

BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG
PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI
BỘ MÔN CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN MÔN HỌC

TÊN ĐỀ TÀI:
HỆ THỐNG PHÂN LOẠI NGƯỜI NGHIÊN RƯỢU DỰA TRÊN DỮ LIỆU EEG

Giảng viên hướng dẫn:

Giảng viên Vũ Thị Hạnh

Sinh viên thực hiện:

Lê Ngọc Tiền

Phan Trần Tường Vy

Nguyễn Hữu Tuấn Phát

Đoàn Anh Vũ

MSSV:

2251068262

2251068284

2351267274

2351267280

Lớp:

S26-65TTNT và S25-64CNTT

TP. HỒ CHÍ MINH, 2026

Mục lục

Mục lục	1
Danh mục các ký hiệu, các chữ viết tắt	3
Danh mục các bảng	4
Danh mục các hình vẽ, đồ thị	5
LỜI CẢM ƠN	6
MỞ ĐẦU	8
Chương 1 TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT	9
1.1 Cơ sở lý thuyết về Điện não đồ (EEG)	9
1.1.1 Hệ thống vị trí điện cực 10-20	9
1.1.1.1 Vùng Trán (Frontal - F)	9
1.1.1.2 Vùng Trung tâm (Central - C)	12
1.1.1.3 Vùng Thái dương (Temporal - T)	13
1.1.1.4 Vùng Đỉnh (Parietal - P)	14
1.1.1.5 Vùng Chẩm (Occipital - O)	15
1.1.1.6 Các kênh đặc biệt (Special Channels)	15
1.2 Mục tiêu nghiên cứu	15
1.3 Bài toán đặt ra	16
Chương 2 DỮ LIỆU VÀ PHƯƠNG PHÁP ĐỀ XUẤT	17
2.1 Tổng quan quy trình thực hiện	17
2.1.1 Đặc tả kỹ thuật Đầu vào - Đầu ra	18
2.2 Mô tả bộ dữ liệu SMNI_CMI	18
2.3 Phân tích số liệu và Thăm dò (EDA)	18
2.3.1 Phân bố nhãm và Kiểm tra rò rỉ (Data Leakage)	18
2.3.2 Cấu hình khử nhiễu EOG (ICA)	19
2.3.3 Phân tích ngoại lai và Phân phối đặc trưng (Outliers)	19
2.3.4 Độ quan trọng đặc trưng (Feature Importance)	19
2.4 Quy trình tiền xử lý dữ liệu (Preprocessing)	20
2.4.1 Dữ liệu thô ban đầu (Raw Input)	20
2.4.2 Bước 1: Chuẩn hóa và Chuyển đổi (Normalization & Pivoting)	20
2.4.3 Bước 2: Xử lý kênh hỏng và giá trị thiếu (Cleaning)	20
2.4.4 Bước 3: Khử nhiễu EOG bằng thuật toán ICA (Artifact Removal)	20
2.4.5 Bước 4: Xử lý ngoại lai và Chuẩn hóa đặc trưng (Advanced Outlier Handling)	21
2.4.6 Dữ liệu đầu ra sau tiền xử lý (Output)	21
2.5 Phương pháp trích xuất đặc trưng	21
2.5.1 Đặc trưng miền Tần số (Frequency Domain)	21
2.5.2 Biểu diễn Thời gian - Tần số (Spectrogram)	22
2.6 Các mô hình máy học đề xuất	22

2.6.1	Nhóm mô hình sử dụng đặc trưng dạng bảng (Bandpower/Time-domain)	22
2.6.2	Nhóm mô hình sử dụng đặc trưng hình ảnh (Spectrogram)	23
Chương 3	THỰC NGHIỆM VÀ KẾT QUẢ ĐẠT ĐƯỢC	24
3.1	Thiết lập thực nghiệm	24
3.2	Kết quả chi tiết từng mô hình	24
3.2.1	Nhóm mô hình sử dụng đặc trưng Bandpower (Tabular Data)	24
3.2.1.1	Logistic Regression (PhanTranTuonVy)	24
3.2.1.2	Multi-layer Perceptron (DoanAnhVu)	24
3.2.1.3	XGBoost Classifier (DoanAnhVu)	25
3.2.1.4	K-Nearest Neighbors (LeNgocTien)	25
3.2.1.5	Random Forest (NguyenHuuTuanPhat)	25
3.2.2	Nhóm mô hình sử dụng ảnh Spectrogram (Image Data)	25
3.2.2.1	CNN - Spectrogram (NguyenHuuTuanPhat)	26
3.2.2.2	CRNN - Spectrogram (NguyenHuuTuanPhat)	26
3.2.2.3	Swin Transformer - Transfer Learning (DoanAnhVu)	26
3.3	Tổng hợp và So sánh	26
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN		28
TÀI LIỆU THAM KHẢO		29

Danh mục các ký hiệu và chữ viết tắt

EEG Electroencephalogram (Điện não đồ)

UCI University of California, Irvine

ML Machine Learning

CNN Convolutional Neural Network

Danh sách bảng

3.1	Báo cáo phân loại - Logistic Regression	24
3.2	Ma trận nhầm lẫn - Logistic Regression	24
3.3	Báo cáo phân loại - MLP	25
3.4	Ma trận nhầm lẫn - MLP	25
3.5	Báo cáo phân loại - XGBoost	25
3.6	Báo cáo phân loại - KNN	25
3.7	Ma trận nhầm lẫn - KNN	25
3.8	Báo cáo phân loại - Random Forest	26
3.9	Báo cáo phân loại - CNN	26
3.10	Báo cáo phân loại - CRNN	26
3.11	Báo cáo phân loại - Swin Transformer	26
3.12	Bảng so sánh hiệu năng các mô hình (Weighted Avg)	27

Danh sách hình vẽ

LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện bài tập kết thúc môn Khai Phá Dữ Liệu, chúng tôi đã nhận được rất nhiều sự quan tâm, chỉ dẫn và hỗ trợ quý báu từ các thầy cô trong Phân hiệu Trường Đại học Thủy Lợi. Đây là nền tảng quan trọng giúp chúng tôi có thể tiếp thu, rèn luyện và vận dụng kiến thức vào thực tế, từ đó hoàn thành được bài tập này.

Đặc biệt, chúng tôi xin gửi lời cảm ơn sâu sắc đến Cô Vũ Thị Hạnh – giảng viên trực tiếp giảng dạy và hướng dẫn môn học. Cô không chỉ truyền đạt những kiến thức chuyên môn một cách rõ ràng, dễ hiểu mà còn tận tình giải đáp thắc mắc, định hướng phương pháp tiếp cận vấn đề, cũng như chia sẻ nhiều kinh nghiệm thực tiễn quý giá. Chính sự tận tâm và nhiệt huyết của Cô đã giúp chúng tôi có thêm động lực, sự tự tin và tinh thần trách nhiệm trong quá trình nghiên cứu và hoàn thiện bài tập.

Chúng tôi cũng xin cảm ơn Phân hiệu Trường Đại học Thủy Lợi đã cung cấp môi trường học tập và các cơ sở vật chất cần thiết. Xin cảm ơn tập thể lớp S26-65TTNT đã luôn đồng hành, chia sẻ và giúp đỡ lẫn nhau.

Mặc dù đã rất cố gắng, nhưng do hạn chế về thời gian và kiến thức, đồ án khó tránh khỏi những thiếu sót. Chúng tôi rất mong nhận được sự đóng góp ý kiến từ Cô và các bạn.

TP. Hồ Chí Minh, ngày 14 tháng 01 năm 2026

Trân trọng

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Thành phố Hồ Chí Minh, ngày 15 tháng 01 năm 2026

Ký và ghi rõ họ tên

MỞ ĐẦU

Nghiện rượu hiện là một trong những vấn đề sức khỏe cộng đồng nghiêm trọng trên toàn cầu, không chỉ gây ra các bệnh lý về gan, tim mạch mà còn để lại những di chứng nặng nề lên hệ thần kinh trung ương. Việc lạm dụng rượu trong thời gian dài dẫn đến suy giảm khả năng nhận thức, rối loạn trí nhớ và làm thay đổi cấu trúc cũng như chức năng hoạt động của não bộ. Chẩn đoán sớm và chính xác các tác động thần kinh này là yếu tố then chốt để đưa ra các phác đồ điều trị và phục hồi hiệu quả.

Điện não đồ, viết tắt là EEG, là một phương pháp thăm dò chức năng không xâm lấn, được sử dụng rộng rãi để ghi lại hoạt động điện sinh lý của não bộ. Bằng cách đặt các điện cực lên bề mặt da đầu, EEG thu thập các tín hiệu điện thế sinh ra từ sự giao tiếp giữa các tế bào thần kinh. Tín hiệu EEG mang thông tin phong phú về trạng thái hoạt động của não, được đặc trưng bởi các dải tần số khác nhau như Delta (0.5-4Hz), Theta (4-8Hz), Alpha (8-13Hz), Beta (13-30Hz) và Gamma (>30Hz).

Trong nghiên cứu y học, EEG được xem là "cửa sổ" để quan sát não bộ hoạt động theo thời gian thực. Đối với bệnh nhân nghiện rượu, các nghiên cứu đã chỉ ra sự bất thường trong tín hiệu EEG, đặc biệt là sự giảm đồng bộ hóa giữa các vùng não và sự thay đổi công suất trong các dải tần số cụ thể khi thực hiện các tác vụ nhận thức thị giác. Tuy nhiên, tín hiệu EEG thường rất phức tạp, đa chiều và chứa nhiều nhiễu, khiến việc phân tích thủ công bằng mắt thường trở nên khó khăn, tốn thời gian và phụ thuộc nhiều vào chủ quan của bác sĩ chuyên khoa.

Sự phát triển của Khai phá dữ liệu và Học máy đã mở ra hướng đi mới trong việc phân tích tín hiệu y sinh. Đồ án này tập trung vào việc áp dụng các kỹ thuật học máy tiên tiến để tự động hóa việc phân tích tín hiệu EEG, nhằm phân loại chính xác giữa nhóm người nghiện rượu và nhóm đối chứng (người khỏe mạnh). Mục tiêu của đồ án không chỉ dừng lại ở việc so sánh hiệu năng các mô hình, mà còn hướng tới việc xây dựng một hệ thống trực quan hỗ trợ các bác sĩ trong việc chẩn đoán, góp phần nâng cao chất lượng chăm sóc sức khỏe tâm thần.

Chương 1

TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT

1.1 Cơ sở lý thuyết về Điện não đồ (EEG)

Điện não đồ (EEG) ghi lại sự dao động điện thế của não bộ thông qua các điện cực gắn trên da đầu. Để đảm bảo tính nhất quán trong nghiên cứu lâm sàng, vị trí các điện cực thường được xác định theo hệ thống quốc tế 10-20.

1.1.1 Hệ thống vị trí điện cực 10-20

Các điện cực trong hệ thống 10-20 được phân bố trên toàn bộ da đầu để bao phủ các thùy não quan trọng. Mỗi vị trí điện cực được ký hiệu bằng một hoặc hai chữ cái (đại diện cho vùng não) và một con số (đại diện cho vị trí trái/phải).

- Số lẻ:** Bán cầu não Trái.
- Số chẵn:** Bán cầu não Phải.
- Z (Zero):** Đường giữa (Midline).

Dưới đây là chi tiết chức năng và đặc điểm tín hiệu của các nhóm điện cực có trong bộ dữ liệu SMNI_CMI:

1.1.1.1 Vùng Trán (Frontal - F)

Vùng trán liên quan chặt chẽ đến chức năng điều hành, ra quyết định, kiểm soát hành vi và cảm xúc.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
FPZ	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi (tùy vị trí trái/phải/đường giữa).	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Vùng trán cũng dễ bị ảnh hưởng bởi nháy mắt và nhiều cơ mặt.	Giảm hoạt động có thể gặp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng (tuỳ bối cảnh đo).
FP1	Vùng trước trán trái (prefrontal). Thường liên quan chú ý, điều hành hành vi và kiểm soát nhận thức.	Tăng hoạt động vùng trước trán thường xuất hiện khi tăng yêu cầu chú ý/điều hành; EEG vùng trán cũng dễ bị ảnh hưởng bởi nháy mắt/EMG.	Giảm hoạt động/dao động đặc trưng có thể liên quan giảm chú ý hoặc thay đổi trạng thái tinh túng (tuỳ bối cảnh).
FP2	Vùng trước trán phải (prefrontal). Thường liên quan chú ý, kiểm soát nhận thức và điều hòa cảm xúc.	Tăng hoạt động có thể gặp khi nhiệm vụ đòi hỏi tập trung/căng thẳng; cần cảnh giác nhiều cơ mặt và nháy mắt.	Giảm hoạt động có thể phản ánh giảm mức tham gia nhiệm vụ hoặc thay đổi trạng thái tinh túng.

Tiếp trang sau...

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
AF3	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
AF4	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F7	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F5	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F3	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F1	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
FZ	Đường giữa thùy trán (frontal midline). Hay dùng để quan sát hoạt động liên quan chú ý/kiểm soát nhận thức.	Tăng hoạt động vùng trán giữa thường gấp khi tăng tải nhận thức (cognitive control) tùy nhiệm vụ.	Giảm có thể xuất hiện khi thư giãn hoặc giảm yêu cầu nhiệm vụ.
F2	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F4	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.
F6	Nhóm điện cực vùng trán (frontal). Thường liên quan chú ý, kiểm soát nhận thức, lập kế hoạch và điều hành hành vi.	Tăng hoạt động có thể liên quan tăng tải nhận thức/chú ý hoặc căng thẳng. Dễ bị ảnh hưởng bởi nháy mắt.	Giảm hoạt động có thể gấp khi thư giãn/giảm yêu cầu nhiệm vụ hoặc thay đổi mức tinh túng.

Tiếp trang sau...

1.1.1.2 Vùng Trung tâm (Central - C)

Vùng trung tâm (C) và Trung tâm-Đỉnh (CP) nằm trên rãnh trung tâm, ngăn cách thùy trán và thùy đỉnh. Khu vực này chủ yếu liên quan đến chức năng vận động và cảm giác cơ thể (Sensorimotor).

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
C5	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác; cũng có thể bị nhiễu bởi EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
C3	Vùng trung tâm trái (central), gần vỏ vận động/cảm giác thân thể bên phải.	Tăng hoạt động liên quan vận động/cảm giác (hoặc chuẩn bị vận động) tùy bối cảnh; có thể nhạy với nhiều cơ.	Giảm có thể gặp khi nghỉ hoặc khi không có hoạt động vận động/cảm giác rõ.
C1	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác; cũng có thể bị nhiễu bởi EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CZ	Đường giữa vùng trung tâm, nằm gần vùng vận động/cảm giác thân thể hai bên.	Có thể tăng khi nhiệm vụ vận động/chuẩn bị vận động; cũng có thể phản ánh nhiều cơ.	Giảm khi nghỉ hoặc khi tín hiệu ít thành phần liên quan vận động.
C2	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác; cũng có thể bị nhiễu bởi EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
C4	Vùng trung tâm phải (central), gần vỏ vận động/cảm giác thân thể bên trái.	Tăng có thể liên quan vận động/cảm giác hoặc nhiều cơ vùng đầu.	Giảm có thể gặp khi nghỉ hoặc ít vận động.
C6	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP5	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP3	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP1	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
CPZ	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP2	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP4	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.
CP6	Nhóm điện cực vùng trung tâm (central/rolandic). Gần vùng vỏ vận động và cảm giác thân thể.	Tăng hoạt động có thể liên quan vận động/chuẩn bị vận động hoặc phản ứng cảm giác/nhiều EMG.	Giảm hoạt động có thể gặp khi nghỉ/ngừng vận động hoặc ít kích thích cảm giác.

1.1.1.3 Vùng Thái dương (Temporal - T)

Nằm hai bên thái dương, chịu trách nhiệm xử lý thông tin thính giác (nghe), ngôn ngữ (chủ yếu bán cầu trái) và trí nhớ.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
T7	Thùy thái dương trái (temporal). Thường liên quan thính giác/ngôn ngữ và trí nhớ tùy bối cảnh.	Tăng có thể liên quan hoạt động thính giác/ngôn ngữ hoặc nhiều cơ vùng thái dương (nhai, nói).	Giảm có thể gặp khi ít hoạt động liên quan thính giác/ngôn ngữ.
T8	Thùy thái dương phải (temporal). Thường liên quan thính giác và một số xử lý cảm xúc/nhận diện tuỳ bối cảnh.	Tăng có thể liên quan hoạt động thính giác hoặc nhiều cơ vùng thái dương.	Giảm có thể gặp khi ít hoạt động liên quan.
TP7	Nhóm điện cực vùng thái dương (temporal/temporo-parietal). Thường liên quan thính giác, ngôn ngữ và trí nhớ (tùy bên trái/phải).	Tăng hoạt động có thể liên quan xử lý thính giác/ngôn ngữ hoặc nhiều cơ (nhai, nói) vùng thái dương.	Giảm hoạt động có thể gặp khi ít hoạt động thính giác/ngôn ngữ hoặc thay đổi trạng thái.
TP8	Nhóm điện cực vùng thái dương (temporal/temporo-parietal). Thường liên quan thính giác, ngôn ngữ và trí nhớ (tùy bên trái/phải).	Tăng hoạt động có thể liên quan xử lý thính giác/ngôn ngữ hoặc nhiều cơ (nhai, nói) vùng thái dương.	Giảm hoạt động có thể gặp khi ít hoạt động thính giác/ngôn ngữ hoặc thay đổi trạng thái.

1.1.1.4 Vùng Đỉnh (Parietal - P)

Thùy đỉnh xử lý thông tin cảm giác (xúc giác, đau, nhiệt), nhận thức không gian và định hướng.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
P7	Nhóm điện cực vùng đỉnh (parietal/centro-parietal). Thường liên quan tích hợp cảm giác, chú ý không gian và xử lý thông tin đa giác quan.	Tăng hoạt động có thể liên quan tăng xử lý chú ý/cảm giác hoặc yêu cầu nhiệm vụ.	Giảm hoạt động có thể gặp khi thư giãn hoặc giảm yêu cầu xử lý cảm giác/chú ý.
P5	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
P3	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
P1	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PZ	Đường giữa thùy đỉnh (parietal). Hay liên quan xử lý cảm giác, chú ý không gian, tích hợp thông tin.	Tăng có thể liên quan tăng xử lý chú ý/cảm giác tùy nhiệm vụ.	Giảm có thể gặp khi thư giãn hoặc giảm yêu cầu xử lý.
P2	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
P4	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
P6	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
P8	Nhóm điện cực vùng đỉnh. Tương tự P7.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PO7	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PO5	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PO3	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
POZ	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PO4	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
PO6	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.
PO8	Nhóm điện cực vùng đỉnh-chẩm. Liên quan tích hợp cảm giác/thị giác.	Tăng khi xử lý chú ý/cảm giác.	Giảm khi thư giãn.

1.1.1.5 Vùng Chẩm (Occipital - O)

Nằm phía sau gáy, là trung tâm xử lý thị giác chính của não bộ.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
O1	Thùy chẩm trái (occipital), vùng thị giác. Hay thấy alpha rõ ở vùng chẩm khi nhắm mắt/thư giãn.	Khi nhắm mắt/thư giãn, alpha vùng chẩm thường tăng; khi nhìn/hoạt động thị giác có thể thay đổi.	Khi mở mắt/tập trung thị giác, alpha chẩm thường giảm.
OZ	Nhóm điện cực vùng chẩm (occipital/parieto-occipital). Liên quan xử lý thị giác; alpha vùng chẩm thường nổi bật khi nhắm mắt/thư giãn.	Alpha chẩm thường tăng khi nhắm mắt/thư giãn; hoạt động thị giác có thể thay đổi phổ biến theo trạng thái.	Khi mở mắt/tập trung thị giác, alpha vùng chẩm thường giảm.
O2	Thùy chẩm phải (occipital), vùng thị giác. Tương tự O1.	Nhắm mắt/thư giãn thường làm alpha vùng chẩm tăng.	Mở mắt/tập trung thị giác thường làm alpha giảm.

1.1.1.6 Các kênh đặc biệt (Special Channels)

Một số kênh trong bộ dữ liệu SMNI_CMI không thuộc chuẩn 10-20 hoặc là kênh tham chiếu/bổ sung.

Kênh	Tóm tắt chức năng	Khi tăng hoạt động	Khi giảm hoạt động
ND	Kênh "ND" xuất hiện trong dataset, không phải điện cực 10–20 chuẩn.	Không diễn giải theo vùng não. Coi là tín hiệu đặc biệt.	Không gán ý nghĩa y học. Có thể ẩn/loại khỏi phân tích scalp.
X	Kênh ký hiệu "X", không phải chuẩn 10-20.	Vì không có tọa độ scalp chuẩn, không diễn giải theo vùng não.	Không diễn giải theo vùng y học. Kiểm tra file nguồn nếu cần.
Y	Kênh ký hiệu "Y", không phải chuẩn 10-20.	Không có tọa độ scalp chuẩn để gán vào vùng não cụ thể.	Không diễn giải theo vùng.

Trong bài toán phân loại người nghiện rượu, các kênh vùng trán (F) và trung tâm (C) thường chứa nhiều thông tin quan trọng do rượu ảnh hưởng mạnh đến khả năng ức chế và điều khiển hành vi của thùy trán.

1.2 Mục tiêu nghiên cứu

Mục tiêu chính của đề tài là xây dựng một quy trình khép kín (pipeline) để phân loại tự động trạng thái nghiện rượu dựa trên tín hiệu EEG. Cụ thể:

1. Phân tích đặc điểm tín hiệu EEG của hai nhóm đối tượng: Nghiện rượu (Alcoholic) và Kiểm soát (Control).
2. Áp dụng các kỹ thuật Học máy và Học sâu để mô hình hóa dữ liệu.
3. Xây dựng ứng dụng Web minh họa, hỗ trợ bác sĩ tải lên tệp tín hiệu và nhận kết quả chẩn đoán nhanh chóng.

1.3 Bài toán đặt ra

Đây là bài toán **Phân loại nhị phân (Binary Classification)**.

- **Đầu vào:** Dữ liệu thô EEG đa kênh (64 kênh) từ bộ dữ liệu SMNI_CMI. Tín hiệu được thu thập theo thời gian từ các phiên đo (trial).
- **Đầu ra:** Nhận dự đoán $y \in \{0, 1\}$.
 - **0:** Nhóm đối chứng (Control - 'c').
 - **1:** Nhóm nghiện rượu (Alcoholic - 'a').

Chương 2

DỮ LIỆU VÀ PHƯƠNG PHÁP ĐỀ XUẤT

2.1 Tổng quan quy trình thực hiện

Hệ thống (Pipeline) được thiết kế theo luồng xử lý tuần tự: **Đọc dữ liệu** → **Tiền xử lý** → **Trích xuất đặc trưng** → **Huấn luyện** → **Đánh giá** → **Lưu model**. Mục tiêu cốt lõi là tạo ra các đặc trưng ổn định, dễ diễn giải và đảm bảo không có rò rỉ dữ liệu giữa tập huấn luyện (train) và kiểm thử (test). Chi tiết các bước như sau:

1. Bước 1: Đọc và Chuẩn hóa dữ liệu đầu vào

- Dữ liệu đầu vào: Các file CSV EEG, mỗi file chứa dữ liệu của 1 phiên đo (trial).
- Chuẩn hóa: Tên cột được chuyển về chữ thường và thay dấu cách bằng dấu gạch dưới (snake_case).
- Tạo nhãn: Ánh xạ nhãn 'a' (Alcoholic) → 1, 'c' (Control) → 0.
- Gom nhóm: Chuyển đổi dữ liệu bằng thành ma trận (kênh × thời gian) cho mỗi trial.

2. Bước 2: Xử lý dữ liệu thiếu và Chống rò rỉ (Date Leakage)

- **Xử lý missing:** Loại bỏ các kênh có tỷ lệ thiếu dữ liệu > 5%. Thực hiện nội suy (interpolate) theo thời gian, sau đó điền đầy (bfill/ffill).
- **Chống rò rỉ dữ liệu:** Tính mã băm MD5 cho mỗi file trong tập train và test. Loại bỏ ngay lập tức các file trong tập test nếu nội dung trùng khớp với bất kỳ file nào trong tập train. Điều này giúp kết quả đánh giá thực tế và khách quan hơn.

3. Bước 3: Trích xuất đặc trưng (Feature Extraction)

- **Welch PSD:** Ước lượng mật độ phổ công suất theo tần số.
- **Bandpower:** Tích phân PSD trong từng dải tần đặc trưng (Delta, Theta, Alpha, Beta, Gamma).
- **Ví dụ để diễn giải:** Feature bp_beta_FP1 là năng lượng dải Beta (13–30Hz) tại kênh FP1. Feature bp_alpha_02 là năng lượng dải Alpha (8–13Hz) tại kênh O2.
- **Ý nghĩa:** Mô hình không nhìn vào "tín hiệu thô" mà quan sát phân bố năng lượng theo dải tần ở từng vùng não.
- **Spectrogram:** (Cho Deep Learning) Tạo ảnh phổ PSD theo thời gian để quan sát sự biến thiên động.

4. Bước 4: Huấn luyện mô hình

- **Nhóm ML cổ điển:** Huấn luyện trên vector đặc trưng Bandpower. Bao gồm: Logistic Regression, Random Forest, KNN, MLP và XGBoost.
- **Nhóm Deep Learning:** Huấn luyện trên dữ liệu hình ảnh Spectrogram. Bao gồm: Custom CNN và EfficientNet (Transfer Learning).

5. Bước 5: Đánh giá và Lưu trữ Artifacts

- **Đánh giá:** Sử dụng các metrics (Accuracy, AUC) và Confusion Matrix.
- **Lưu trữ:** Model huấn luyện xong được lưu vào thư mục saved_models/.

2.1.1 Đặc tả kỹ thuật Đầu vào - Đầu ra

- **Input:** File CSV EEG thô (nhiều kênh), mỗi file chứa 1 trial + nhãn phân loại.
- **Output (ML Model):** Vector đặc trưng **Bandpower** có kích thước (số dải tần × số kênh).
- **Output (DL Model):** Ảnh **Spectrogram** có kích thước (tần số × thời gian).
- **Đầu ra hệ thống (Artifacts):**
 - Model files: .joblib (sklearn/xgboost) hoặc .keras (tensorflow).
 - Scaler: scaler_bp.joblib (dùng để chuẩn hóa dữ liệu mới).
 - Metadata: metrics.json (kết quả đánh giá), eda.json (thống kê dữ liệu).

2.2 Mô tả bộ dữ liệu SMNI_CMI

Bộ dữ liệu được tổ chức thành hai thư mục chính phục vụ cho quá trình huấn luyện và kiểm thử:

- **SMNI_CMI_TRAIN:** Tập dữ liệu dùng để huấn luyện mô hình.
- **SMNI_CMI_TEST:** Tập dữ liệu dùng để kiểm tra đánh giá.

Mỗi tệp tin trong bộ dữ liệu là một file CSV tương ứng với một phiên đo (trial), chứa thông tin của 64 kênh điện cực theo thời gian. Nguồn dữ liệu gốc được lấy từ **UCI Machine Learning Repository** (do Begleiter và cộng sự thu thập), bao gồm hai nhóm đối tượng chính là 'a' (Alcoholic) và 'c' (Control).

2.3 Phân tích số liệu và Thăm dò (EDA)

Trước khi huấn luyện mô hình, chúng tôi thực hiện phân tích dữ liệu khám phá (EDA) để hiểu rõ phân bố, phát hiện bất thường và lựa chọn chiến lược tiền xử lý phù hợp.

2.3.1 Phân bố nhãn và Kiểm tra rò rỉ (Data Leakage)

Dữ liệu được chia thành hai tập Train và Test. Chúng tôi đã kiểm tra sự cân bằng nhãn và rò rỉ dữ liệu giữa hai tập này.

- **Cân bằng nhãn:** Tỉ lệ giữa nhóm bệnh (Alcoholic - a) và nhóm chứng (Control - c) khá cân bằng.
 - **Train:** Class 'c': 233, Class 'a': 235 (Tổng 468 files).
 - **Test:** Class 'c': 240, Class 'a': 240 (Tổng 480 files).
- **Kiểm tra rò rỉ (Leakage):**
 - Tổng số file Train: 468.
 - Tổng số file Test: 480.
 - Số file Test có tên trùng với Train: 468 (trùng basename).
 - **Trùng nội dung (MD5 Check):** Phát hiện **24** files trong tập Test có nội dung hoàn toàn trùng khớp với tập Train. Các file này đã được loại bỏ khỏi quá trình đánh giá để đảm bảo tính khách quan.

2.3.2 Cấu hình khử nhiễu EOG (ICA)

Kỹ thuật khử nhiễu mắt (EOG) sử dụng ICA là một bước tùy chọn nâng cao. Trong các thực nghiệm cơ sở, bước này có thể được tắt (OFF) để đánh giá hiệu suất trên dữ liệu thô. Cấu hình chi tiết khi kích hoạt:

- **High-pass Filter:** 1.00 Hz (loại bỏ nhiễu tần số thấp và trôi dòng).
- **ICA Components:** 12 thành phần độc lập.
- **Cơ chế:** Tính tương quan (correlation) giữa các thành phần ICA và tín hiệu trung bình vùng trán (FP1, FP2, FZ...).
- **Ngưỡng (Threshold):** 0.35. Các thành phần có tương quan cao hơn ngưỡng này sẽ bị coi là nhiễu mắt và bị loại bỏ.

2.3.3 Phân tích ngoại lai và Phân phối đặc trưng (Outliers)

Chúng tôi đã phân tích thống kê các đặc trưng Bandpower trên tập Train để xác định chiến lược chuẩn hóa.

- **Skewness (Độ lệch):** Các kênh vùng trán và dải Gamma có độ lệch phân phối cao nhất (vd: bp_beta_FP1, bp_gamma_FP2, bp_alpha_FP1).
- **Outlier (IQR & Z-Score):** Nhiều giá trị ngoại lai được phát hiện ở các đặc trưng như bp_gamma_F8, bp_gamma_T8, bp_gamma_FP1.
- **Giải pháp:** Do sự hiện diện dày đặc của các giá trị ngoại lai (outliers), chúng tôi quyết định sử dụng **RobustScaler** (dựa trên trung vị và khoảng từ phân vị IQR) thay vì StandardScaler (dựa trên trung bình và phương sai) để chuẩn hóa dữ liệu đầu vào cho các mô hình Machine Learning.

2.3.4 Độ quan trọng đặc trưng (Feature Importance)

Dựa trên các mô hình cây quyết định (Random Forest và XGBoost) cùng mô hình tuyến tính (Logistic Regression), các đặc trưng sau đây được đánh giá là quan trọng nhất trong việc phân loại:

Feature	Độ quan trọng	Ý nghĩa
bp_theta_POZ	0.078289	Vùng Đỉnh-Chẩm, dải Theta.
bp_beta_AF2	0.031861	Vùng Trước trán, dải Beta.
bp_theta_P3	0.031101	Vùng Đỉnh trái, dải Theta.
bp_delta_P4	0.027445	Vùng Đỉnh phải, dải Delta.
bp_beta_FC6	0.027313	Vùng Trán-Trung tâm phải.
bp_theta_P4	0.025066	Vùng Đỉnh phải.
bp_gamma_FC6	0.023958	Vùng Trán-Trung tâm phải.
bp_beta_F1	0.020705	Vùng Trán trái.

Kết quả cho thấy các sóng chậm (Theta, Delta) ở vùng Đỉnh-Chẩm (P, PO, O) và sóng nhanh (Beta, Gamma) ở vùng Trán (F, AF, FC) đóng vai trò then chốt trong việc nhận diện đặc điểm nghiện rượu.

2.4 Quy trình tiền xử lý dữ liệu (Preprocessing)

Dựa trên mã nguồn thực tế của hệ thống (trong module `src.preprocessing`), quy trình xử lý dữ liệu thường được thực hiện qua các bước tuần tự nghiêm ngặt để đảm bảo chất lượng tín hiệu đầu vào cho các mô hình học máy.

2.4.1 Dữ liệu thô ban đầu (Raw Input)

Mỗi tệp tin đầu vào là một file CSV chứa dữ liệu của một lần đo (trial). Cấu trúc dữ liệu dạng "long-format" với các cột chính:

- `trial number`: Số thứ tự của phiên đo.
- `sensor position`: Tên kênh điện cực (ví dụ: FP1, CZ, P3...).
- `sample num`: Chỉ số thời gian của mẫu (0-255).
- `sensor value`: Giá trị điện thế (μV) đo được.
- `subject identifier`: Nhãn của đối tượng ('a': Alcoholic, 'c': Control).

2.4.2 Bước 1: Chuẩn hóa và Chuyển đổi (Normalization & Pivoting)

Hàm `normalize_long_df` thực hiện việc chuẩn hóa các tên cột và mã hóa nhãn:

- Nhãn dữ liệu được chuyển đổi: 'a' → 1 (Bệnh lý) và 'c' → 0 (Bình thường).
- Chuyển đổi cấu trúc bảng: Dữ liệu được xoay trực (pivot) từ dạng dọc sang dạng ma trận thời gian thực sử dụng hàm `build_epoch_matrix_from_long`. Kết quả là một ma trận X có kích thước ($N_{channels} \times N_{samples}$), trong đó $N_{samples}$ được cố định là 256 điểm (tương ứng 1 giây tín hiệu ở tần số 256Hz).

2.4.3 Bước 2: Xử lý kênh hỏng và giá trị thiêu (Cleaning)

Trong quá trình thu tín hiệu, một số điện cực có thể bị mất kết nối hoặc nhiễu. Hệ thống thực hiện các bước:

1. **Loại bỏ kênh lỗi**: Tính toán tỉ lệ giá trị thiêu (NaN) trên từng kênh. Các kênh có tỉ lệ mất dữ liệu lớn hơn 5% (> 0.05) sẽ bị loại bỏ khỏi ma trận.
2. **Nội suy (Interpolation)**: Các giá trị thiêu rải rác còn lại được lấp đầy bằng phương pháp nội suy tuyến tính (linear interpolation) theo trục thời gian, kết hợp với điền giá trị lùi (backward fill) và điền (forward fill) ở các biên.

2.4.4 Bước 3: Khử nhiễu EOG bằng thuật toán ICA (Artifact Removal)

Nhiễu do chuyển động mắt (EOG - Electrooculogram) là thành phần nhiễu lớn nhất trong EEG. Hàm `eog_clean_epoch_matrix` thực hiện quy trình khử nhiễu tự động như sau:

- **Lọc thông cao (Highpass Filtering)**: Tín hiệu được lọc với tần số cắt f_c (thường là 0.5-1Hz) để loại bỏ nhiễu trôi đường nền (baseline wander) trước khi đưa vào ICA.
- **Tạo tín hiệu EOG tham chiếu**: Tính trung bình tín hiệu của các kênh vùng trán (Frontal channels: FP1, FP2, AF7, AF8...) để tạo ra một bản mẫu nhiễu mắt đặc trưng.

- **Phân rã ICA (FastICA):** Sử dụng thuật toán FastICA để phân rã tín hiệu đa kênh thành các thành phần độc lập (Independent Components - ICs). Số lượng thành phần được thiết lập tối đa là 12 hoặc bằng số lượng kênh.
- **Loại bỏ thành phần nhiễu:** Tính độ tương quan Pearson giữa từng thành phần IC và tín hiệu EOG tham chiếu. Nếu hệ số tương quan $|r| \geq threshold$ (ngưỡng cài đặt), thành phần đó được xác định là nhiễu mắt và bị gán trọng số bằng 0.
- **Tái tạo tín hiệu:** Tín hiệu EEG sạch (Artifact-free EEG) được tái tạo lại từ các thành phần độc lập còn lại.

2.4.5 Bước 4: Xử lý ngoại lai và Chuẩn hóa đặc trưng (Advanced Outlier Handling)

Sau khi trích xuất đặc trưng Bandpower, dữ liệu thường chứa các giá trị ngoại lai (outliers) do nhiễu hoặc đặc điểm sinh lý bất thường. Chúng tôi áp dụng các kỹ thuật xử lý nâng cao:

1. **Kẹp giá trị ngoại lai (Outlier Clipping):** Để hạn chế ảnh hưởng của các giá trị cực đoan, kỹ thuật *quantile clipping* được áp dụng cho từng đặc trưng. Tất cả các giá trị nằm ngoài khoảng phân vị [1%, 99%] (tính trên tập Train) sẽ được gán (clip) về giá trị biên tương ứng (Lo_q và Hi_q).
2. **Kiểm tra dịch chuyển phân phối (Distribution Shift Check):** Trước khi đưa vào mô hình, chúng tôi thực hiện kiểm tra độ lệch phân phối giữa tập Train và Test thông qua so sánh giá trị trung bình (mean shift) và trực quan hóa không gian dữ liệu bằng PCA. Điều này giúp đảm bảo tập Test có đặc điểm tương đồng với tập Train.
3. **Chuẩn hóa (Normalization):** Sử dụng **RobustScaler** (thay vì MinMaxScaler hay StandardScaler) để chuẩn hóa các đặc trưng. RobustScaler sử dụng trung vị (median) và khoảng tứ phân vị (IQR) để thay đổi tỷ lệ dữ liệu, giúp mô hình bền vững hơn trước sự hiện diện của outliers.

2.4.6 Dữ liệu đầu ra sau tiền xử lý (Output)

Kết quả cuối cùng của quá trình tiền xử lý là hai định dạng dữ liệu sạch sẵn sàng cho huấn luyện:

- **Cho Machine Learning:** Vector đặc trưng Bandpower đã được kẹp biên (clipped) và chuẩn hóa (scaled).
- **Cho Deep Learning:** Ma trận Spectrogram sạch (đã log-transform) kích thước $64 \times N_{freq} \times N_{time}$.

2.5 Phương pháp trích xuất đặc trưng

Mặc dù EEG là chuỗi tín hiệu theo thời gian, nhưng các thông tin đặc trưng quan trọng nhất thường nằm ở miền tần số. Do đó, phương pháp tiếp cận của chúng tôi tập trung khai thác:

2.5.1 Đặc trưng miền Tần số (Frequency Domain)

Sử dụng phương pháp Welch để tính Mật độ phổ công suất (PSD). Từ đó tính toán **Bandpower** - năng lượng trung bình trong các dải tần não bộ đặc trưng: Delta, Theta, Alpha, Beta và Gamma. Đây là các đặc trưng số học mạnh mẽ cho các mô hình máy học dạng bảng.

2.5.2 Biểu diễn Thời gian - Tân số (Spectrogram)

Để quan sát sự biến thiên của năng lượng theo thời gian (điều mà Bandpower trung bình không thể hiện được), chúng tôi sử dụng **Spectrogram**. Phổ năng lượng được chuyển sang thang đo \log_{10} để làm nổi bật các đặc trưng nhỏ. Hình ảnh Spectrogram này là đầu vào lý tưởng cho các mô hình Học sâu (Deep Learning) như CNN.

2.6 Các mô hình máy học đề xuất

Dựa trên đặc thù của hai loại đặc trưng đã trích xuất, chúng tôi chia quá trình thực nghiệm thành hai nhóm mô hình riêng biệt:

2.6.1 Nhóm mô hình sử dụng đặc trưng dạng bảng (Bandpower/Time-domain)

Các mô hình này nhận đầu vào là vector đặc trưng số học (Numerical Features) bao gồm các giá trị Bandpower (Delta, Theta, Alpha, Beta, Gamma) và các thông kê thời gian. Dữ liệu được chuẩn hóa sử dụng **RobustScaler** để giảm ảnh hưởng của ngoại lai (outliers).

- **Logistic Regression:**
 - Solver: `liblinear` (tối ưu cho bộ dữ liệu nhỏ/trung bình).
 - Max Iterations: 5000.
 - Vai trò: Đóng vai trò là baseline tuyển tính đơn giản nhất.
- **K-Nearest Neighbors (KNN):**
 - Số láng giềng (k): 7.
 - Vai trò: Baseline phi tham số, dựa trên độ tương đồng cục bộ.
- **Random Forest Classifier:** Mô hình Ensemble dựa trên Bagging.
 - Số lượng cây ($n_estimators$): 300 (cấu hình cơ sở) hoặc được tối ưu qua GridSearch trong khoảng [200, 400].
 - Độ sâu cây (max_depth): Không giới hạn (None) hoặc giới hạn [10, 20] để kiểm soát overfitting.
 - Số luồng (n_jobs): -1 (sử dụng tối đa CPU).
- **XGBoost Classifier:** Sử dụng kỹ thuật Gradient Boosting để tối ưu hóa hàm mất mát (LogLoss). Các tham số được tinh chỉnh thông qua Grid Search:
 - $n_estimators$: [300, 500].
 - Learning rate: [0.05, 0.1] - Tốc độ học chậm giúp mô hình hội tụ tốt hơn.
 - Max Depth: [4, 6] - Giới hạn độ sâu.
 - Subsample: [0.8, 1.0] - Tỉ lệ lấy mẫu dữ liệu ngẫu nhiên cho mỗi cây.
 - Colsample_bytree: [0.8, 1.0] - Tỉ lệ lấy mẫu đặc trưng ngẫu nhiên.
- **Multi-layer Perceptron (MLP):** Mạng nơ-ron truyền thẳng (Feed-forward Neural Network) cho dữ liệu bảng.

- Kiến trúc: 2 lớp ẩn với số neurons lần lượt là (256, 128).
- Hàm kích hoạt: ReLU.
- Solver: Adam.
- Regularization (alpha): [1e-4, 1e-3].
- Max Iterations: 350-500.
- Early Stopping: True (dừng sớm nếu loss không giảm).

2.6.2 Nhóm mô hình sử dụng đặc trưng hình ảnh (Spectrogram)

Đầu vào là các ảnh phổ (Spectrogram) biểu diễn tín hiệu Time-Frequency dưới dạng hình ảnh 2D. Các giá trị biên độ phổ được chuyển sang thang đo Logarit và chuẩn hóa (Z-score normalization).

- **Custom CNN (Convolutional Neural Network):** Mô hình CNN tiêu chuẩn với 3 lớp tích chập (Conv2D) nhằm trích xuất đặc trưng không gian từ ảnh phổ.
 - Kiến trúc: 3 khối Conv-BatchNorm-ReLU-Pool.
 - Số filters: $32 \rightarrow 64 \rightarrow 128$ (hoặc tùy chỉnh 32-64).
 - Phân loại: Lớp Dense kết hợp Dropout để đưa ra dự đoán.
- **CRNN (Convolutional Recurrent Neural Network):** Mô hình lai ghép giữa mạng tích chập và mạng hồi quy, nhằm tận dụng khả năng trích xuất đặc trưng hình ảnh của CNN và khả năng ghi nhớ chuỗi thời gian của RNN.
 - **Feature Extractor:** Sử dụng các lớp Conv2D để trích xuất bản đồ đặc trưng từ Spectrogram.
 - **Temporal Modeler:** Đặc trưng sau đó được đưa vào lớp **Bidirectional LSTM** để học sự phụ thuộc theo thời gian (temporal dependencies) của các mẫu sóng não.
 - **Ưu điểm:** Phù hợp với dữ liệu EEG có cấu trúc thời gian-tần số phức tạp.
- **Swin Transformer (Transfer Learning):** Sử dụng kiến trúc Vision Transformer hiện đại dựa trên cơ chế cửa sổ trượt (Shifted Windows).
 - **Model:** swin-tiny-patch4-window7-224.
 - **Cơ chế:** Fine-tuning mô hình đã được huấn luyện trước trên ImageNet. Áp dụng chiến lược "Unfreeze" dần dần các tầng (Gradual Unfreezing) và sử dụng Cosine Decay Scheduler để tinh chỉnh trọng số phù hợp với ảnh phổ EEG.

Chương 3

THỰC NGHIỆM VÀ KẾT QUẢ ĐẠT ĐƯỢC

3.1 Thiết lập thực nghiệm

Quá trình đánh giá được thực hiện trên tập kiểm thử (Test Set) độc lập bao gồm **456 mẫu** (sau khi đã loại bỏ các mẫu trùng lặp và lỗi). Phân bố nhãn trong tập kiểm thử khá cân bằng:

- **Class 0 (Control)**: 223 mẫu.
- **Class 1 (Alcoholic)**: 233 mẫu.

Việc đánh giá được thực hiện khách quan dựa trên các chỉ số: Accuracy, Precision, Recall, F1-Score và ROC-AUC.

3.2 Kết quả chi tiết từng mô hình

3.2.1 Nhóm mô hình sử dụng đặc trưng Bandpower (Tabular Data)

Nhóm này sử dụng đầu vào là vector đặc trưng năng lượng phổ (bandpower) được trích xuất từ các dải tần số.

3.2.1.1 Logistic Regression (PhanTranTuonVy)

Độ chính xác (Accuracy): 96.27% | ROC-AUC: 0.9875

Đây là mô hình đạt kết quả tốt nhất, cho thấy tính tuyển tính cao của dữ liệu.

Bảng 3.1: Báo cáo phân loại - Logistic Regression

Nhóm	Precision	Recall	F1-Score	Support
Class 0 (c)	0.9518	0.9731	0.9623	223
Class 1 (a)	0.9737	0.9528	0.9631	233
Macro avg	0.9627	0.9629	0.9627	456
Weighted avg	0.9630	0.9627	0.9627	456

Bảng 3.2: Ma trận nhầm lẫn - Logistic Regression

	Dự đoán 0	Dự đoán 1
Thực tế 0	217 (TP)	6 (FP)
Thực tế 1	11 (FN)	222 (TN)

3.2.1.2 Multi-layer Perceptron (DoanAnhVu)

Độ chính xác: 94.96% | ROC-AUC: 0.9819

Mạng MLP xấp xỉ Logistic Regression, chứng tỏ khả năng học phi tuyển tốt.

Bảng 3.3: Báo cáo phân loại - MLP

Nhóm	Precision	Recall	F1-Score	Support
Class 0 (c)	0.9505	0.9462	0.9483	223
Class 1 (a)	0.9487	0.9528	0.9507	233
Macro avg	0.9496	0.9495	0.9495	456

Bảng 3.4: Ma trận nhầm lẩn - MLP

	Dự đoán 0	Dự đoán 1
Thực tế 0	211	12
Thực tế 1	11	222

3.2.1.3 XGBoost Classifier (DoanAnhVu)

Độ chính xác: 91.67% | ROC-AUC: 0.9683

Mô hình hoạt động tốt nhưng bị ảnh hưởng bởi nhiễu nên thấp hơn Logistic Regression.

Bảng 3.5: Báo cáo phân loại - XGBoost

Nhóm	Precision	Recall	F1-Score	Support
Class 0 (c)	0.8936	0.9417	0.9170	223
Class 1 (a)	0.9412	0.8927	0.9163	233
Weighted avg	0.9179	0.9167	0.9167	456

3.2.1.4 K-Nearest Neighbors (LeNgocTien)

Độ chính xác: 90.79% | ROC-AUC: 0.9701

KNN có độ chính xác tốt nhưng Precision lớp 1 thấp hơn Recall.

Bảng 3.6: Báo cáo phân loại - KNN

Nhóm	Precision	Recall	F1-Score	Support
Class 0 (c)	0.9502	0.8565	0.9009	223
Class 1 (a)	0.8745	0.9571	0.9139	233
Weighted avg	0.9115	0.9079	0.9076	456

Bảng 3.7: Ma trận nhầm lẩn - KNN

	Dự đoán 0	Dự đoán 1
Thực tế 0	191	32
Thực tế 1	10	223

3.2.1.5 Random Forest (NguyenHuuTuanPhat)

Độ chính xác: 84.65% | ROC-AUC: 0.9321

Random Forest có kết quả thấp nhất trong nhóm Bandpower.

3.2.2 Nhóm mô hình sử dụng ảnh Spectrogram (Image Data)

Nhóm này sử dụng đầu vào là ảnh phổ thời gian-tần số (Spectrogram).

Bảng 3.8: Báo cáo phân loại - Random Forest

Nhóm	Precision	Recall	F1-Score	Support
Class 0 (c)	0.8097	0.8969	0.8511	223
Class 1 (a)	0.8900	0.7983	0.8416	233
Weighted avg	0.8507	0.8465	0.8462	456

3.2.2.1 CNN - Spectrogram (NguyenHuuTuanPhat)

Độ chính xác: 83.33% | ROC-AUC: 0.9238

Mô hình CNN cơ bản cho kết quả ở mức khá.

Bảng 3.9: Báo cáo phân loại - CNN

Nhóm	Precision	Recall	F1-Score	Support
Class 0	0.8296	0.8296	0.8296	223
Class 1	0.8369	0.8369	0.8369	233
Weighted avg	0.8333	0.8333	0.8333	456

3.2.2.2 CRNN - Spectrogram (NguyenHuuTuanPhat)

Độ chính xác: 85.75% | ROC-AUC: 0.9384

Việc bổ sung lớp LSTM (CRNN) đã giúp cải thiện hiệu suất so với CNN thuần túy (tăng khoảng 2.4% Accuracy), khẳng định tầm quan trọng của thông tin chuỗi thời gian trong tín hiệu EEG.

Bảng 3.10: Báo cáo phân loại - CRNN

Nhóm	Precision	Recall	F1-Score	Support
Class 0	0.8465	0.8655	0.8559	223
Class 1	0.8684	0.8498	0.8590	233
Weighted avg	0.8577	0.8575	0.8575	456

3.2.2.3 Swin Transformer - Transfer Learning (DoanAnhVu)

Độ chính xác: 82.24% | ROC-AUC: 0.8955

Mô hình Swin Transformer, mặc dù là kiến trúc SOTA trong thị giác máy tính, nhưng chưa đạt kết quả cao trên tập dữ liệu này. Có thể do sự khác biệt miền (Domain Shift) giữa ảnh tự nhiên và Spectrogram là quá lớn, và lượng dữ liệu chưa đủ để Fine-tune hiệu quả các tầng Attention phức tạp.

Bảng 3.11: Báo cáo phân loại - Swin Transformer

Nhóm	Precision	Recall	F1-Score	Support
Class 0	0.8737	0.7444	0.8039	223
Class 1	0.7857	0.8970	0.8377	233
Weighted avg	0.8287	0.8224	0.8211	456

3.3 Tổng hợp và So sánh

Bảng dưới đây tóm tắt hiệu năng của tất cả các mô hình đã thực nghiệm:

Bảng 3.12: Bảng so sánh hiệu năng các mô hình (Weighted Avg)

Mô hình	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LogRegression	96.27%	0.9630	0.9627	0.9627	0.9875
MLP (ANN)	94.96%	0.9496	0.9496	0.9496	0.9819
XGBoost	91.67%	0.9179	0.9167	0.9167	0.9683
KNN	90.79%	0.9115	0.9079	0.9076	0.9701
CRNN (Spectrogram)	85.75%	0.8577	0.8575	0.8575	0.9384
CNN (Spectrogram)	83.33%	0.8333	0.8333	0.8333	0.9238
Random Forest	84.65%	0.8507	0.8465	0.8462	0.9321
Swin Transformer	82.24%	0.8287	0.8224	0.8211	0.8955

Nhận xét chung: Các mô hình máy học cổ điển (Logistic Regression, MLP, KNN, XGBoost) sử dụng đặc trưng Bandpower vẫn giữ vị thế dẫn đầu với độ chính xác trên 90%. Trong nhóm Deep Learning, mô hình CRNN cho kết quả tốt nhất (85.75%), vượt qua CNN thường và Swin Transformer, cho thấy việc kết hợp thông tin không gian (CNN) và thời gian (LSTM) là hướng đi đúng đắn cho dữ liệu EEG dạng ảnh phổ. Swin Transformer dù mạnh mẽ nhưng chưa thích nghi tốt với tập dữ liệu nhỏ.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Dự án đã xây dựng thành công quy trình xử lý và phân loại tín hiệu EEG toàn diện để nhận diện trạng thái nghiện rượu (Alcoholism) từ bộ dữ liệu SMNI_CMI. Thông qua thực nghiệm trên 7 mô hình học máy và học sâu, chúng tôi rút ra các kết luận chính sau:

- **Hiệu quả của đặc trưng Bandpower:** Phương pháp trích xuất đặc trưng truyền thống (Bandpower) kết hợp với các mô hình máy học cổ điển (Logistic Regression, MLP, XGBoost) đạt hiệu suất vượt trội (Accuracy > 90%, ROC-AUC > 0.96) so với phương pháp Học sâu trên ảnh Spectrogram.
- **Logistic Regression và MLP dẫn đầu:** Với độ chính xác 96.27% và 94.96%, hai mô hình này cho thấy tín hiệu EEG sau khi được trích xuất đặc trưng năng lượng (Delta, Theta, Alpha, Beta, Gamma) có tính phân tách tuyến tính rất cao.
- **Hạn chế của Deep Learning trên tập dữ liệu nhỏ:** Với kích thước tập dữ liệu khoảng 900 mẫu, các mô hình CNN và EfficientNet (Transfer Learning) chưa thể phát huy tối đa khả năng học trích xuất đặc trưng tự động, dẫn đến kết quả thấp hơn (75% - 85%).
- **Đặc điểm sinh lý thần kinh:** Kết quả phân tích EDA và độ quan trọng đặc trưng (Feature Importance) chỉ ra rằng sự thay đổi sóng điện não tập trung mạnh ở các dải tần thấp (Delta, Theta) tại vùng vỏ não thị giác (Chẩm - Occipital) và trung tâm (Central), phù hợp với các nghiên cứu y văn về tác động của cồn lên não bộ.

2. Đóng góp của đề tài

- Xây dựng được pipeline xử lý dữ liệu chuẩn hóa: từ khâu làm sạch, xử lý nhiễu mắt (EOG Artifacts) bằng ICA, trị ngoại lai (Outlier Clipping) đến trích xuất đặc trưng.
- Giải quyết triệt để vấn đề rò rỉ dữ liệu (Data Leakage) thường gặp trong các bài toán y sinh, đảm bảo kết quả đánh giá là trung thực và tin cậy.
- Tích hợp thành công các mô hình vào ứng dụng Web (Sử dụng FastAPI và ReactJS), cho phép tải lên file EEG và nhận kết quả chẩn đoán trực quan trong thời gian thực.

3. Hướng phát triển

Để nâng cao hơn nữa độ chính xác và tính ứng dụng thực tiễn, nhóm đề xuất các hướng nghiên cứu tiếp theo:

- **Mở rộng dữ liệu:** Thu thập thêm dữ liệu từ nhiều nguồn khác nhau để tăng số lượng mẫu huấn luyện, tạo điều kiện cho các mô hình Deep Learning (như EEGNet, Conformer) phát huy hiệu quả.
- **Khai thác đặc trưng kết nối (Connectivity):** Bên cạnh bandpower (năng lượng cục bộ), việc nghiên cứu sự đồng bộ tín hiệu giữa các vùng não (Coherence, Phase Locking Value - PLV) có thể mang lại những đặc trưng mới giúp phân biệt rõ nét hơn trạng thái bệnh lý.

- **Tối ưu hóa thời gian thực:** Triển khai mô hình nén (Model Distillation) để chạy trực tiếp trên các thiết bị đeo EEG cá nhân (Wearable EEG) hoặc điện thoại thông minh, hỗ trợ giám sát và cảnh báo sớm.

TÀI LIỆU THAM KHẢO

- [1] Begleiter, H. (1999). *EEG Database*. UCI Machine Learning Repository. [<http://archive.ics.uci.edu/ml/datasets/EEG+Database>]. Irvine, CA: University of California, School of Information and Computer Science.
- [2] Begleiter, H., Porjesz, B., Bihari, B., & Kissin, B. (1984). "Event-related brain potentials in boys at risk for alcoholism". *Science*, 225(4669), 1493-1496.
- [3] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [4] FastAPI Documentation. [<https://fastapi.tiangolo.com/>]. Truy cập ngày 14/01/2026.
- [5] Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.
- [6] React Documentation. [<https://reactjs.org/>]. Truy cập ngày 14/01/2026.
- [7] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., & Litke, A. (1995). "Event related potentials during object recognition tasks". *Brain Topography*, 9(1), 1-15.