HỆ THỐNG CẢI TIẾN NÂNG CAO HIỆU NĂNG GIAO DIỆN NÃO - MÁY TÍNH THÔNG QUA VIỆC GIẢI MÃ DỮ LIỆU BỊ MẤT CỦA TÍN HIỆU ĐIỆN NÃO ĐỒ KHÔNG XÂM LẤN

AN IMPROVEMENT FRAMEWORK FOR NONINVASIVE EEG-BASED BRAIN - COMPUTER INTERFACES PERFORMANCE VIA ENCODING MISSING SIGNALS

Dương Thanh Linh[1], Lương Duy Đức[2], Nguyễn Thị Ngọc Anh[3*]

[1]Trường Đại học Bình Dương
[2]Học viên cao học ngành Hệ thống Thông tin, Trường Đại học Sư phạm - Đại học Đà Nẵng
[3]Trường Đại học Sư phạm - Đại học Đà Nẵng

*Tác giả liên hệ: ngocanhnt@ued.udn.vn

(Nhận bài: 05/5/2022; Chấp nhận đăng: 25/6/2022)

Tóm tắt - Phương pháp đề xuất trong bài báo nhằm mục đích nắm bắt các mô hình tối ưu dựa trên hai đặc điểm chính trong chuỗi thời gian điện não đồ (EEG) liên tục: Động lực thông qua khám phá các hành vi phát triển theo thời gian và các mối tương quan bằng cách xác định mối quan hệ tiềm ẩn giữa nhiều tín hiệu não. Từ những khai thác này, phương pháp được đề xuất trích xuất thành công khai thác các biến ẩn và phát hiện ra động lực của chúng để khôi phục tự động các giá trị còn thiếu. Các thử nghiệm mô phỏng chứng minh rằng phương pháp được đề xuất cung cấp hiệu suất tái tạo tốt hơn lên đến 67% so với phương pháp phân tích suy biến cho giá trị bị mất (MSVD) và phương pháp nội suy. Sau đó, thử nghiệm phân loại chuyển động trên dữ liệu hoàn chỉnh, dữ liệu bị thiếu và dữ liệu khôi phục theo phương pháp đề xuất cho kết quả chính xác lần lượt là 92,15%, 73,19% và 86,18%, điều này chứng minh tính khả thi trong việc ứng dụng của phương pháp đề xuất.

Từ khóa - Điện não đồ (EEG); dữ liệu bị mất; Kalman Filter; phân tích suy biến cho giá trị bị mất (MSVD).

Abstract - The purpose of the proposed method in this article is to capture the optimal patterns that based on two main characteristics in the coevolving Electroencephalogram (EEG) time series including Dynamics via discovering temporal evolving behaviors and correlations by identifying the implicit relationships among multiple brain signals. From these exploits, the proposed method successfully identifies a few hidden variables and discovers their dynamics for automatic recovery of the missing values. The experimental simulations demonstrate that the proposed method provides a better reconstruction performance up to 67% improvements over Missing value Singular Values Decomposition (MSVD) and interpolation approaches. Then, we conducted an experiment for classifying movement based on the complete data, the missing data, and the restored one according to the proposed methods; with the exact results of 92.15%, 73.19%, and 86.18% respectively. The results of the experiment proved the feasibility in the application of the proposed method.

Key words – Electroencephalogram (EEG); missing data; Kalman Filter; Missing value Singular Values Decomposition (MSVD).

1. Đặt vấn đề

Điện não đồ (EEG) là một kỹ thuật ghi lại hoạt động điện do não tạo ra bằng cách sử dụng các điện cực. Có hai phương pháp để thu được tín hiệu điện não đồ: (1) Xâm lấn và (2) không xâm lấn. Trong phương pháp xâm lấn các điện cực được đặt trên bề mặt tiếp xúc của não, phương pháp không xâm lấn các điện cực được đặt dọc theo da đầu. Giao diện não - máy tính (BCI) là công nghệ sử dụng các điện cực khác nhau để thu thập các tín hiệu điện sinh học do hoạt động của não tạo ra, sau đó xử lý và phân tích các tín hiệu thông qua máy tính để giải mã các tín hiệu như chuyển động và thị giác, nhằm đạt được sự tương tác giữa người và máy tính. BCI cung cấp một kênh giao tiếp trực tiếp giữa não và thiết bị bên ngoài mà không liên quan đến bất kỳ hoạt động cơ bắp nào. Các hệ thống này hoặc sử dụng hoạt động điện não đồ được ghi lại từ da đầu hoặc hoạt động của các tế bào thần kinh vỏ não riêng lẻ được ghi lại từ các điện cực được cấy ghép. BCI có nhiều ứng dụng như điều khiển bộ phận cơ thể giả, điều hướng robot, điều khiển hệ thống tự động hóa tại nhà, điều khiển các ứng dụng điện thoại di động, điều khiển chuyển động của xe lăn và hệ thống nhận dạng giọng nói.

Trong hầu hết các phân tích dữ liệu chuỗi thời gian, các giá trị bị thiếu do nhiều lý do khác nhau như lỗi của con người hoặc lỗi thiết bị dẫn đến giảm hiệu suất hoặc thậm chí gây ra lỗi hệ thống. Các kỹ thuật phân tích dữ liệu EEG được áp dụng gần đây không chỉ áp đặt thống kê truyền thống mà còn áp dụng phương pháp tổng hợp dựa trên học máy để xử lý các giá trị bị thiếu. Tuy nhiên, các phương pháp này không có khả năng tạo ra các tín hiệu chuỗi thời gian thực tế liên quan đến thông tin tiềm ẩn quan trọng cần thiết để khai thác trong ứng dụng mục tiêu, chẳng hạn như phân loại chuyển động dựa trên điện não đồ.

Để có tập dữ liệu EEG hoàn chỉnh trong thế giới thực là điều gần như không thể. Đặc biệt, trong lĩnh vực y học và chăm sóc sức khỏe, người ta cũng báo cáo rằng phần lớn các bản ghi EEG chứa một số lượng lớn các giá trị bị thiếu. Việc ghi các dữ liệu không thành công có thể là do sự cố của thiết bị ghi, bị mất bản ghi hoặc do nhầm lẫn trong việc gắn điện cực. Ngoài ra, rất khó để ghi lại dữ liệu điện não đồ hoàn chỉnh, vì các yêu cầu nghiêm ngặt của môi trường ghi hoặc các đối tượng tham gia. Do đó, hầu hết các ứng dụng sử dụng bộ dữ liệu bị thiếu các giá

[1] Binh Duong University (Duong Thanh Linh)
[2] Master's student of Information Systems major, The University of Danang - University of Science and Education (Luong Duy Duc)
[3] The University of Danang - University of Science and Education (Nguyen Thi Ngoc Anh)

trị đều có thể đưa ra kết quả sai hoặc chuẩn đoán không chính xác.

Để xử lý những thách thức trên, bài báo đề xuất một cách tiếp cận mới phù hợp với dữ liệu có sẵn khi các giá trị bị thiếu. Mục tiêu chính của là khai thác các mối tương quan và phát triển hành vi của nhiều điện cực bằng cách tự động xác định một vài biến ẩn, sau đó khai thác động lực của chúng để giải quyết vấn đề thiếu khi quan sát. Sự tương quan ngụ ý rằng các kích thước quan sát được của nhiều điện cực không phụ thuộc. Do đó, các giá trị bị thiếu có thể được suy ra từ những giá trị khác thông qua các biến ẩn. Hành vi đang phát triển biểu thị rằng các giá trị còn thiếu có thể được ước tính một cách hiệu quả dựa trên quan sát của những "người hàng xóm" về các lần tích tắc tiếp theo và theo xu hướng di chuyển của chúng. Để đánh giá hiệu quả của phương pháp được đề xuất bằng cách xem xét các khía cạnh của độ chính xác, độ tin cậy và độ phức tạp. Bài báo này chứng minh hiệu suất của việc khôi phục cho các dữ liệu bị thiếu liên tiếp trên hai thực các bộ dữ liệu khác nhau của tín hiệu EEG. Phương pháp được đề xuất có thể hiệu quả để nắm bắt một vài biến ẩn một cách tự động cũng như để minh họa một cách ngắn gọn cách tìm hiểu động lực học của chúng để tính toán các giá trị bị thiếu liên tiếp. Hơn nữa, thời gian tính toán của nó tăng tuyến tính với thời gian của các chuỗi. Sau đó, so sánh kết quả của phương pháp được đề xuất với các phương pháp MSVD và phương pháp nội suy.

2. Mô tả bài toán phục hồi dữ liệu EEG bị mất

Thông thường các tín hiệu điện não đồ bị suy giảm vì các lý do khác nhau như ngắt kết nối điện cực với cơ thể hoặc tín hiệu nhiễu. Việc khôi phục dữ liệu bị thiếu trong các ứng dụng thực tế là rất cần thiết, vì nó có ảnh hưởng tiêu cực đến độ chính xác của việc phân loại, dẫn đến việc các ứng dụng đưa ra kết quả không chính xác [1-2]. Thông thường dữ liệu bị mất có hai loại: (1) Không có cấu trúc và (2) có cấu trúc. Dữ liệu bị mất không có cấu trúc nghĩa là các giá trị dữ liệu bị mất trên các chỉ số ngẫu nhiên của dữ liệu quan sát. Trong khi dữ liệu bị mất có cấu trúc là một phần dữ liệu từ các cảm biến cụ thể bị thiếu. Trong các ứng dụng thực tế, dữ liệu thường bị mất theo cách có cấu trúc, do đó nghiên cứu này đề xuất thiết lập mô hình dữ liệu bị thiếu có cấu trúc, dữ liệu EEG được sắp xếp theo nhiều chiều như một tensor [3-4] để bảo toàn tính chất đa chiều của dữ liệu.

Một trong những giải pháp đơn giản nhất hiện có là thay thế mỗi giá trị bị thiếu bằng các phương pháp như tính giá trị trung bình thích hợp. Một phương pháp xen kẽ khác để lấp đầy các giá trị bị thiếu là phương pháp nội suy, có liên quan đến việc xử lý các phần tử bị thiếu bằng cách sử dụng khớp nối đường cong, được gọi là nội suy tuyến tính và splines. Chi tiết về các cách tiếp cận này và khả năng áp dụng của chúng có thể được tìm thấy trong [5-6]. Tuy nhiên, các phương pháp này trở nên không phù hợp hoặc gặp thách thức lớn khi khoảng cách quan sát của các giá trị bị mất quá lớn. Hơn nữa, những cách tiếp cận này loại bỏ hoàn toàn bất kỳ mối quan hệ nào giữa các biến theo thời gian.

Trong những nghiên cứu gần đây để xử lý các giá trị bị mất bằng cách sử dụng mô hình thống kê được gọi là hệ thống động lực học tuyến tính, có thể được sử dụng để ước tính giá trị cho các điểm thời gian bị thiếu. Trong [7] đã chứng minh tính hiệu quả của phương pháp của họ bằng cách khám phá mô hình hệ thống động lực học tuyến tính để biểu hiện gen với các giá trị bị thiếu. Trong [8] sử dụng bộ lọc Kalman, dự đoán vị trí điểm đánh dấu bị khuyết trên tập dữ liệu chuyển động của con người. Trong [9-10] chỉ ra cách sử dụng vị trí người tạo trước đó và mô hình bộ xương để ước tính các vị trí điểm đánh dấu bị thiếu bằng cách sử dụng bộ lọc Kalman mở rộng. Tuy nhiên, các mô hình này trở nên không hiệu quả khi các vị trí bị thiếu của điểm đánh dấu được giữ trong một thời gian dài.

Với những lý do trên, bài báo nghiên cứu đề xuất một phương thức mới được xây dựng để tính các giá trị còn thiếu trong chuỗi thời gian EEG để khôi phục dữ liệu bị thiếu một cách tự động.

3. Đề xuất mô hình

3.1. Hệ thống động lực học tuyến tính

EEG là dữ liệu đa chiều vì nhiều điện cực được sử dụng để ghi lại hoạt động điện dọc theo bề mặt da đầu. Một hệ thống động lực học có thể được mô hình hóa bởi một chuỗi các tín hiệu EEG đa chiều, ký hiệu là $Y = \{y_1, y_2, \dots, y_T\}$, trong đó mỗi vectơ y_t biểu thị dữ liệu tại mỗi thời điểm đánh dấu $\mathrm{t} = 1, 2, \dots \mathrm{T}$ của chiều m. Điều này có nghĩa là dữ liệu từ chuỗi thời gian EEG có thể được trình bày bằng ma trận Y_{mxT} với biến m và thời gian quan sát T. Nhóm tác giả xem dữ liệu chuỗi thời gian EEG thu được từ các tín hiệu EEG trong một hệ thống động lực học như vậy. Sau đó, xây dựng một mô hình thống kê để biểu diễn trạng thái của các biến ẩn đang phát triển thành một phép biến đổi tuyến tính dẫn đến các chuỗi thời gian số được quan sát. Mô hình có thể tìm hiểu động lực của dữ liệu chuỗi thời gian [11]. Nó nắm bắt mối tương quan giữa nhiều điện cực bằng cách chọn một số biến ẩn thích hợp. Đặc biệt, hệ thống động lực học tuyến tính (LDS) cho chuỗi thời gian EEG đa chiều được mô hình hóa bằng các phương trình sau:

$$z_1 = \mu_0 + \omega_0 \quad (1a)$$

$$z_{n+1} = A.z_n + \omega_n \quad (1b)$$

$$y_n = C.z_n + \varepsilon_n \quad (1c)$$

Trong đó, $\theta = \{\mu_0, Q_0, A, Q, C, R\}$ tập các tham số μ_0 là trạng thái ban đầu cho các biến ẩn của toàn hệ thống. Vectơ y_n và z_n lần lượt biểu thị các chuỗi dữ liệu quan sát và các biến ẩn tại thời điểm t. Ma trận chuyển đổi A liên quan đến sự chuyển đổi trạng thái từ tích tắc thời gian hiện tại sang đánh dấu thời gian tiếp theo có nhiễu $\{\omega_n\}$. Ma trận C là phép chiếu quan sát với nhiễu $\{\varepsilon_n\}$ tại mỗi thời điểm t, nghĩa là dãy các biến ẩn z_n đang phát triển theo thời gian tích với ma trận chuyển đổi tuyến tính A [12]. Hơn nữa, các chuỗi dữ liệu quan sát y_n được tạo ra từ chuỗi các biến ẩn này với ma trận chiếu tuyến tính C. Tất cả nhiễu ω_0, ω_i và ε_i(i=1...T) là các biến ngẫu nhiên có phân phối chuẩn không trung bình với các ma trận hiệp phương sai Q_0, Q và R tương ứng. Trong mô hình, chỉ quan sát hệ thống được trình bày, trạng thái và tất cả các biến nhiễu đều bị ẩn. Định nghĩa và mô tả toán học của các ký hiệu được sử dụng trong hệ thống được thể hiện trong Bảng 1.

Bảng 1. *Định nghĩa và mô tả toán học*

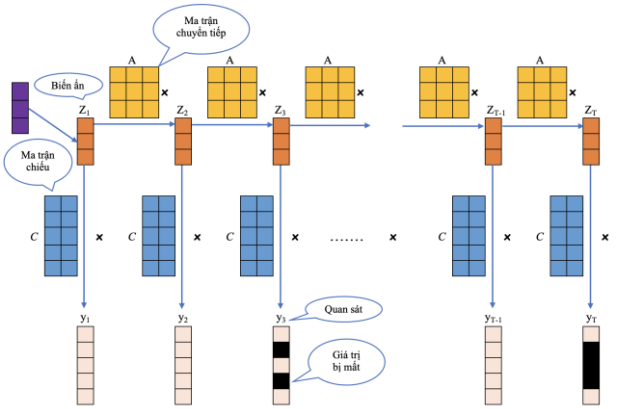
Ký hiệu	Định nghĩa và mô tả toán học
Y	Chuỗi quan sát đa chiều, $m \times T$
m	Chiều của chuỗi quan sát
T	Khoảng thời gian của chuỗi
W	Ma trận chỉ dẫn thiếu giá trị, $m \times T$
H	Thứ nguyên của các biến ẩn
μ_0	Trạng thái ban đầu cho biến ẩn, $H \times 1$
A	Ma trận chuyển đổi, $H \times H$
C	Ma trận chiếu từ trạng thái ẩn sang trạng thái quan sát, $m \times H$
Q	Hiệp phương sai chuyển tiếp, $H \times H$
Q_0	Hiệp phương sai ban đầu, $H \times H$
R	Hiệp phương sai của phép chiếu, $m \times m$
Z	Một chuỗi các biến ẩn, $\{z_1, z_2, ..., z_T\}$
θ	Một tập hợp bao gồm tất cả các thông số mô hình cần thiết, $\theta = \{\mu_0, Q_0, A, Q, C, R\}$

3.2. Thiết lập mô hình đề xuất khi thiếu các giá trị

Vấn đề thiếu tích tắc thời gian sẽ được mô hình hệ thống đề xuất xây dựng trong dữ liệu EEG lần đầu tiên. Trong thí nghiệm, xem xét tập hợp chuỗi thời gian Y có m chiều và độ dài T với các số đo bị mất; các giá trị bị thiếu của các quan sát được chỉ ra bởi ma trận W. Ma trận W của quan sát bị thiếu có cùng kích thước với Y và được xác định như dưới đây:

$$\boldsymbol{W(t,i)} = \begin{cases} \mathbf{1} & \textit{Nếu quan sát được Y theo chiều } \boldsymbol{i} \textit{ tại thời điểm } \boldsymbol{t} \\ \mathbf{0} & \textit{Nếu không quan sát được Y theo chiều } \boldsymbol{i} \textit{ tại thời điểm } \boldsymbol{t} \end{cases} \quad (2)$$

Trình tự thời gian được mô hình hóa dựa trên LDS, như đã thấy trong các phương trình 1 và 2, với một ma trận bị thiếu W [12]. Nhóm tác giả sử dụng thuật toán tối đa hóa kỳ vọng (Expectation Maximization) để đưa ra các vị trí bị thiếu thông qua ước tính kỳ vọng của thuật toán về các giá trị bị thiếu, $E[Y_{miss}Y_{obs}]$, điều kiện dựa trên các giá trị quan sát, trong đó Y_{miss} và Y_{obs} lần lượt là tập hợp các biến cho các giá trị bị thiếu và tập hợp các giá trị quan sát trong chuỗi Y.



Hình 1. *Kiến trúc thiết lập mô hình đề xuất*

Để xử lý vấn đề thiếu các giá trị, mục tiêu chính là khai thác các mẫu có ý nghĩa thông qua việc tự động xác định một vài biến ẩn để động lực học của chúng sẽ được phát hiện để giải quyết vấn đề thiếu quan sát. Nghiên cứu này tập trung vào việc khai thác khả năng kết nối động của các tín hiệu não thông qua hai đặc tính cụ thể: Tính tương quan và tính liên tục theo thời gian. Để đáp ứng vấn đề này, cần phải mô hình hóa động lực học và các dạng ẩn của chuỗi thời gian quan sát bằng cách sử dụng chuỗi các biến trạng thái ẩn Z. Để mô hình hóa các mối tương quan, mô hình sử dụng chuỗi dữ liệu, bao gồm cả giá trị quan sát được và giá trị bị thiếu, được tạo ra từ nhiều biến ẩn thông qua phép chiếu tuyến tính MxT ma trận chiếu tuyến tính C tại mỗi thời điểm, được hiển thị trong Hình 1 trong đó H là số biến ẩn.

Mặt khác, để mô hình hóa thuộc tính liên tục theo thời gian, vì các biến ẩn phụ thuộc thời gian với các giá trị được xác định từ lần đánh dấu thời gian trước đó. Điều này có nghĩa là ma trận A có liên quan đến sự chuyển đổi trạng thái của các biến ẩn theo thời gian, mô tả cách các trạng thái tiến lên theo thời gian. Như vậy, điểm thời gian tiếp theo chỉ phụ thuộc vào điểm thời gian hiện tại. Trong trường hợp này, trước tiên đặt trạng thái ban đầu cho các biến ẩn tại thời điểm bắt đầu với tập các tham số $\theta = \{\mu_0, Q_0, A, Q, C, R\}$. Trong hệ thống, phân phối chung của Y_{obs}, Y_{miss} và Z bởi phương trình sau:

$$P(Y_{miss}, Y_{obs}\ and\ Z) = P(z_1).\prod_{i=2}^{T} P(z_i|z_{i-1}).\prod_{i=1}^{T} P(y_i|z_i) \quad (3)$$

Để đạt được những mục tiêu trên, mô hình đề xuất được đưa ra để tìm ra giải pháp tối ưu nhằm tối đa hóa khả năng ghi nhật ký dự kiến của chuỗi quan sát liên quan đến các tham số của mô hình $\theta = \{\mu_0,\ Q_0,\ A,\ Q,\ C,\ R\}$, các biến ẩn $\hat{z}_n = E[z_n], n = 1 \dots T$, và các quan sát bị thiếu $E[Y_{miss}Y_{obs}]$.

Trong thực tế, để đạt được ước lượng tham số, cần phải để tìm ra khả năng tối đa xảy ra $L(\theta) = P(Y_{obs})$. Tuy nhiên, người ta biết rằng rất khó để tối đa hóa khả năng dữ liệu khi có các giá trị bị thiếu. Do đó, mức độ giống như ước lượng khả năng cực đại của chuỗi quan sát trên tham số θ được tối đa hóa bằng cách sử dụng thuật toán Expectation-Maximization (EM) [12], lặp đi lặp lại bước tối đa hóa để dự đoán khả năng hoàn chỉnh như trong phương trình 4. Để đạt được ước tính khả năng xảy ra tối đa của các tham số mô hình, phương pháp EM để học LDS được sử dụng. Thuật toán lặp lại giữa việc tính toán kỳ vọng có điều kiện của các biến ẩn thông qua thuật toán tiến lùi (forward-backward) trong E-step và cập nhật các tham số mô hình để tối đa hóa khả năng của nó trong M-step để ước tính các giá trị bị thiếu [12].

$$L(\theta; Y) = E_{Y,Z|\theta}\Big[-(z_1 - \mu_0)^T Q_0^{-1}(z_1 - \mu_0) - \sum_{t=2}^{T}(z_t - A \cdot z_{t-1})^T Q^{-1}(z_t - A \cdot z_{n-1}) - \sum_{t=1}^{T}(y_t - C \cdot z_t)^T R^{-1}(y_n - C \cdot y_n)\Big] \quad (4)$$

Tóm lại, phương pháp đề xuất được thực hiện để đạt được các thông số tốt nhất $\theta = \{\mu_0, Q_0, A, Q, C, R\}$ cho mô hình. Phương pháp được áp dụng trong bài báo này tiến hành ba bước chính: Kỳ vọng, khôi phục các giá trị bị thiếu và tối đa hóa. Chi tiết hơn, thuật toán đầu tiên đoán một tập hợp ban đầu của các tham số mô hình trong bước kỳ vọng. Sử dụng bộ lọc Kalman và làm mịn Kalman để ước tính các biến ẩn dựa trên quan sát và các tham số hiện tại cho mỗi lần lặp. Ý tưởng chung là sử dụng thuật toán tiến-lùi

để tính toán các kỳ vọng sau của các biến ẩn, $E(z_n|Y;\theta)$, đánh dấu bằng tích tắc dựa trên tính toán của lần trước đánh dấu. Với dữ liệu có các giá trị bị thiếu, ước lượng tìm phân phối biên cho các biến trạng thái ẩn sau khi khởi tạo các giá trị còn thiếu là một số ngẫu nhiên bằng phương pháp nội suy. Cả hai phân phối trước có điều kiện trong mô hình đều là Gaussian, do đó đánh dấu sau tính đến thời điểm hiện tại là $p(z_n\, y_1,...,y_T)$, cũng được Gaussian đưa ra bởi:

$$\hat{\alpha}(z_n) = N(\mu_0, V_n) \quad (5)$$

Chúng ta thu được các phương trình suy diễn tiến-lùi sau đây. Các giá trị ở đây là μ_n , V_n và P_{n-1}, được đưa ra bởi:

$$P_{n-1} = A.V_{n-1}.A^T + Q \quad (6)$$

$$K_n = P_{n-1}.C^T(C.P_{n-1}.C^T + R)^{-1} \quad (7)$$

$$V_n = (I - K_n).P_{n-1}) \quad (8)$$

$$\mu_n = A.\mu_{n-1} + K_n.(y_n - C.A.\mu_{n-1}) \quad (9)$$

Các giá trị ban đầu được cho bởi các phương trình sau:

$$\boldsymbol{K_1 = Q_0 C^T (G Q_0 C^T + R)^{-1}} \quad (10)$$

$$\boldsymbol{\mu_1 = \mu_0 + K_1(y_1 - C.A.\mu_0)} \quad (11)$$

$$\boldsymbol{V_1 = (I - K_1).Q_0} \quad (12)$$

Làm mịn bao gồm một đệ quy mở đầu, sau đó đệ quy lùi. Trong bước tiếp theo, các giá trị của phương trình bộ lọc Kalman được lưu trữ. Trong bước lùi, các giá trị này sau đó được sử dụng để khởi tạo các phương trình Kalman mượt mà hơn được đưa ra bởi:

$$\boldsymbol{\hat{\mu}_n = \mu_n + J_n.(\hat{\mu}_{n+1} - A.\mu_n)} \quad (13)$$

$$\boldsymbol{\hat{V}_n = V_n + J_n.(\hat{V}_{n+1} - P_n).J_n^T} \quad (14)$$

$$\boldsymbol{J_n = V_n.A^T.P_n^{-1}} \quad (15)$$

Các kỳ vọng được lấy từ phân phối cận biên sau $p(z_n|y_1,...,y_T)$ từ việc những quan điểm phổ biến. Do đó, các kỳ vọng nhận được bằng các phương trình sau:

$$E[z_n] = \hat{\mu}_n \quad (16)$$

$$E[z_n z_{n-1}^T] = J_{n-1}\hat{V}_n + \hat{\mu}_n\hat{\mu}_{n-1}^T \quad (17)$$

$$E[z_n z_n^T] = \hat{V}_n + \hat{\mu}_n\hat{\mu}_n^T \quad (18)$$

Trong bước khôi phục, các giá trị bị thiếu được khôi phục bằng cách sử dụng thuộc tính Markov trong mô hình đồ họa từ ước lượng các biến ẩn, Hình 1 với các phương trình sau:

$$\boldsymbol{E[Y_{miss}|Y_{obs}, Z; \theta] = C.E[Z]_{\{i,j\}},\ \ \{i,j\} \in W} \quad (19)$$

Trong bước tối đa hóa, thuật toán cập nhật tham số θ^{new} bằng cách tối đa hóa khả năng ghi nhật ký dự kiến bằng cách sử dụng một số thống kê đầy đủ từ phân phối sau. Để ước tính các thông số, số mũ giống như nhật ký dự kiến $L(\theta;Y)$ trong phương trình 4, liên quan đến các thành phần của θ^{new} được tối đa hóa. Lấy các đạo hàm của phương trình 4 và biến chúng thành 0 sẽ cho các kết quả sau:

$$\mu_0^{new} = E[z_1] \quad (20)$$

$$Q_0^{new} = E[z_1 z_1^T] - E[z_1]E[z_1^T] \quad (21)$$

$$A^{new} = (\sum_{n=2}^{T} E[z_n z_{n-1}^T])(\sum_{n-1}^{T-1} E[z_n z_n^T]^{-1}) \quad (22)$$

$$Q^{new} = \frac{1}{T-1}\sum_{n=2}^{T}(E[z_n z_n^T] - A^{new}E[z_{n-1}z_n^T] - E[z_n z_{n-1}^T](A^{new})^T + A^{new}E[z_n z_{n-1}^T](A^{new})^T \quad (23)$$

$$C^{new} = (\sum_{n=1}^{N} y_n[z_n^T])(\sum_{n=1}^{T} E[z_n z_n^T])^{-1} \quad (24)$$

$$R^{new} = \frac{1}{T}\sum_{n=1}^{T}(y_n y_n^T - C^{new}E[z_n]y_n^T - y_n E[z_n^T](C^{new})^T + C^{new}E[z_n y_n^T](C^{new})^T) \quad (25)$$

Về tổng thể, phương pháp đề xuất để giải quyết vấn đề thiếu giá trị trong chuỗi thời gian EEG có thể được tóm tắt như sau:

- Ước tính các biến ẩn Z (E-step): Với các tham số cố định, θ và Y chứa các giá trị bị thiếu, quy trình tiến-lùi để ước tính hậu nghiệm $P(Z|Y;\theta)$ và số liệu thống kê đầy đủ của nó $E(z_n|Y;\theta), E(z_n z'_n|Y;\theta), E(z_n z'_{n+1}|Y;\theta)$ được sử dụng.

- Khôi phục các giá trị bị thiếu: Đã cho Z cố định, các giá trị bị thiếu $Y_{miss}E(Y_{miss}|Z;\theta)$ sử dụng $E(z_n|Y;\theta)$ được ước tính.

Cập nhật thông số mô hình (M-step): Với Y và Z cố định, các thông số mô hình mới, $\theta^{new} \leftarrow argmax\ E[\log(Y, Z, \theta]$ được ước tính.

4. Đánh giá thực nghiệm

4.1. Dữ liệu thực nghiệm

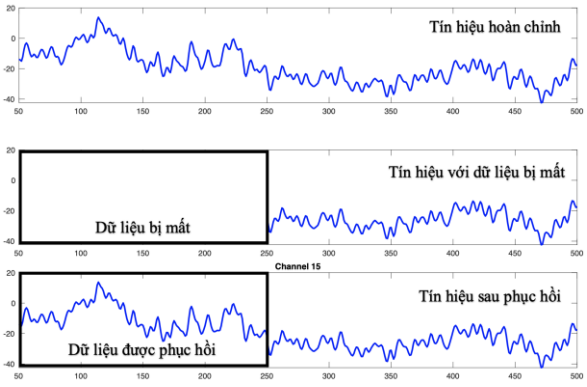
Trong nghiên cứu này, nhóm tác giả nghiên cứu dựa trên 2 bộ dữ liệu EEG:

- Bộ dữ liệu hình ảnh chuyển động ABSP EEG [13]. Các tín hiệu điện não đồ trong tập dữ liệu được ghi lại từ các đối tượng khỏe mạnh về mặt huyết thanh. Mô hình BCI dựa trên gợi ý bao gồm 2/3 nhiệm vụ hình ảnh vận động, cụ thể là tưởng tượng chuyển động của tay trái (Left Hand - LH), tay phải (Right Hand - RH) và cả hai chân (feet - F). Một số buổi thí nghiệm vào các ngày khác nhau được ghi lại cho một số thí nghiệm, dữ liệu của mỗi thí nghiệm được lưu trữ trong một tệp dữ liệu tương ứng. Có 31 tệp bao gồm các tín hiệu EEG cho các thí nghiệm riêng biệt;

- Bộ dữ liệu hình ảnh chuyển động BCI III (4a) [14]. Bộ dữ liệu này được cung cấp bởi Fraunhofer FIRST, Nhóm Phân tích Dữ liệu Thông minh (Klaus-Robert Müller, Benjamin Blankertz), và Cơ sở Benjamin Franklin của Charité - Đại học Y Berlin, Khoa Thần kinh, Nhóm Khoa học Thần kinh (Gabriel Curio) và được xuất bản tại web: *https://www.bbci.de/competition/iii/desc_IVa.html*.

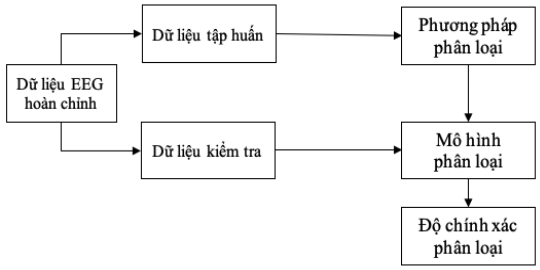
4.2. Mô hình thực nghiệm

Sự xuất hiện của giá trị bị mất trong bất kỳ dữ liệu thực tế nào cũng có ảnh hưởng tiêu cực đến kết quả phân tích của các thuật toán. Nghiên cứu này tập trung vào loại dữ liệu bị mất có cấu trúc vì đây là dạng dữ liệu bị mất hầu như luôn tồn tại trong thực tế. Bài báo đề xuất khôi phục dữ liệu bị mất có cấu trúc bằng cách sử dụng các phương pháp phân tích nhân tử dựa trên tensor nhằm bảo toàn tính chất đa chiều của dữ liệu và khôi phục dữ liệu bị mất một cách hiệu quả. Để hình dung một cách tổng quan mô hình thực nghiệm bài toán khôi phục dữ liệu EEG bị mất. Quan sát Hình 2 mô phỏng các tín hiệu EEG hoàn chỉnh (EEG gốc của bộ dữ liệu), dữ liệu bị mất một đoạn và dữ liệu được phục hồi. Hình 2(a) cho thấy, tín hiệu EEG hoàn chỉnh từ bộ dữ liệu EEG. Hình 2(b) cho thấy trường hợp một khoảng giá trị từ một kênh ngẫu nhiên của tín hiệu điện não đồ bị mất. Sau đó, tín hiệu EEG được phục hồi thông qua phương pháp đề xuất được thể hiện trong Hình 2(c).

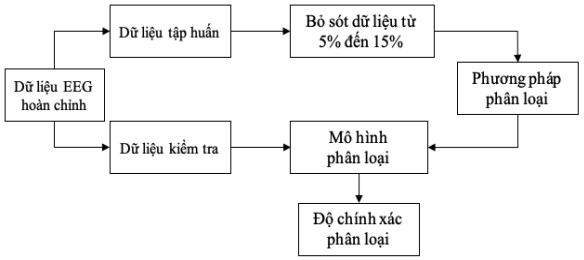


Hình 2. *Mô phỏng các tín hiệu EEG gốc, bị mất và được phục hồi*

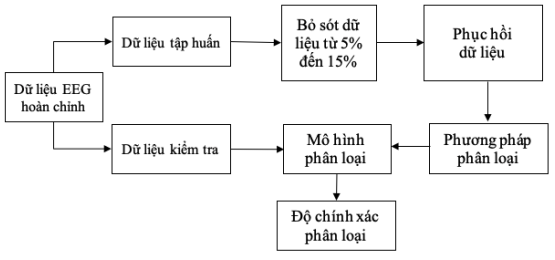
Mục đích của việc phục hồi dữ liệu bị mất là cải thiện độ chính xác của phân loại. Để đạt được độ chính xác phân loại cao, các tín hiệu nhiễu đã được loại bỏ khỏi điện não đồ. Nhóm tác giả đề xuất ba mô hình để so sánh hiệu suất kết quả phân loại. Mô hình thứ nhất, sử dụng các phương pháp phân loại trên tập dữ liệu hoàn chỉnh như được mô phỏng trong Hình 3. Mô hình thứ hai, cố tình bỏ mất dữ liệu từ 5% đến 15% và sau đó áp dụng các phương pháp phân loại như trong Hình 4. Trong mô hình thứ ba, khôi phục dữ liệu đã mất và sau đó sử dụng các phép phân loại như Hình 5. Kết quả lý tưởng nhất là hiệu suất phân loại trên dữ liệu được khôi phục phải bằng hoặc tốt hơn hiệu suất phân loại trên dữ liệu hoàn chỉnh. Trong khi đó, hiệu suất phân loại trên dữ liệu bị thiếu sẽ kém hơn so với hai mô hình còn lại.



Hình 3. *Mô hình phân loại trên dữ liệu EEG hoàn chỉnh*



Hình 4. *Mô hình phân loại trên dữ liệu EEG với 5% đến 15% giá trị bị mất*



Hình 5. *Mô hình phân loại trên dữ liệu điện não đồ đã phục hồi bằng phương pháp đề xuất*

4.3. Kết quả thực nghiệm

Để tiến hành thử nghiệm nghiên cứu, hai khía cạnh được xem xét để đánh giá hoạt động hiệu quả của phương pháp đề xuất đối với phương pháp MSVD và phương pháp nội suy (Interpolation). So sánh được thực hiện dựa trên khả năng ước tính những lượng giá trị khác nhau của các mục khác nhau có giá trị bị mất. Đối với mỗi thiết lập thử nghiệm, bài báo đã tạo các vị trí quan sát bị mất liên tiếp khác nhau trên các kênh ngẫu nhiên của bộ dữ liệu ABSP và BCI III (4a). Các thí nghiệm được lặp lại 10 lần để tránh ảnh hưởng ngẫu nhiên. Tính toán trung bình của sai số bình phương trung bình (MSE) để đánh giá chất lượng của phương pháp được đề xuất. MSE được tính theo công thức: $\sum\|\tilde{y}_t - y_t\|^2 / \sum\|y_t\|^2$, trong đó t biểu thị mỗi lần đánh dấu, $\tilde{y}_t$ là dữ liệu được tái tạo và y là dữ liệu đầu vào.

Bảng 2. *Lỗi cấu trúc lại đối với các tỷ lệ khác nhau của các giá trị bị mất 5%, 10% và 15%*

Dataset	Tệp	Phương pháp	Lỗi tái cấu trúc đối với dữ liệu bị mất so với dữ liệu gốc (MSE)			Trung bình MSE
			5%	10%	15%	
ABSP	*subA*	Đề xuất	**0,0039**	0,00602	0,0807	0,0302
		MSVD	0,0284	0,0509	0,0934	0,0576
		Interpolation	0,03463	0,07357	0,0926	0,0669
	subB	Đề xuất	**0,0015**	0,00523	0,00738	0,0047
		MSVD	0,02203	0,03374	0,09193	0,0492
		Interpolation	0,03959	0,08728	0,0966	0,0745
BCI III (4a)	*aw*	Đề xuất	**0,01109**	0,0738	0,07459	0,0532
		MSVD	0,13903	0,22025	0,29322	0,2175
		Interpolation	0,16643	0,28596	0,38245	0,2783
	ay	Đề xuất	**0,03013**	0,17195	0,30328	0,1685
		MSVD	0,19056	0,29601	0,99752	0,4947
		Interpolation	0,38527	0,60137	0,73021	0,5723

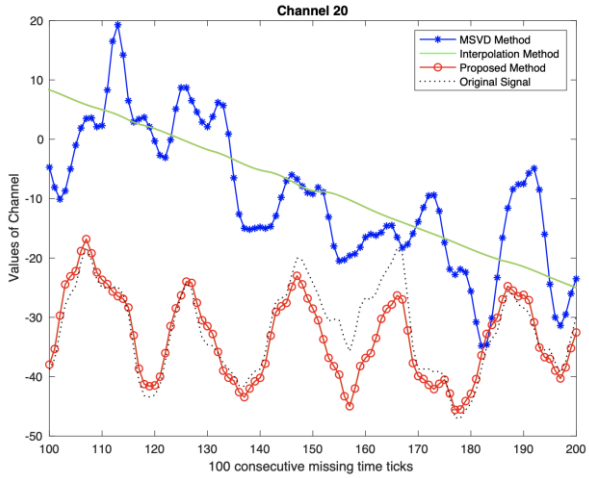
Đối với mỗi tập dữ liệu trong mỗi thử nghiệm, nguyên tắc của Fukunaka [15,16] được sử dụng như một công cụ để đạt được số h thích hợp cho kích thước ẩn của mô hình bằng phương pháp suy biến của dữ liệu ban đầu $Y=U\,x\,S\,x\,V^T$. Trong đó, cả U và V là ma trận trực chuẩn, S là ma trận đường chéo với các giá trị kỳ dị trên đường chéo. Để có được số h, các giá trị số ít nhỏ thường được đặt bằng *0*. Do đó, sắp xếp thứ tự các giá trị kỳ dị và sau đó chọn h ở giá trị có phân vị thứ 98 của tổng các giá trị kỳ dị bình phương.

Đầu tiên so sánh dựa trên sự khác biệt giữa các sai số xây dựng lại của ba phương pháp ở các giá trị bị thiếu lần lượt là 5%, 10% và 15%, được thể hiện trong Bảng 2. Trong tất cả các trường hợp, cả phương pháp đề xuất và phương pháp MSVD đều sử dụng cùng một số của các biến ẩn với 98% năng lượng; Độ dài trung bình của các giá trị còn thiếu liên tiếp là 35 điểm thời gian. Bảng 2 chứng minh rằng, trong tất cả các lượng dữ liệu bị thiếu khác nhau trong phạm vi 5%, 10% và 15%, sai số tái tạo của phương pháp đề xuất cho kết quả tốt nhất, có sai số nhỏ hơn so với phương pháp nội suy và phương pháp MSVD.

Cụ thể, trong tập dữ liệu ABSP, phương pháp được đề xuất đưa ra *0,0039* và *0,0105* sai số tái tạo trung bình, thấp hơn lần lượt so với phương pháp MSVD và phương pháp nội suy. Trên tập dữ liệu này, nó cho thấy sự cải thiện

khoảng 67% và 75% so với phương pháp MSVD và nội suy. Tương tự, trong thực nghiệm phục hồi dữ liệu bị mất với tập dữ liệu BCI III (4a) với các tệp *aw* và *ay*, hiệu suất của phương pháp được đề xuất cũng cho thấy việc tái tạo được cải thiện 69% và 74% so với hiệu suất của phương pháp MSVD và nội suy.

Để kiểm tra tính hiệu quả của phương pháp được đề xuất, Hình 3 cho thấy khả năng phục hồi dữ liệu trên kênh 20 của bộ dữ liệu BCI III (4a) theo ba cách tiếp cận. Trong tất cả các trường hợp, sự phục hồi tốt nhất được thực hiện với các điểm thời gian mất tích liên tiếp. Trong hình, đường nét chấm biểu thị các tín hiệu ban đầu (Original Signal), đường nét liền ký hiệu dấu sao (✱) các tín hiệu được tái tạo bằng phương pháp MSVD, đường nét liền là tín hiệu được phục hồi bằng phương pháp nội suy (Interpolation Method) và đường nét liền ký hiệu dấu tròn (○) mô tả các tín hiệu được phục hồi của phương pháp đề xuất. Nó cho thấy rằng phương pháp đề xuất (ký hiệu ○) đạt được sự tái tạo tốt nhất vì nó đạt rất gần với các tín hiệu ban đầu so với phương pháp MSVD và phương pháp nội suy.



Hình 3. *Khôi phục tín hiệu so sánh với tín hiệu ban đầu (giả sử khôi phục 100 điểm thời gian mất liên tiếp)*

Bảng 3. *So sánh kết quả thực nghiệm phân loại với dữ liệu giả định bị mất và dữ liệu sau khi phục hồi*

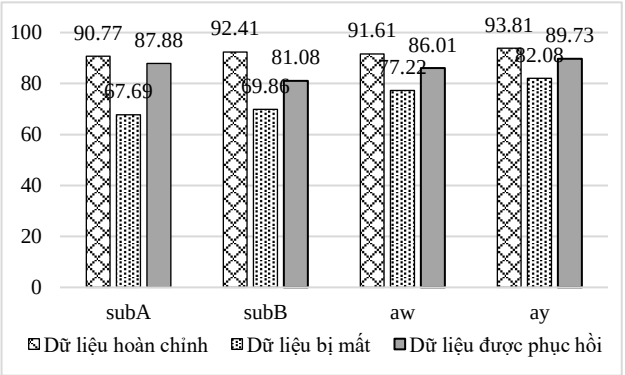
Dataset	Tệp	Phân loại với dữ liệu bị mất so với dữ liệu gốc				Phân loại với dữ liệu sau khi được phục hồi			
		5%	10%	15%	Trung bình	5%	10%	15%	Trung bình
ABSP	*subA*	78,08	63,46	61,54	67,69	89,71	87,29	86,63	87,88
	subB	86,42	70,68	52,47	69,86	90,57	84,78	67,88	81,08
BCI III (4a)	*aw*	84,40	75,39	71,88	77,22	90,60	85,81	81,61	86,01
	ay	84,99	85,32	65,93	78,75	91,7	90,87	86,63	89,73

Tiếp theo, nhóm tác giả thử nghiệm phân loại với bộ dữ liệu ABSP với hai đối tượng: *subA_6chan_2LR_s1 (subA)* và *subB_6chan_2LR (subB)*. Tập dữ liệu *subA* bao gồm 130 thử nghiệm hình ảnh động cơ BCI EEG cho đối tượng A. Tất cả các tín hiệu EEG được ghi lại trong thời gian 3 giây ở tần số lấy mẫu 256 Hz trên 6 kênh bằng bộ khuếch đại gTec. Mỗi thử nghiệm được gán một nhãn theo hình ảnh động cơ bên trái hoặc bên phải. Tập dữ liệu *subB* bao gồm 162 lần đó giả định dữ liệu của bộ ABSP bị mất lần lượt theo cấu trúc 5%, 10% và 15%, sau khi nhận dạng trên dữ liệu bị giả định bị mất ta thu được kết quả lần lượt là 78,08%, 63,46% và 61,54 cho tệp dữ liệu *subA*, với trung bình cộng là 67,69%. Sau đó, tiến hành phục hồi dữ liệu bằng phương pháp đề xuất ta thu được kết quả nhận dạng với trung bình cộng là 87,88% cải thiện được 20,19%.

Bảng 4. *Độ chính xác của phương pháp phân loại chuyển động với dữ liệu EEG*

Dataset	Tệp	Dữ liệu EEG hoàn chỉnh	Trungbình cộng dữ liệu EEG bị mất	Trung bình cộng của dữ liệu EEG sau phục hồi
ABSP	*subA*	90,77	67,69	87,88
	subB	92,41	69,86	81,08
BCI III (4a)	*aw*	91,61	77,22	86,01
	ay	93,81	77,98	89,73

Bộ dữ liệu BCI III (4a) chỉ chọn 7 kênh (51-57) từ 118 kênh của dữ liệu đầy đủ để minh họa hoạt động, tiến hành phân loại theo phương thức CSP (Common Spatial Pattern) thu được độ chính xác là 91,61%, 93,81% cho các tệp *aw*, *ay* hoàn chỉnh. Ta thu được kết quả phân loại sau khi phục hồi dữ liệu EEG bị mất lần lượt theo cấu trúc 5%, 10% và 15%, lần lượt là với trung bình cộng 86,01% với 89,73% cho tệp *aw* và *ay*. Các thực nghiệm đều cho kết quả cải thiện mức độ chính xác ít nhất 8,89% của thuật toán phân loại áp dụng trên dữ liệu được phục hồi so với dữ liệu bị mất.



Hình 4. *Hiệu quả nhận dạng của việc phục hồi dữ liệu EEG*

Hình 4 cho thấy, hiệu quả nhận dạng của việc phục hồi dữ liệu EEG. Thuật toán phân loại hoạt động cho kết quả gần giống với dữ liệu gốc ban đầu trên dữ liệu EEG đã được khôi phục. Điều này chứng tỏ khả năng ứng dụng của phương pháp phục hồi dữ liệu đề xuất.

5. Kết luận

Trong bài báo này, đã đề xuất một phương pháp để giải quyết vấn đề các giá trị bị thiếu liên tiếp cho chuỗi thời gian

EEG, đặc biệt là về việc tái tạo và phục hồi lại chúng. Phương pháp giải quyết các giá trị bị thiếu liên tiếp của chuỗi thời gian EEG thực đa chiều. Trong hầu hết các trường hợp, phương pháp này cung cấp kết quả tốt nhất so với các kỹ thuật thay thế như nội suy và MSVD. Việc khôi phục lại dữ liệu EEG cho kết quả gần giống trên dữ liệu EEG ban đầu theo phương pháp đề xuất chứng tỏ khả năng ứng dụng của phương pháp đề xuất.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Quỹ Phát triển Khoa học và Công nghệ Quốc gia (NAFOSTED) trong đề tài mã số 102.01-2020.27.

TÀI LIỆU THAM KHẢO

[1] Dornhege, Guido, et al. "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms", *IEEE transactions on biomedical engineering* 51.6 (2004): 993-1002.

[2] Horst, Reiner, and Panos M. Pardalos, eds. *Handbook of global optimization*. Vol. 2. Springer Science & Business Media, 2013.

[3] Kousarrizi, Mohammad Reza Nazari, et al. "Feature extraction and classification of EEG signals using Wavelet transform, SVM and artificial neural networks for brain computer interfaces", *2009 international joint conference on bioinformatics, systems biology and intelligent computing*. IEEE, 2009.

[4] Lacy, Seth L., and Dennis S. Bernstein. "Subspace identification with guaranteed stability using constrained optimization", *IEEE Transactions on automatic control* 48.7 (2003): 1259-1263.

[5] Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)", *Artificial Intelligence Review* 53.2 (2020): 1487-1509.

[6] Lin, Wan-Ju, et al. "Evaluation of deep learning neural networks for surface roughness prediction using vibration signal analysis", *Applied Sciences* 9.7 (2019): 1462.

[7] Mehra, Mani, V. K. Mehra, and V. K. Ahmad. *Wavelets theory and its applications*. Springer Singapore, 2018.

[8] Mistry, Krupal Sureshbai, et al. "An SSVEP based brain computer interface system to control electric wheelchairs", 2018 *IEEE International Instrumentation and Measurement Technology Conference* (I2MTC). IEEE, 2018.

[9] Ramachandra, K. V. *Kalman filtering techniques for radar tracking*. CRC Press, 2018.

[10] Shivappa, Vinay Kumar Karigar, et al. "Home automation system using brain computer interface paradigm based on auditory selection attention", 2018 *IEEE international instrumentation and measurement technology conference* (I2MTC). IEEE, 2018.

[11] Akmal, Muhammad, Syed Zubair, and Hani Alquhayz. "Classification analysis of tensor-based recovered missing EEG data", *IEEE Access* 9 (2021): 41745-41756.

[12] Campbell, Andrew, et al. "NeuroPhone: brain-mobile phone interface using a wireless EEG headset", *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds*. 2010.

[13] Cichocki, A., and Q. Zhao. "EEG motor imagery dataset", *Tech. Rep., Laboratory for Advanced Brain Signal Processing, BSI, RIKEN, Saitama, Japan* (2011).

[14] Dang, Lujuan, et al. "Kernel Kalman filtering with conditional embedding and maximum correntropy criterion", *IEEE Transactions on Circuits and Systems I: Regular Papers* 66.11 (2019): 4265-4277.

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, Calif, USA, 1990.

[16] L. Li, B. Aditya Prakash, and C. Faloutsos, "Parsimonious linear fingerprint for time series," *Proceedings of the VLDB Endowment*, vol. 3, pp. 385–396, 2010.

[17] Phan, A.H. NFEA: Tensor Toolbox for Feature Extraction and Applications; Technical Report; Lab for Advanced Brain Signal Processing, Brain Science Institute RIKEN: Hirosawa Wako City, Japan, 2011.

[18] Thi, Ngoc Anh Nguyen, Hyung-Jeong Yang, and Sun-Hee Kim. "Exploiting patterns for handling incomplete coevolving eeg time series", *International Journal of Contents* 9.4 (2013): 1-10.

[19] Tong, Hanghang, and Lei Ying. "NetDyna: mining networked coevolving time series with missing values", 2019 *IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.