: sigma=0.02561, C=1
## - Fold04: sigma=0.02561, C=1
## + Fold05: sigma=0.02561, C=1
## - Fold05: sigma=0.02561, C=1
## + Fold06: sigma=0.02561, C=1
## - Fold06: sigma=0.02561, C=1
## + Fold07: sigma=0.02561, C=1
## - Fold07: sigma=0.02561, C=1
## + Fold08: sigma=0.02561, C=1
## - Fold08: sigma=0.02561, C=1
## + Fold09: sigma=0.02561, C=1
## - Fold09: sigma=0.02561, C=1
## + Fold10: sigma=0.02561, C=1
## - Fold10: sigma=0.02561, C=1
## Aggregating results
## Fitting final model on full training set
```

So sánh kết quả của các phiên bản SVM

```
## Call:
## summary.resamples(object = svm_pca_comparison)
##
## Models: SVM_Full, SVM_12PCA, SVM_9PCA, SVM_4PCA
## Number of resamples: 10
##
## ROC
##                Min.   1st Qu.    Median      Mean   3rd Qu.
   Max. NA's
## SVM_Full  0.9966919 0.9971386 0.9983939 0.9980388 0.9986007
   0.9991804    0
## SVM_12PCA 0.9894480 0.9940299 0.9958242 0.9951131 0.9972512
   0.9977404    0
## SVM_9PCA  0.9911168 0.9924496 0.9948975 0.9945748 0.9968229
   0.9974155    0
## SVM_4PCA  0.9857042 0.9909580 0.9937825 0.9925862 0.9950175
   0.9972014    0
##
## Sens
##                Min.   1st Qu.    Median      Mean   3rd Qu.
   Max. NA's
## SVM_Full  0.9836957 0.9864130 0.9891304 0.9883152 0.9904891
   0.9918478    0
## SVM_12PCA 0.9714674 0.9769022 0.9796196 0.9794837 0.9806386
   0.9891304    0
## SVM_9PCA  0.9660326 0.9714674 0.9735054 0.9750000 0.9809783
   0.9823370    0
## SVM_4PCA  0.9605978 0.9646739 0.9687500 0.9694293 0.9748641
   0.9823370    0
##
## Spec
##                Min.   1st Qu.    Median      Mean   3rd Qu.
   Max. NA's
## SVM_Full  0.9402174 0.9551630 0.9728261 0.9679348 0.9823370
```

```
   0.9891304     0
## SVM_12PCA 0.9293478 0.9524457 0.9646739 0.9614130 0.9714674
   0.9782609     0
## SVM_9PCA  0.9402174 0.9524457 0.9592391 0.9570652 0.9619565
   0.9782609     0
## SVM_4PCA  0.9347826 0.9510870 0.9565217 0.9559783 0.9605978
   0.9782609     0
```

Mô hình	ROC (TB)	Sensitivity	Specificity
SVM_Full	**0.9980**	**0.9883**	**0.9679**
SVM_12PCA	0.9951	0.9795	0.9614
SVM_9PCA	0.9946	0.9750	0.9571
SVM_4PCA	0.9926	0.9694	0.9560

Hình 25: So sánh hiệu năng SVM với và không dùng PCA

Nhận xét:

- **SVM_Full** cho hiệu năng cao nhất, cho thấy việc giữ nguyên toàn bộ đặc trưng mang lại kết quả tối ưu.
- Khi số chiều giảm dần qua PCA, hiệu năng có xu hướng giảm theo, đặc biệt rõ ở chỉ số ROC và Sensitivity.
- Tuy nhiên, các phiên bản như **SVM_12PCA** hoặc **SVM_9PCA vẫn duy trì hiệu năng ở mức cao**, với ROC trên 0.994 và Sensitivity trên 0.97 → cho thấy kết quả **vẫn đủ mạnh để được xem là chấp nhận được trong nhiều tình huống ứng dụng thực tế**.

Tác động của PCA đến hiệu năng và tính chấp nhận được

- PCA giúp giảm số chiều đầu vào, từ đó rút ngắn thời gian huấn luyện, giảm tài nguyên tính toán, và hạn chế quá khớp khi số đặc trưng ban đầu lớn.
- Dù hiệu năng có giảm nhẹ, các mô hình PCA vẫn cho kết quả tương đối ổn định và có thể đủ tốt cho các hệ thống triển khai thực tế cần tốc độ và hiệu quả cao hơn.
- Như vậy, PCA là một sự đánh đổi hợp lý giữa độ chính xác và hiệu suất tính toán, nhất là khi mô hình được tích hợp vào các hệ thống thời gian thực hoặc tài nguyên hạn chế.

Kết luận:

- Nếu mục tiêu ưu tiên là độ chính xác tuyệt đối trong môi trường yêu cầu cao (như y tế), **SVM_Full** vẫn là lựa chọn tốt nhất.
- Tuy nhiên, các phiên bản **SVM_PCA** với **9–12 thành phần chính** hoàn toàn có thể **được chấp nhận** trong các tình huống thực tế, khi hiệu năng chỉ giảm nhẹ nhưng mô hình trở nên nhẹ hơn, nhanh hơn và dễ triển khai hơn.

⇒ Do đó, PCA là lựa chọn khả thi trong thực hành nếu cần cân bằng giữa độ chính xác và hiệu suất tính toán.

6.4 Đánh giá trên tập kiểm tra

Dựa trên các phân tích trước đó, hai mô hình tiềm năng nhất được lựa chọn là **SVM_Full** (có hiệu năng cao nhất) và **SVM_12PCA** (đạt được sự cân bằng hợp lý giữa hiệu năng và độ phức tạp mô hình). Trong phần này, chúng tôi sẽ tiến hành đánh giá cả hai mô hình trên tập dữ liệu kiểm tra nhằm đưa ra quyết định lựa chọn cuối cùng.

Đánh giá model **SVM_full**:

Kết quả của mô hình **SVM_Full** trên tập Test

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    NonSeizure Seizure
##   NonSeizure        1827      13
##   Seizure             13     447
##
##                 Accuracy : 0.9887
##                   95% CI : (0.9835, 0.9926)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : <2e-16
##
##                    Kappa : 0.9647
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9717
##             Specificity : 0.9929
##          Pos Pred Value : 0.9717
##          Neg Pred Value : 0.9929
##              Prevalence : 0.2000
##          Detection Rate : 0.1943
##    Detection Prevalence : 0.2000
##       Balanced Accuracy : 0.9823
##
##        'Positive' Class : Seizure
```

Nhận xét:

- **Độ chính xác tổng thể (Accuracy):** Đạt 98.87% với khoảng tin cậy 95% là (98.35%, 99.26%) → cho thấy mô hình hoạt động ổn định và đáng tin cậy.
- **Độ nhạy (Sensitivity – phát hiện Seizure):** 97.17% → mô hình phát hiện đúng phần lớn các trường hợp thực sự bị động kinh, giảm thiểu nguy cơ bỏ sót bệnh nhân.
- **Độ đặc hiệu (Specificity – nhận diện NonSeizure):** 99.29% → mô hình rất chính xác trong việc xác định người không bị bệnh, giúp hạn chế báo động giả.
- **Chỉ số Kappa:** 0.9647 → phản ánh mức độ đồng thuận rất cao giữa dự đoán và thực tế, vượt xa mức ngẫu nhiên.
- **Balanced Accuracy:** 98.23% → trung bình giữa Sensitivity và Specificity, phù hợp khi dữ liệu mất cân bằng.
- **McNemar's Test p-value = 1:** Cho thấy không có sự bất đối xứng đáng kể giữa hai loại lỗi (False Positive và False Negative) → dự đoán hai chiều cân bằng.

Dánh giá mô hình **SVM_12PCA**:
Kết quả của mô hình **SVM_12PCA** trên tập Test

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    NonSeizure Seizure
##   NonSeizure        1812      21
##   Seizure             28     439
##
##                 Accuracy : 0.9787
##                   95% CI : (0.9719, 0.9842)
##     No Information Rate : 0.8
##     P-Value [Acc > NIR] : <2e-16
##
##                    Kappa : 0.9338
##
##  Mcnemar's Test P-Value : 0.3914
##
##             Sensitivity : 0.9543
##             Specificity : 0.9848
##          Pos Pred Value : 0.9400
##          Neg Pred Value : 0.9885
##              Prevalence : 0.2000
##          Detection Rate : 0.1909
##    Detection Prevalence : 0.2030
##       Balanced Accuracy : 0.9696
##
##        'Positive' Class : Seizure
```

Nhận xét:

- **Accuracy:** Mô hình đạt độ chính xác 97.87% (CI 95%: 97.19% – 98.42%) → vẫn ở mức rất cao và đáng tin cậy.
- **Sensitivity (phát hiện Seizure):** 95.43% → mô hình phát hiện tốt phần lớn ca bệnh, tuy có giảm nhẹ so với **SVM_Full** (97.17%).
- **Specificity (phân biệt NonSeizure):** 98.48% → vẫn giữ độ đặc hiệu rất cao, chỉ giảm khoảng 0.8% so với **SVM_Full**.
- **Kappa = 0.9338:** Thể hiện độ phù hợp cao giữa dự đoán và thực tế, cho thấy mô hình đáng tin cậy.
- **Balanced Accuracy = 96.96%:** Vẫn nằm trong ngưỡng cao, phù hợp với dữ liệu không cân bằng.
- **McNemar's test p-value = 0.3914:** Không có sự mất cân bằng đáng kể giữa hai loại lỗi → mô hình dự đoán ổn định theo cả hai hướng.

Chỉ số	SVM_Full	SVM_12PCA
Accuracy	98.87%	97.87%
Sensitivity	97.17%	95.43%
Specificity	99.29%	98.48%
Kappa	0.9647	0.9338

Hình 26: So sánh nhanh với SVM$_Full$

⇒ Hiệu năng giảm nhẹ, nhưng vẫn nằm trong ngưỡng chấp nhận được, đặc biệt là khi xét đến lợi ích tính toán từ việc giảm chiều (bộ dữ liệu nhỏ gọn hơn, huấn luyện nhanh hơn).

ChiSo	SVM_Full	SVM_12PCA
Accuracy	0.9887	0.9787
Sensitivity	0.9717	0.9543
Specificity	0.9929	0.9848
F1	0.9717	0.9471

Hình 27: So sánh hiệu năng trên tập Test

So sánh hiệu năng trên tập kiểm tra

Bảng dưới đây trình bày các chỉ số đánh giá hiệu năng của hai mô hình SVM được lựa chọn, bao gồm: SVM_Full (huấn luyện trên toàn bộ tập đặc trưng gốc) và SVM_12PCA (huấn luyện trên 12 thành phần chính sau PCA).

Chỉ số	SVM_Full	SVM_12PCA
Accuracy	0.9887	0.9787
Sensitivity	0.9717	0.9543
Specificity	0.9929	0.9848
F1-score	0.9717	0.9471

Hình 28: So sánh hiệu năng trên tập kiểm tra

Nhận xét:

- **SVM_Full** vượt trội hơn ở tất cả các chỉ số, đặc biệt là về Accuracy (0.9887 so với 0.9787) và F1-score (0.9717 so với 0.9471).
- Độ nhạy (Sensitivity) và độ đặc hiệu (Specificity) của **SVM_Full** cũng cao hơn, cho thấy mô hình này vừa phát hiện tốt các trường hợp dương tính, vừa hạn chế được báo động giả.
- Tuy nhiên, **SVM_12PCA** vẫn thể hiện hiệu năng khá tốt, trong khi sử dụng số chiều đặc trưng ít hơn đáng kể → phù hợp với các bài toán cần giảm thiểu độ phức tạp mô hình hoặc tài nguyên tính toán.

Kết luận: Với mục tiêu tối ưu hiệu năng dự đoán, mô hình **SVM_Full** là lựa chọn phù hợp nhất để triển khai trên tập dữ liệu hiện tại. Tuy nhiên, mô hình **SVM_12PCA** vẫn có thể được cân nhắc trong các tình huống yêu cầu đơn giản hóa mô hình hoặc giảm thiểu chi phí tính toán.

Dưới đây là phiên bản tối ưu hơn của phần kết luận, được chỉnh sửa để đảm bảo văn phong học thuật, cô đọng nhưng sâu sắc, và nhấn mạnh đồng thời đóng góp kỹ thuật lẫn ý nghĩa ứng dụng thực tiễn của nghiên cứu:

7 Kết luận

Nghiên cứu này đã chứng minh rằng từ các đoạn tín hiệu EEG ngắn hạn, hoàn toàn có thể xây dựng một hệ thống tự động nhận diện cơn động kinh với độ chính xác và độ tin cậy cao, mở ra giải pháp khả thi cho việc hỗ trợ chẩn đoán thay thế phương pháp thủ công truyền thống vốn phụ thuộc nhiều vào chuyên gia.

Cốt lõi của hệ thống là quy trình tiền xử lý và trích xuất đặc trưng được thiết kế kỹ lưỡng. Việc khai thác thông tin từ nhiều miền - bao gồm miền thời gian, miền tần số và miền wavelet - đã cho phép làm nổi bật các đặc điểm tín hiệu đặc trưng cho trạng thái co giật, ngay cả khi chúng không dễ nhận biết từ dữ liệu thô. Các phân tích thống kê đa biến (MANOVA) và trực quan hóa phi tuyến (t-SNE, UMAP) đã xác nhận sự phân tách rõ ràng giữa hai trạng thái, cho thấy tiềm năng phân loại rất cao của bộ đặc trưng thu được.

Để đối phó với không gian đặc trưng có chiều cao, nghiên cứu đã tích hợp Phân tích Thành phần Chính (PCA) như một bước rút gọn chiều hiệu quả. Dù mô hình sử dụng dữ liệu sau PCA (SVM_12PCA) có độ chính xác thấp hơn đôi chút, hiệu suất vẫn duy trì ở mức rất cao (97.87%), cho thấy PCA là một lựa chọn hợp lý trong các tình huống yêu cầu giảm thiểu độ phức tạp tính toán, ví dụ như triển khai thực tế trên thiết bị đeo hoặc hệ thống nhúng.

Trong số các mô hình được đánh giá, **Máy Vector Hỗ trợ với bộ đặc trưng đầy đủ (SVM_Full)** nổi bật với hiệu năng vượt trội: đạt **độ chính xác (accuracy) 98.87%, độ nhạy (sensitivity) 97.17%, và độ đặc hiệu (specificity) 99.29% trên tập kiểm tra**. Mức hiệu suất này không chỉ cho thấy khả năng phân loại mạnh mẽ, mà còn đặc biệt phù hợp trong bối cảnh lâm sàng – nơi cả hai yếu tố: phát hiện chính xác và giảm thiểu báo động giả, đều có ý nghĩa quan trọng.

Bên cạnh giá trị học thuật, kết quả nghiên cứu còn mở ra các hướng ứng dụng đầy triển vọng: từ công cụ hỗ trợ chuyên gia thần kinh trong đánh giá nhanh, đến nền tảng cho các hệ thống cảnh báo sớm tại giường bệnh hoặc thiết bị theo dõi cá nhân cho bệnh nhân động kinh. Trong tương lai, nghiên cứu có thể được mở rộng theo hướng: kiểm thử trên tập dữ liệu đa trung tâm, triển khai theo thời gian thực, hoặc tích hợp các mô hình học sâu (deep learning) nhằm hướng đến khả năng tự động hóa toàn diện quá trình phân tích tín hiệu EEG.

Tài liệu tham khảo

Tài liệu

[1] Nguyễn, Thị Mộng Ngọc, Đinh, Ngọc Thanh, và Đặng, Đức Trọng. (2025). *Thống kê nhiều chiều*. Nhà xuất bản Đại học Quốc gia TP.HCM.

[2] Maaten, L. van der, & Hinton, G. (2008). *Visualizing Data using t-SNE*. Journal of Machine Learning Research, 9(11).

[3] McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. ArXiv:1802.03426.

[4] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

[5] Welch, P. (1967). *The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms*. IEEE Transactions on Audio and Electroacoustics, 15(2), 70–73.

[6] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.

[7] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.

[8] Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.