**ĐẠI HỌC QUỐC GIA HÀ NỘI TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**Nguyễn Ngọc Thái**

**NGHIÊN CỨU BÀI TOÁN PHÁT HIỆN ĐỘNG ĐẤT SỬ DỤNG DỮ LIỆU CẢM BIẾN GIA TỐC**

**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Công nghệ Kỹ thuật Cơ điện tử**

**HÀ NỘI - 2025**

**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Nguyễn Ngọc Thái**

**NGHIÊN CỨU BÀI TOÁN PHÁT HIỆN ĐỘNG ĐẤT SỬ DỤNG DỮ LIỆU CẢM BIẾN GIA TỐC**

**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Công nghệ Kỹ thuật Cơ điện tử**

**Cán bộ hướng dẫn : TS. Hoàng Văn Mạnh**

**Cán bộ đồng hướng dẫn:**

**HÀ NỘI - 2025**

## I. ****MỞ ĐẦU****

### 1.1. Lý do chọn đề tài

* Tình hình thiên tai động đất tại Việt Nam và thế giới

## Tình hình động đất tại Việt Nam

### Hoạt động địa chấn gần đây

* **Tháng 3/2025**: Việt Nam ghi nhận **26 trận động đất** với độ lớn từ 2,5 đến 3,7, tập trung chủ yếu tại huyện Kon Plông, tỉnh Kon Tum (23 trận) và huyện Nam Trà My, tỉnh Quảng Nam (3 trận). Đặc biệt, ngày 31/3, huyện Kon Plông xảy ra 5 trận động đất liên tiếp với độ lớn từ 2,6 đến 3,1. ​[Báo điện tử Tiền Phong+2https://baomoi.com+2Báo Sức khỏe & Đời sống+2](https://baomoi.com/viet-nam-xay-ra-26-tran-dong-dat-trong-thang-3-c51861990.epi?utm_source=chatgpt.com)
* **Ngày 28/3/2025**: Mặc dù chỉ ghi nhận một trận động đất có độ lớn 3,0 tại Kon Plông, nhưng do ảnh hưởng từ trận động đất mạnh 7,7 độ tại Myanmar, người dân ở một số tòa nhà cao tầng tại Hà Nội và TP.HCM cảm nhận rõ rung lắc. ​[Báo điện tử Tiền Phong+7Báo Sức khỏe & Đời sống+7https://baomoi.com+7](https://suckhoedoisong.vn/viet-nam-xay-ra-26-tran-dong-dat-trong-thang-3-169250401111949296.htm?utm_source=chatgpt.com)

### Các trận động đất đáng chú ý trong thế kỷ qua

* **Năm 1935**: Trận động đất mạnh 6,8 độ xảy ra tại lòng chảo Điện Biên.​[Báo điện tử Tiền Phong+1Báo điện tử Tiền Phong+1](https://tienphong.vn/dong-dat-ha-noi-tag78493.tpo?utm_source=chatgpt.com)
* **Năm 1983**: Động đất mạnh 6,7 độ tại Tuần Giáo, Điện Biên.​[Báo điện tử Dân Trí+3Báo điện tử Tiền Phong+3Báo điện tử Dân Trí+3](https://tienphong.vn/nhung-tran-dong-dat-lon-nhat-o-viet-nam-the-ky-qua-post1729297.tpo?utm_source=chatgpt.com)
* **Năm 2001**: Trận động đất 5,3 độ tại TP. Điện Biên Phủ gây thiệt hại khoảng 210 tỷ đồng.​[Báo điện tử Dân Trí+3Báo điện tử Tiền Phong+3thanhnien.vn+3](https://tienphong.vn/nhung-tran-dong-dat-lon-nhat-o-viet-nam-the-ky-qua-post1729297.tpo?utm_source=chatgpt.com)
* **Năm 2020**: Ngày 27/7, động đất mạnh 5,3 độ tại huyện Mộc Châu, Sơn La, gây rung lắc cho nhiều tỉnh miền Bắc, bao gồm Hà Nội. Hàng trăm căn nhà tại vùng tâm chấn bị hư hại. ​[Báo điện tử Tiền Phong+1Báo điện tử Tiền Phong+1](https://tienphong.vn/nhung-tran-dong-dat-lon-nhat-o-viet-nam-the-ky-qua-post1729297.tpo?utm_source=chatgpt.com)

### Nguy cơ và cảnh báo

Theo các chuyên gia, Việt Nam nằm trong khu vực có hoạt động địa chấn, đặc biệt tại các tỉnh Tây Bắc và miền Trung. Việc theo dõi và cập nhật thông tin về động đất là cần thiết để có biện pháp phòng ngừa và ứng phó kịp thời. ​

## Tình hình động đất trên thế giới

### Trận động đất tại Myanmar (28/3/2025)

* **Cường độ**: 7,7 độ.​[YouTube+8TUOI TRE ONLINE+8Báo điện tử Dân Trí+8](https://tuoitre.vn/dong-dat-myanmar-thoi-gian-vang-da-het-20250402075108925.htm?utm_source=chatgpt.com)
* **Thiệt hại**: Hơn 2.700 người thiệt mạng, hàng nghìn người bị thương và mất tích. Hơn 3.000 tòa nhà bị hư hại, bao gồm nhiều di sản văn hóa quan trọng. ​[Báo điện tử Dân Trí+4TUOI TRE ONLINE+4TUOI TRE ONLINE+4](https://tuoitre.vn/dong-dat-myanmar-thoi-gian-vang-da-het-20250402075108925.htm?utm_source=chatgpt.com)
* **Ảnh hưởng khu vực**: Rung chấn từ trận động đất này được cảm nhận tại nhiều quốc gia Đông Nam Á như Thái Lan, Việt Nam, Lào. ​[Báo điện tử Dân Trí+1Báo điện tử Tiền Phong+1](https://dantri.com.vn/khoa-hoc/nhin-lai-tham-hoa-dong-dat-vua-xay-ra-o-myanmar-qua-bieu-do-dia-chan-20250329061322083.htm?utm_source=chatgpt.com)

### Cảnh báo siêu động đất tại Nhật Bản

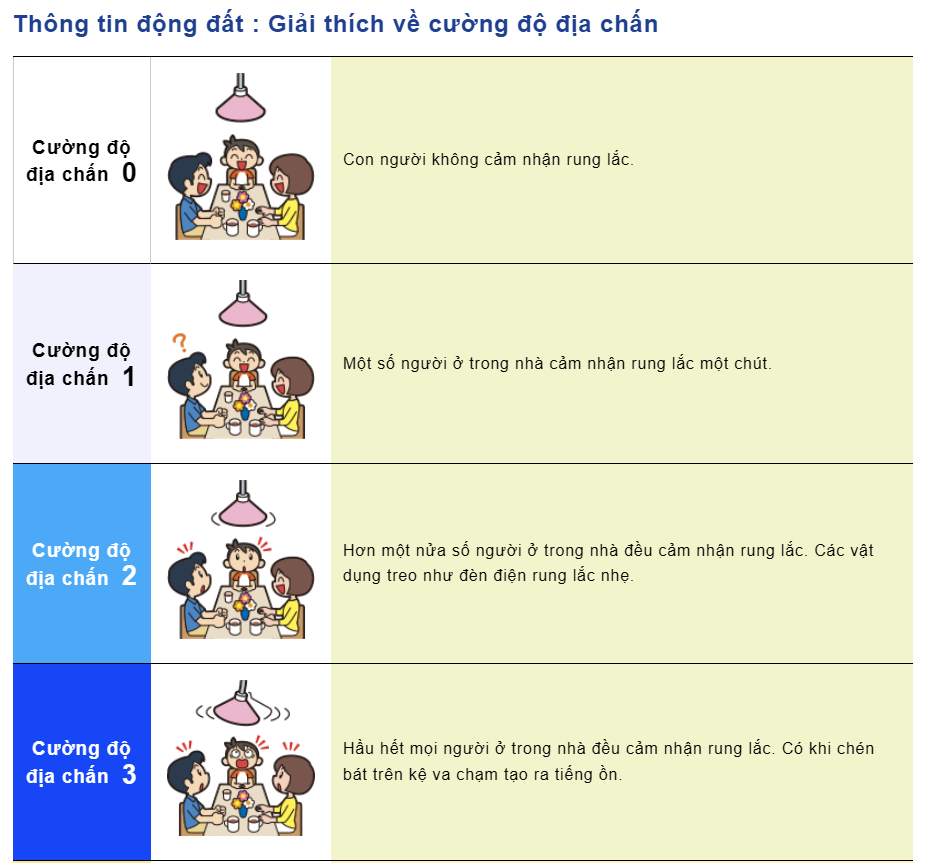
Chính phủ Nhật Bản cảnh báo về nguy cơ xảy ra siêu động đất tại rãnh Nankai, có thể gây thiệt hại nghiêm trọng:​[TUOI TRE ONLINE](https://tuoitre.vn/nhat-ban-canh-bao-sieu-dong-dat-khien-gan-300-000-nguoi-chet-thiet-hai-1-800-ti-usd-20250331161754589.htm?utm_source=chatgpt.com)

* **Dự báo thiệt hại**: Gần 300.000 người thiệt mạng, thiệt hại kinh tế lên tới 1.800 tỷ USD.​[TUOI TRE ONLINE](https://tuoitre.vn/nhat-ban-canh-bao-sieu-dong-dat-khien-gan-300-000-nguoi-chet-thiet-hai-1-800-ti-usd-20250331161754589.htm?utm_source=chatgpt.com)
* **Khu vực ảnh hưởng**: 31 tỉnh của Nhật Bản có thể chịu rung lắc mạnh và sóng thần trên 3m. ​[TUOI TRE ONLINE](https://tuoitre.vn/nhat-ban-canh-bao-sieu-dong-dat-khien-gan-300-000-nguoi-chet-thiet-hai-1-800-ti-usd-20250331161754589.htm?utm_source=chatgpt.com)

### Các trận động đất khác

Động đất là hiện tượng tự nhiên xảy ra thường xuyên trên toàn cầu, đặc biệt tại các khu vực nằm trên vành đai lửa Thái Bình Dương như Nhật Bản, Indonesia, và khu vực Andes ở Nam Mỹ. Việc theo dõi và nghiên cứu về động đất giúp giảm thiểu thiệt hại và nâng cao khả năng ứng phó của cộng đồng.​

**Lưu ý**: Thông tin về động đất có thể thay đổi theo thời gian. Để cập nhật thông tin mới nhất, nên theo dõi các cơ quan chức năng và trang tin tức uy tín.





* Vai trò của việc phát hiện sớm động đất
* Những hạn chế của hệ thống cảnh báo truyền thống
* Tiềm năng ứng dụng Machine Learning vào phân loại tín hiệu động đất

### 1.2. Mục tiêu nghiên cứu

* Xây dựng mô hình học máy để phân biệt tín hiệu động đất và nhiễu
* Đánh giá hiệu quả mô hình với các dữ liệu thực tế
* Triển khai hệ thống có khả năng dự đoán tự động qua giao diện người dùng

### 1.3. Đối tượng và phạm vi nghiên cứu

* Dữ liệu: cảm biến gia tốc 3 trục (X, Y, Z)
* Mô hình học sâu (CRNN - Convolutional Recurrent Neural Network)
* Chỉ phân loại 2 lớp: **Động đất (EQ)** và **Nhiễu (Noise)**
* Không bao gồm dự báo cường độ hay vị trí địa lý

### 1.4. Phương pháp nghiên cứu

* Xử lý tín hiệu thời gian
* Trích xuất đặc trưng hoặc đưa thẳng dữ liệu thô
* Huấn luyện mô hình học máy/học sâu
* Đánh giá bằng các chỉ số: Accuracy, F1-score, Confusion Matrix

Động đất, một trong những hiện tượng tự nhiên có sức tàn phá khủng khiếp nhất, luôn là mối đe dọa thường trực đối với sự an toàn của con người và sự ổn định của xã hội. Sự rung chuyển đột ngột của vỏ Trái Đất không chỉ gây ra những tổn thất nặng nề về người và tài sản mà còn để lại những hậu quả lâu dài về mặt kinh tế, xã hội và môi trường. Việc nghiên cứu và phát triển các phương pháp phát hiện động đất sớm, hiệu quả là vô cùng quan trọng để giảm thiểu tối đa những thiệt hại do thảm họa này gây ra.

Theo thống kê từ Cục Khảo sát Địa chất Hoa Kỳ (USGS), mỗi năm trên thế giới xảy ra khoảng 20.000 trận động đất, trong đó có khoảng 18 trận động đất mạnh từ 7.0 độ Richter trở lên. Chỉ tính riêng trong thế kỷ 21, đã có hàng triệu người thiệt mạng do động đất. Trận động đất Tohoku năm 2011 tại Nhật Bản, với cường độ 9.0 độ Richter, đã gây ra sóng thần tàn phá nặng nề, cướp đi sinh mạng của hơn 18.000 người và gây thiệt hại kinh tế ước tính lên đến 360 tỷ USD. Hay trận động đất năm 2010 tại Haiti, dù có cường độ thấp hơn (7.0 độ Richter), nhưng do điều kiện xây dựng kém, đã làm hơn 200.000 người thiệt mạng. Những con số này là minh chứng rõ ràng cho sức tàn phá khủng khiếp của động đất và tầm quan trọng của việc cảnh báo sớm.

Tại Việt Nam, mặc dù nằm trong khu vực có hoạt động địa chấn không quá mạnh so với Vành đai lửa Thái Bình Dương, nhưng vẫn tiềm ẩn nguy cơ xảy ra động đất ở một số khu vực như Tây Bắc, Tây Nguyên và ven biển miền Trung. Theo Viện Vật lý Địa cầu, trong vòng 100 năm qua, Việt Nam đã ghi nhận nhiều trận động đất có cường độ từ 5.0 đến 6.8 độ Richter. Trận động đất Điện Biên năm 1935, với cường độ 6.8 độ Richter, đã gây ra thiệt hại đáng kể cho các công trình xây dựng trong khu vực. Gần đây, các trận động đất nhỏ xảy ra tại Kon Tum, Lai Châu cũng cho thấy hoạt động địa chấn vẫn diễn ra và cần được theo dõi sát sao.

Động đất không chỉ gây ra những thiệt hại trực tiếp về người và tài sản mà còn ảnh hưởng sâu sắc đến các hoạt động kinh tế, xã hội. Các công trình giao thông, hệ thống điện, nước, các cơ sở y tế và giáo dục có thể bị phá hủy, gây ra sự gián đoạn trong hoạt động sản xuất và sinh hoạt của người dân. Hơn nữa, động đất còn gây ra những tác động tiêu cực đến môi trường như sạt lở đất, lũ quét, ô nhiễm nguồn nước, ảnh hưởng đến hệ sinh thái và đa dạng sinh học. Theo Ngân hàng Thế giới, thiệt hại do thiên tai, trong đó có động đất, có thể làm giảm GDP của một quốc gia tới 1% mỗi năm.

Trong bối cảnh đó, việc phát hiện sớm động đất là vô cùng quan trọng để đưa ra cảnh báo kịp thời, giúp người dân có thời gian sơ tán và chuẩn bị ứng phó. Các hệ thống cảnh báo động đất hiện tại ở Việt Nam, mặc dù đã có những bước tiến đáng kể, vẫn còn nhiều hạn chế về độ chính xác và khả năng phủ sóng. Hệ thống trạm quan trắc địa chấn hiện tại còn thưa thớt, đặc biệt là ở các khu vực có nguy cơ động đất cao. Do đó, việc nghiên cứu và phát triển các hệ thống cảnh báo động đất hiệu quả hơn, với khả năng phủ sóng rộng hơn và độ chính xác cao hơn, là một nhu cầu cấp thiết.

Trong những năm gần đây, sự phát triển của công nghệ cảm biến, đặc biệt là cảm biến gia tốc, đã mở ra những hướng đi mới trong việc phát hiện động đất. Cảm biến gia tốc có khả năng ghi nhận những dao động nhỏ nhất của mặt đất, từ đó cung cấp dữ liệu quan trọng để phân tích và nhận diện dấu hiệu của động đất. Với sự phổ biến của các thiết bị di động thông minh, việc tích hợp cảm biến gia tốc vào hệ thống cảnh báo động đất có thể tạo ra một mạng lưới giám sát rộng khắp, giúp chúng ta phát hiện động đất một cách nhanh chóng và chính xác hơn. Theo một nghiên cứu của Đại học Stanford, việc sử dụng mạng lưới cảm biến gia tốc trên điện thoại thông minh có thể giúp phát hiện động đất với độ chính xác tương đương với các trạm quan trắc địa chấn truyền thống.

Đề tài "Sử dụng dữ liệu cảm biến để phát hiện động đất" tập trung vào việc nghiên cứu khả năng ứng dụng của cảm biến gia tốc trong việc phát hiện động đất. Mục tiêu chính của đề tài là xây dựng một hệ thống có khả năng phân tích dữ liệu cảm biến gia tốc để nhận diện và cảnh báo sớm động đất. Việc nghiên cứu đề tài này không chỉ có ý nghĩa khoa học mà còn mang tính ứng dụng thực tiễn cao, góp phần vào việc nâng cao khả năng cảnh báo và giảm thiểu thiệt hại do động đất gây ra.

Cảm biến gia tốc là một thiết bị đo lường gia tốc, tức là tốc độ thay đổi của vận tốc. Chúng được sử dụng rộng rãi trong nhiều lĩnh vực như công nghiệp, ô tô, điện tử tiêu dùng và nghiên cứu khoa học. Trong nghiên cứu động đất, cảm biến gia tốc được sử dụng để ghi nhận và phân tích dao động của mặt đất. Dữ liệu từ cảm biến gia tốc có thể cung cấp thông tin về cường độ, tần số và hướng của sóng địa chấn, từ đó giúp các nhà khoa học xác định được vị trí và cường độ của động đất. Các loại cảm biến gia tốc MEMS (Micro-Electro-Mechanical Systems) ngày càng trở nên phổ biến do kích thước nhỏ gọn, chi phí thấp và độ nhạy cao.

Việc sử dụng cảm biến gia tốc trong hệ thống cảnh báo động đất có nhiều ưu điểm. Thứ nhất, cảm biến gia tốc có chi phí thấp và dễ dàng triển khai, cho phép xây dựng một mạng lưới giám sát rộng khắp. Thứ hai, cảm biến gia tốc có độ nhạy cao, có thể ghi nhận được những dao động nhỏ nhất của mặt đất. Thứ ba, cảm biến gia tốc có khả năng thu thập dữ liệu thời gian thực, cho phép phát hiện động đất một cách nhanh chóng. Theo một nghiên cứu của Đại học California, Berkeley, việc sử dụng cảm biến gia tốc trên điện thoại thông minh có thể giúp phát hiện động đất trong vòng vài giây sau khi xảy ra.

Tuy nhiên, việc sử dụng cảm biến gia tốc trong việc phát hiện động đất cũng gặp phải một số thách thức. Một trong những thách thức lớn nhất là nhiễu tín hiệu. Dữ liệu từ cảm biến gia tốc có thể bị ảnh hưởng bởi nhiều yếu tố như rung động của các phương tiện giao thông, hoạt động xây dựng và các nhiễu điện từ. Do đó, cần có các phương pháp xử lý tín hiệu hiệu quả để loại bỏ nhiễu và trích xuất thông tin hữu ích. Theo một nghiên cứu của Viện Công nghệ Massachusetts (MIT), việc sử dụng các thuật toán học máy có thể giúp cải thiện độ chính xác của việc phát hiện động đất từ dữ liệu cảm biến gia tốc.

Trong tương lai, việc nghiên cứu và phát triển các hệ thống cảnh báo động đất dựa trên cảm biến gia tốc cần tập trung vào việc cải thiện độ chính xác và độ tin cậy của hệ thống. Cần có các phương pháp xử lý dữ liệu thông minh hơn để phân biệt giữa các dao động do động đất và các dao động do các nguồn khác. Đồng thời, cần nghiên cứu và phát triển các thuật toán cảnh báo sớm hiệu quả hơn, giúp người dân có thời gian sơ tán và chuẩn bị ứng phó kịp thời. Việc kết hợp dữ liệu từ nhiều nguồn cảm biến khác nhau, như GPS, địa chấn kế và cảm biến áp suất, cũng có thể giúp nâng cao độ chính xác của việc phát hiện động đất.

**CHƯƠNG 1: GIỚI THIỆU**

1.1 **Lý do chọn đề tài**

* Tầm quan trọng của phát hiện động đất.
* Những thiệt hại do động đất gây ra.
* Ứng dụng của cảm biến gia tốc trong việc phát hiện động đất.

1.2 **Mục tiêu đề tài**

* Thu thập dữ liệu từ cảm biến MPU6050.
* Xây dựng mô hình phát hiện động đất bằng Machine Learning.
* Đánh giá độ chính xác của mô hình.

1.3 **Phạm vi nghiên cứu**

* Sử dụng dữ liệu thu được từ cảm biến MPU6050.
* Áp dụng các thuật toán Machine Learning cơ bản.
* Không nghiên cứu dự báo động đất trong tương lai.

1.4 **Phương pháp nghiên cứu**

* Thu thập dữ liệu cảm biến MPU6050.
* Tiền xử lý dữ liệu.
* Huấn luyện mô hình Machine Learning.
* Đánh giá mô hình bằng các tiêu chí phù hợp.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

2.1 **Tổng quan về động đất**

* Khái niệm về động đất.
* Nguyên nhân và cách thức xảy ra động đất.
* Các phương pháp phát hiện động đất hiện nay.

2.2 **Giới thiệu cảm biến MPU6050**

* Nguyên lý hoạt động của MPU6050.
* Thông số kỹ thuật của MPU6050.
* Cách thu thập dữ liệu từ MPU6050.

2.3 **Machine Learning trong phát hiện động đất**

* Giới thiệu về Machine Learning.
* Các thuật toán có thể áp dụng (SVM, Random Forest, Neural Networks,...).
* Các tiêu chí đánh giá mô hình (Accuracy, Precision, Recall, F1-score).

**CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG**

3.1 **Tổng quan hệ thống**

* Mô hình phát hiện động đất.
* Các thành phần của hệ thống.

3.2 **Thu thập dữ liệu cảm biến**

* Cấu trúc dữ liệu thu thập từ MPU6050.
* Thời gian và tần suất lấy mẫu dữ liệu.
* Các yếu tố ảnh hưởng đến dữ liệu.

3.3 **Tiền xử lý dữ liệu**

* Lọc nhiễu dữ liệu cảm biến.
* Chuẩn hóa và xử lý dữ liệu bị thiếu.
* Chuyển đổi dữ liệu về định dạng phù hợp cho mô hình.

3.4 **Lựa chọn thuật toán Machine Learning**

* So sánh các thuật toán tiềm năng.
* Lựa chọn thuật toán phù hợp nhất.

3.5 **Thiết kế mô hình phát hiện động đất**

* Mô tả kiến trúc mô hình.
* Quy trình huấn luyện và kiểm thử.

**CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT QUẢ**

4.1 **Huấn luyện mô hình**

* Chia tập dữ liệu (Train, Validation, Test).
* Quá trình huấn luyện mô hình.

4.2 **Đánh giá mô hình**

* Độ chính xác của mô hình trên tập kiểm thử.
* Phân tích kết quả dự đoán đúng và sai.

4.3 **Thảo luận kết quả**

* So sánh mô hình với các phương pháp khác.
* Những ưu điểm và hạn chế của mô hình.

**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

5.1 **Kết luận**

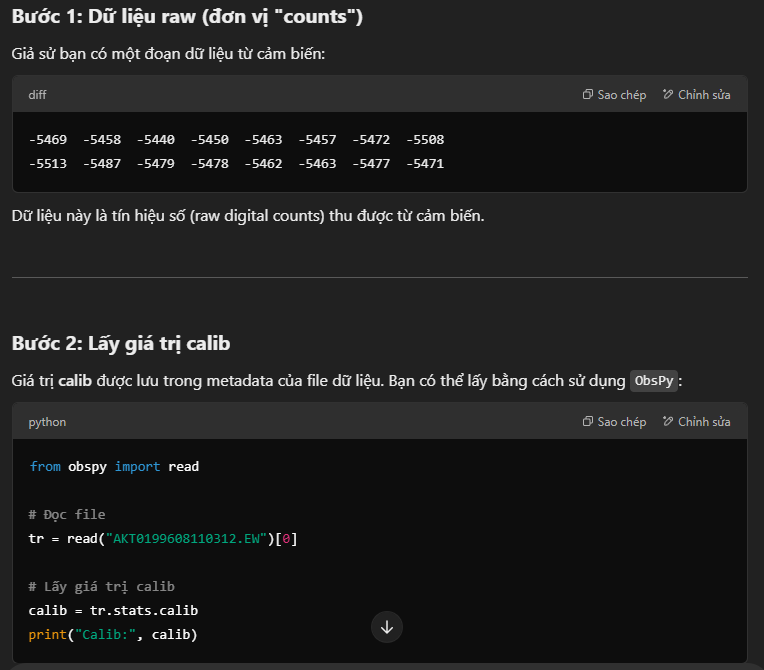
* Tóm tắt lại các kết quả đạt được.
* Khả năng ứng dụng thực tế.

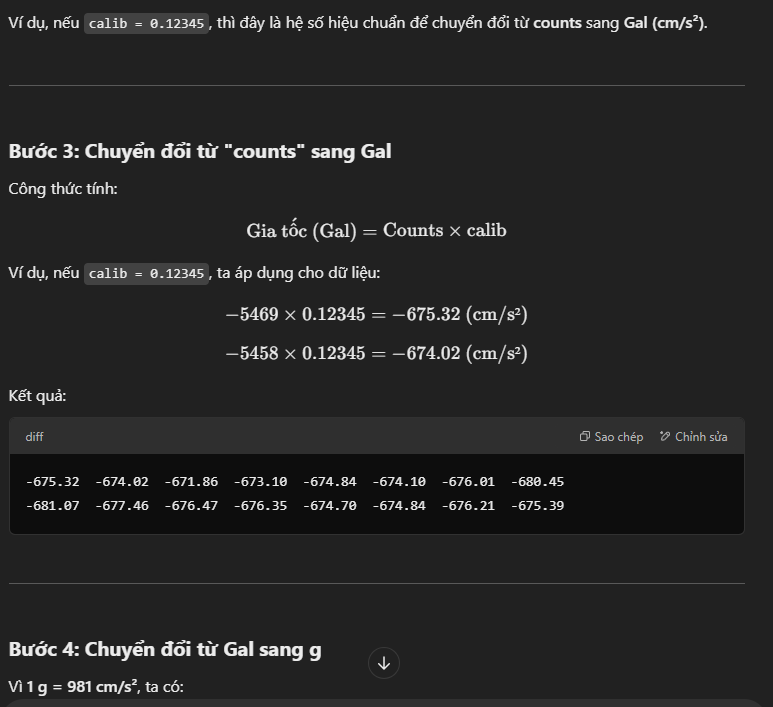
5.2 **Hướng phát triển**

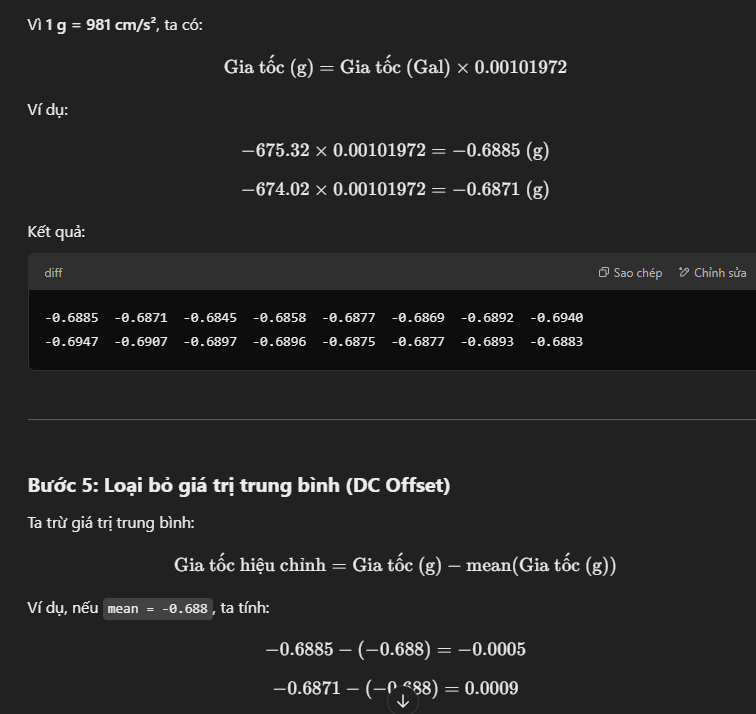
* Cải thiện độ chính xác của mô hình.
* Thu thập dữ liệu thực tế từ nhiều vị trí khác nhau.
* Ứng dụng mô hình vào hệ thống cảnh báo động đất.

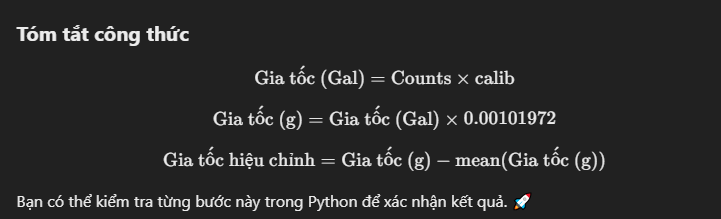
**TÀI LIỆU THAM KHẢO**

* Danh sách tài liệu, sách, bài báo khoa học và nguồn tham khảo.









Tỷ lệ **scale\_factor = 3920 / 6182761** có thể được sử dụng trong trường hợp dữ liệu **chưa được calib** hoặc cần hiệu chỉnh lại đơn vị của dữ liệu.

**1️⃣ Ý Nghĩa của scale\_factor = 3920 / 6182761**

* **3920**: Có thể là hệ số chuyển đổi của cảm biến hoặc giá trị chuẩn hóa nào đó trong thiết bị đo.
* **6182761**: Có thể là giá trị đếm (count) từ cảm biến gia tốc kỹ thuật số, thường là dữ liệu thô trước khi chuyển thành đơn vị vật lý.
* **Tỷ lệ này có thể dùng để chuyển từ giá trị đo thô (count) sang gia tốc thực tế** (thường là m/s² hoặc g).

**2️⃣ Khi nào cần scale\_factor?**

* Nếu **dữ liệu từ cảm biến là số đếm (count)**, không có giá trị calib trong metadata, bạn cần nhân với scale\_factor để chuyển sang đơn vị gia tốc.
* Nếu đã có **calib trong metadata**, thì không cần scale\_factor nữa, vì calib đã tự động chuyển đổi về đơn vị đúng.

**3️⃣ Khi nào KHÔNG cần scale\_factor?**

* Nếu tr.stats.calib đã có giá trị hợp lệ, bạn chỉ cần:

python

Sao chépChỉnh sửa

data[axis] = tr.data \* calib # Chuyển trực tiếp sang gia tốc

* Khi dữ liệu đọc từ K-NET đã chuẩn hóa về **g** hoặc **m/s²**, không cần thêm scale\_factor.

**4️⃣ Kết Luận**

✅ **Nếu dữ liệu có calib → chỉ nhân với calib**.  
✅ **Nếu dữ liệu là raw count → cần scale\_factor để chuyển đổi**.

**Tại sao lại chọn 0.1g hoặc 0.05g?**

* **0.1g**: Thường được xem là ngưỡng tối thiểu để một trận động đất có thể gây ra thiệt hại nhỏ đến vừa.
* **0.05g**: Ngưỡng thấp hơn, có thể được sử dụng để phát hiện cả những trận động đất nhẹ hơn hoặc tín hiệu rung động yếu.

Biên độ dao động của động đất phụ thuộc vào nhiều yếu tố như cường độ động đất, khoảng cách đến tâm chấn, loại mặt đất và cảm biến sử dụng. Tuy nhiên, mình có thể đưa ra một số phạm vi phổ biến dựa trên dữ liệu thực tế:

**🔹 1. Gia tốc dao động (Acceleration - m/s² hoặc gal)**

* **Động đất yếu (< 3.0 độ richter):** 1 - 10 gal (0.01 - 0.1 m/s²)
* **Động đất trung bình (3.0 - 5.0 độ richter):** 10 - 100 gal (0.1 - 1 m/s²)
* **Động đất mạnh (5.0 - 7.0 độ richter):** 100 - 500 gal (1 - 5 m/s²)
* **Động đất cực mạnh (> 7.0 độ richter):** > 500 gal (> 5 m/s²)

📌 **Lưu ý:** 1 gal = 0.01 m/s²

**🔹 2. Biên độ dịch chuyển (Displacement - mm hoặc cm)**

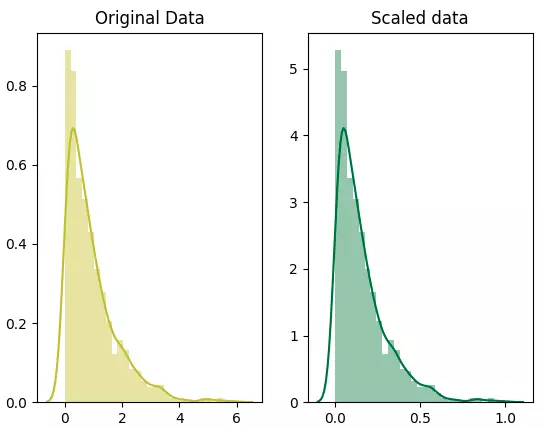
* **Động đất nhỏ:** Biên độ dao động dưới 1 mm
* **Động đất trung bình:** Biên độ dao động khoảng 1 - 10 mm
* **Động đất mạnh:** Có thể lên đến vài cm hoặc thậm chí > 1 m ở vùng gần tâm chấn

**🔹 3. Phân tích phổ tần số**

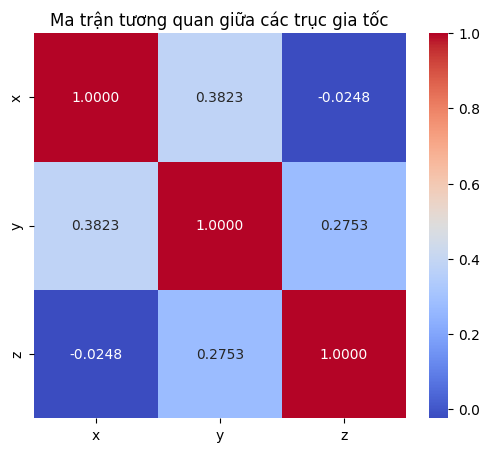
* **Tần số thấp (< 1 Hz):** Chủ yếu xuất hiện trong động đất lớn, tác động lên công trình cao tầng.
* **Tần số trung bình (1 - 10 Hz):** Gây ảnh hưởng đến các công trình trung bình và hệ thống hạ tầng.
* **Tần số cao (> 10 Hz):** Xuất hiện trong động đất nhỏ hoặc các rung động cục bộ.

Min-Max Scaling:

Mục tiêu của phương pháp là đưa các giá trị về gần hơn giá trị mean của features. Phương pháp này đưa các giá trị về 1 khoảng đặc biệt, thường là [0,1][0,1] hoặc [−1,1][−1,1]. Một trong những hạn chế của phương pháp này là khi áp dụng với một khoảng giá trị nhỏ, ta sẽ thu được độ lệch chuẩn nhỏ hơn, điều này làm giảm weight của các outliers trong dữ liệu



khẳng định scaling chỉ làm thay đổi khoảng giá trị chứ không làm thay đổi hình dạng phân phối dữ liệu là chính xác.



Biểu đồ heatmap:

**Phân Tích Ma Trận Tương Quan Giữa Các Trục Gia Tốc (x, y, z)**

**1️⃣ Ý nghĩa của ma trận tương quan**

Ma trận tương quan giúp bạn hiểu mức độ quan hệ giữa các trục gia tốc. Giá trị tương quan nằm trong khoảng **[-1, 1]**:

* **1.0** → Hai trục hoàn toàn đồng biến (tăng cùng nhau).
* **-1.0** → Hai trục hoàn toàn nghịch biến (một trục tăng, trục kia giảm).
* **0** → Không có mối quan hệ tuyến tính giữa hai trục.

**2️⃣ Nhận xét từ dữ liệu của bạn**

| **Tương quan** | **Giá trị** | **Ý nghĩa** |
| --- | --- | --- |
| **x - y** | **0.382** | Có mối tương quan dương nhẹ giữa trục X và Y (chuyển động theo phương ngang có ảnh hưởng đến nhau). |
| **x - z** | **-0.024** | Hầu như không có mối quan hệ giữa X và Z (dao động ngang không ảnh hưởng nhiều đến dao động dọc). |
| **y - z** | **0.275** | Có mối tương quan dương nhẹ giữa trục Y và trục Z (dao động trước/sau có thể ảnh hưởng một phần đến dao động dọc). |

**3️⃣ Giải thích vật lý từ kết quả**

1. **Trục X và Y có tương quan dương nhẹ (0.382)**
   * Điều này có thể chỉ ra rằng khi có rung động hoặc chuyển động ngang trên trục X, nó cũng ảnh hưởng đến trục Y một chút.
   * Đây có thể là dấu hiệu của sự dao động nghiêng hoặc các chuyển động địa chấn theo phương ngang.
2. **Trục X và Z gần như độc lập (-0.024)**
   * Gia tốc trên trục X không ảnh hưởng nhiều đến chuyển động trên trục Z (thẳng đứng).
   * Điều này cho thấy sự rung động theo phương ngang **không gây ra biến động đáng kể về độ cao**.
3. **Trục Y và Z có mối tương quan nhẹ (0.275)**
   * Sự dao động theo trục trước/sau (Y) có ảnh hưởng một phần đến độ cao (Z).
   * Có thể là do chuyển động nghiêng hoặc tác động của mặt đất.

**🎯 Ứng dụng thực tế**

🔹 **Trong phân tích địa chấn**:

* Nếu có một trận động đất, chúng ta có thể xác định hướng dao động chính dựa vào mối quan hệ giữa các trục.
* Nếu một trục có tương quan rất cao với trục khác, nó có thể là dấu hiệu của một **chuyển động mạnh có hướng rõ ràng**.

🔹 **Trong phân tích dao động kết cấu (cầu, tòa nhà)**:

* Nếu trục X và Z có tương quan cao → có thể có sự rung động mạnh theo phương thẳng đứng.
* Nếu trục Y và X có tương quan mạnh → rung lắc có thể chủ yếu theo phương ngang.

🔹 **Trong thiết bị đo rung động (VD: máy móc, ô tô, máy bay)**:

* Nếu có sự tương quan cao giữa X, Y, Z, điều đó có thể chỉ ra hệ thống bị rung động mất cân bằng.
* Nếu trục Z bị ảnh hưởng mạnh, có thể là dấu hiệu của **dao động thẳng đứng do va chạm hoặc xóc**.

**IQR, ZC và CAV được sử dụng để làm gì trong bài toán của bạn?**

Ba đặc trưng **IQR (Interquartile Range), ZC (Zero Crossing), CAV (Cumulative Absolute Velocity)** được sử dụng để trích xuất thông tin từ dữ liệu gia tốc của cảm biến nhằm phân biệt giữa động đất (**earthquake**) và hoạt động con người (**non-earthquake**). Cụ thể:

| **Đặc trưng** | **Công thức** | **Ý nghĩa** |
| --- | --- | --- |
| **IQR (Interquartile Range)** | IQR = Q3 - Q1 | Đo lường độ phân tán của dữ liệu, giúp xác định mức độ dao động mạnh/yếu của tín hiệu. |
| **ZC (Zero Crossing)** | ZC = số lần tín hiệu đổi dấu | Đếm số lần tín hiệu thay đổi từ dương sang âm (hoặc ngược lại), giúp đánh giá tần số dao động. |
| **CAV (Cumulative Absolute Velocity)** | `CAV = ∑ | acceleration |

**1️⃣ IQR - Độ phân tán của tín hiệu**

* **Áp dụng vào đâu?**
  + Tính toán khoảng giữa của dữ liệu gia tốc (giữa **Q1** và **Q3**) trong mỗi cửa sổ thời gian (WINDOW\_SIZE).
* **Tại sao quan trọng?**
  + Động đất thường có biên độ dao động lớn hơn so với hoạt động con người, vì vậy IQR giúp nhận diện mức độ biến thiên của tín hiệu.
* **Cách tính trong code của bạn:**

python

Sao chépChỉnh sửa

np.percentile(segment, 75) - np.percentile(segment, 25) # IQR

**2️⃣ ZC - Đo số lần đổi dấu của tín hiệu**

* **Áp dụng vào đâu?**
  + Tính số lần tín hiệu VS (vector sum gia tốc) đổi dấu trong từng khoảng thời gian (WINDOW\_SIZE).
* **Tại sao quan trọng?**
  + Hoạt động con người có nhiều chuyển động nhỏ, thường có tần số cao hơn (nhiều lần đổi dấu hơn).
  + Động đất có dao động mạnh nhưng chậm hơn, nên ZC có thể giúp phân biệt.
* **Cách tính trong code của bạn:**

python

Sao chépChỉnh sửa

np.sum(np.diff(np.sign(segment)) != 0) # ZC

* + np.sign(segment): Xác định dấu của mỗi điểm trong cửa sổ.
  + np.diff(): Tính sự khác biệt giữa các điểm liên tiếp.
  + != 0: Đếm số lần đổi dấu.

**3️⃣ CAV - Tổng gia tốc tích lũy**

* **Áp dụng vào đâu?**
  + Tính tổng gia tốc trong mỗi khoảng thời gian (WINDOW\_SIZE), chia cho tần số lấy mẫu (SAMPLING\_RATE).
* **Tại sao quan trọng?**
  + Động đất có rung động kéo dài hơn nên **CAV thường cao hơn** so với hoạt động con người.
  + Hoạt động con người có gia tốc thấp hơn và dao động không liên tục, nên CAV nhỏ hơn.
* **Cách tính trong code của bạn:**

python

Sao chépChỉnh sửa

np.sum(segment) / SAMPLING\_RATE # CAV

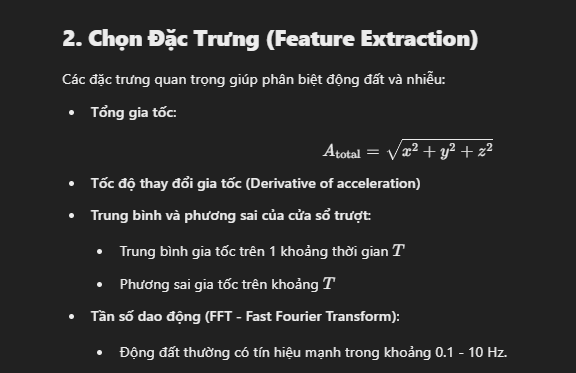
* + np.sum(segment): Tổng giá trị gia tốc trong cửa sổ.
  + / SAMPLING\_RATE: Chuẩn hóa theo tần số lấy mẫu.

**📌 Tổng kết**

Ba đặc trưng này giúp mô tả các khía cạnh khác nhau của tín hiệu gia tốc để phân biệt giữa động đất và hoạt động con người:

* **IQR**: Độ mạnh/yếu của rung động.
* **ZC**: Tần số dao động.
* **CAV**: Mức năng lượng tích lũy của tín hiệu.

📌 **Trong code của bạn, các đặc trưng này được sử dụng làm đầu vào (features) cho mô hình học máy ANN để phân loại động đất hay không.** 🚀



**📌 Dữ liệu của bạn sẽ có 1 hàng hay nhiều hàng mỗi trận?**

🔹 **Câu trả lời:** **Phụ thuộc vào cách bạn tổ chức dữ liệu.**  
Có **2 cách phổ biến** để biểu diễn dữ liệu cho mô hình học máy:

**✅ Cách 1: Mỗi trận động đất là 1 hàng dữ liệu (Aggregate Features)**

* Bạn tính toán **các đặc trưng thống kê và miền tần số** từ toàn bộ tín hiệu của trận động đất.
* Mỗi trận động đất tương ứng với **1 hàng duy nhất** trong tập dữ liệu.

Ví dụ: Nếu một trận động đất kéo dài **10 giây**, với **sampling rate = 100 Hz**, thì bạn có **10 × 100 = 1000 giá trị trên mỗi trục (X, Y, Z)**.  
➡ **Bạn sẽ tính toán các đặc trưng trên toàn bộ 1000 giá trị và chỉ lưu một hàng dữ liệu cho trận đó.**

| **mean(X)** | **mean(Y)** | **mean(Z)** | **std(X)** | **std(Y)** | **std(Z)** | **max(X)** | **min(X)** | **Dominant Freq** | **Energy** | **Magnitude** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.002 | -0.001 | 0.98 | 0.12 | 0.11 | 0.10 | 1.2 | -1.1 | 5.5 Hz | 0.8 | 3.5 |

📌 **Ưu điểm**:

* **Dữ liệu gọn nhẹ** hơn, huấn luyện mô hình nhanh.
* Dễ sử dụng với các mô hình **Random Forest, SVM, XGBoost**, v.v.

📌 **Nhược điểm**:

* Mất thông tin chi tiết theo thời gian.
* Nếu trận động đất kéo dài lâu, có thể mất một số tín hiệu quan trọng.

**✅ Cách 2: Chia trận động đất thành nhiều đoạn nhỏ (Sliding Window)**

* **Thay vì tổng hợp toàn bộ trận động đất thành một hàng duy nhất, bạn chia tín hiệu ra thành nhiều khoảng thời gian nhỏ (cửa sổ thời gian - Sliding Window).**
* Mỗi cửa sổ có độ dài cố định (ví dụ: **1 giây = 100 mẫu**) và có thể **chồng lấn nhau** (overlapping).
* Khi đó, **một trận động đất có thể có nhiều hàng dữ liệu**.

Ví dụ: Nếu trận động đất kéo dài **10 giây**, bạn có thể chia nó thành **10 cửa sổ 1 giây** hoặc chồng lấn để có nhiều hơn.

| **mean(X)** | **mean(Y)** | **mean(Z)** | **std(X)** | **std(Y)** | **std(Z)** | **max(X)** | **min(X)** | **Dominant Freq** | **Energy** | **Magnitude** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.002 | -0.001 | 0.98 | 0.12 | 0.11 | 0.10 | 1.2 | -1.1 | 5.5 Hz | 0.8 | 3.5 |
| 0.003 | -0.002 | 0.97 | 0.10 | 0.09 | 0.12 | 1.1 | -1.0 | 4.2 Hz | 0.7 | 3.5 |
| 0.004 | -0.001 | 0.96 | 0.09 | 0.08 | 0.11 | 1.0 | -0.9 | 3.8 Hz | 0.6 | 3.5 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

📌 **Ưu điểm**:

* **Giữ được nhiều thông tin về sự thay đổi theo thời gian**.
* Tốt hơn khi dùng với **mô hình Deep Learning (CNN, LSTM, Transformer)**.

📌 **Nhược điểm**:

* Tăng số lượng dữ liệu đáng kể, có thể gây khó khăn khi huấn luyện mô hình truyền thống.
* Cần chọn **kích thước cửa sổ hợp lý** để tránh mất thông tin.

**📢 Bạn nên chọn cách nào?**

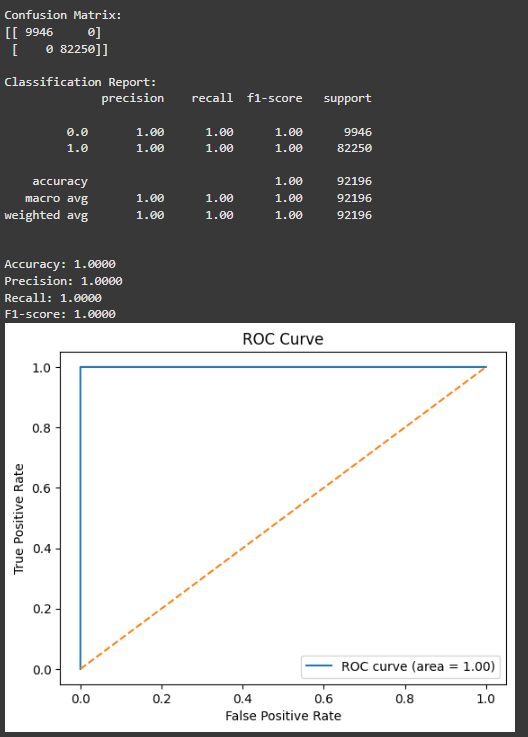
* **Nếu dùng mô hình truyền thống (Random Forest, SVM, XGBoost):**  
  ➝ **Cách 1** (Mỗi trận = 1 hàng) là tối ưu hơn.
* **Nếu muốn dùng Deep Learning (CNN, LSTM, Transformer):**  
  ➝ **Cách 2** (Sliding Window) giúp giữ thông tin theo thời gian.

👉 **Bạn có muốn mình hỗ trợ viết code trích xuất đặc trưng theo một trong hai cách không?**

**Tại sao IQR quan trọng?**

* Nếu **IQR lớn**, điều đó cho thấy tín hiệu dao động mạnh hơn.
* Nếu **IQR nhỏ**, rung động ổn định hơn.
* IQR có thể giúp xác định xem dữ liệu có bị **nhiễu nhiều hay không**, hữu ích trong phân tích động đất.

QUÁ trình training bị overfit:



Tại sao mô hình của tôi lại có độ chính xác cao thế

**ChatGPT đã nói:**

Câu hỏi rất hay! Nếu mô hình của bạn có **độ chính xác cao**, có thể đến từ **một hoặc nhiều yếu tố sau**. Dưới đây là phân tích từ góc độ chuyên môn của một chuyên gia học máy và rung động:

**✅ 1. Đặc trưng phân biệt rõ ràng**

Bạn đã chọn những đặc trưng cực kỳ mạnh như:

* Dominant\_freq\_y: Động đất thường có dải tần thấp, nhiễu có thể ngẫu nhiên hoặc cao hơn.
* Energy\_z: Động đất thường tạo ra năng lượng lớn và phân bố đều hơn.
* ZC\_z: Số lần đổi dấu – nhiễu thường dao động nhiều và ngẫu nhiên hơn so với động đất.

➡️ **Những đặc trưng này phân tách rất tốt giữa hai lớp**, giúp mô hình học rất dễ.

**✅ 2. Dữ liệu sạch và rõ ràng**

* Không có nhiễu nhãn (label noise).
* Không bị mất giá trị hoặc dữ liệu sai.
* Tập dữ liệu cân bằng khá tốt (Earthquake ≈ Noise).

➡️ Dữ liệu chất lượng cao sẽ khiến mô hình dễ đạt hiệu năng tốt.

**✅ 3. Sự khác biệt vật lý giữa Earthquake và Noise**

Trong rung động học:

* **Earthquake** có dạng sóng điển hình, biên độ lớn, năng lượng rõ ràng, thường có dạng **low-frequency burst**.
* **Noise** thường không có quy luật vật lý rõ ràng, dễ bị mô hình học được "dấu vết" khác biệt.

**✅ 4. Mô hình Random Forest rất mạnh**

Random Forest:

* Rất tốt trong việc xử lý đặc trưng không tuyến tính.
* Không dễ bị overfitting như các mô hình khác khi số cây lớn.
* Tự động chọn ra đặc trưng quan trọng nhất.

**🚨 Tuy nhiên: Hãy cẩn trọng!**

Mặc dù độ chính xác cao là tốt, nhưng bạn cần kiểm tra thêm:

* **Dữ liệu test có bị trùng với train không?**
* **Có overfit khi thử với dữ liệu thực tế?**
* **Có outliers nào mô hình không xử lý được không?**

**DỰ ĐOÁN ĐỘ LỚN TRẬN ĐỘNG ĐẤT:**

Bạn đã cung cấp một tập dữ liệu gồm 2.909 trận động đất với 34 đặc trưng đầu vào và 1 cột đích (Magnitude) đại diện cho độ lớn của trận động đất. Dữ liệu bao gồm các đặc trưng thống kê (mean, std, skew, kurtosis...), tần số chiếm ưu thế, năng lượng, v.v. cho ba trục x, y, z.

Nếu bạn muốn xây dựng một mô hình dự đoán độ lớn của trận động đất (Magnitude), các bước chính có thể như sau:

**🧠 Bước 1: Chuẩn bị dữ liệu**

* Kiểm tra phân phối và tương quan các đặc trưng
* Chuẩn hóa dữ liệu nếu cần (đặc biệt quan trọng với các mô hình như SVR, KNN)
* Chia dữ liệu thành tập huấn luyện và kiểm tra

**🤖 Bước 2: Chọn mô hình**

* Hồi quy tuyến tính (Linear Regression)
* Random Forest Regressor
* XGBoost / LightGBM
* MLPRegressor (Neural Network)

**📊 Bước 3: Đánh giá mô hình**

* Dùng các metric như MAE, RMSE, R² score
* Trực quan hóa dự đoán vs giá trị thực tế

Câu hỏi rất hay! Sử dụng nhiều đặc trưng **có thể mang lại lợi ích**, nhưng cũng tiềm ẩn một số **vấn đề** nếu không kiểm soát kỹ. Dưới đây là cái nhìn tổng quát:

**✅ Ưu điểm khi dùng nhiều đặc trưng**

* **Tăng độ chính xác** nếu đặc trưng có liên quan thật sự đến mục tiêu (ở đây là Magnitude)
* **Giúp mô hình học được mối quan hệ phức tạp** giữa các tín hiệu địa chấn
* **Giảm thiếu hụt thông tin** nếu mỗi đặc trưng nắm bắt một khía cạnh riêng của dữ liệu

**⚠️ Những rủi ro tiềm ẩn**

| **Vấn đề** | **Ý nghĩa** |
| --- | --- |
| **Đặc trưng dư thừa (redundancy)** | Nhiều đặc trưng có thể bị lặp thông tin, làm nhiễu mô hình |
| **Đặc trưng không liên quan (irrelevant)** | Làm giảm độ chính xác vì mô hình học những thứ không liên quan |
| **Đa cộng tuyến (multicollinearity)** | Đặc biệt với Linear Regression – nếu 2 đặc trưng quá tương quan, mô hình khó học đúng |
| **Overfitting** | Mô hình học quá chi tiết vào dữ liệu huấn luyện, kém hiệu quả khi dự đoán thực tế |

Câu hỏi rất chuyên sâu và thực tế, và mình sẽ trả lời cả từ **góc nhìn chuyên gia địa chấn học** và **kỹ thuật rung động học**.

### 🌍 ****1. Góc nhìn địa chấn học (Seismology)****

#### ✅ Trục Z (thẳng đứng) là ****rất nhạy**** với động đất

* Trục Z thường là nơi **đầu tiên ghi nhận sóng P (Primary wave)** – loại sóng đầu tiên khi có động đất.
* Sóng P di chuyển **theo phương nén – giãn**, và **trục Z (thẳng đứng)** là nơi cảm nhận rõ nhất lực này.
* Do đó, **Z thường có tín hiệu rõ rệt ngay khi động đất xảy ra**.

#### ✅ Trục X và Y (ngang) thường ghi nhận:

* **Sóng S (Secondary wave)** và **sóng bề mặt (surface waves)** – xuất hiện sau P-wave.
* Những sóng này có biên độ lớn hơn nhưng đến sau, và thường chịu ảnh hưởng bởi cấu trúc địa chất địa phương.

### 🧪 ****2. Góc nhìn kỹ thuật – máy đo gia tốc (Accelerometer)****

| **Trục** | **Nhận diện động đất** | **Ghi chú** |
| --- | --- | --- |
| **Z** | ⭐⭐⭐ Rõ nhất | P-wave đi theo chiều thẳng đứng |
| X/Y | ⭐⭐ Bình thường | Thể hiện sóng S, surface waves |

### 📊 ****3. Thực nghiệm từ machine learning (như bạn đang làm)****

Các mô hình học máy thường cũng phát hiện rằng:

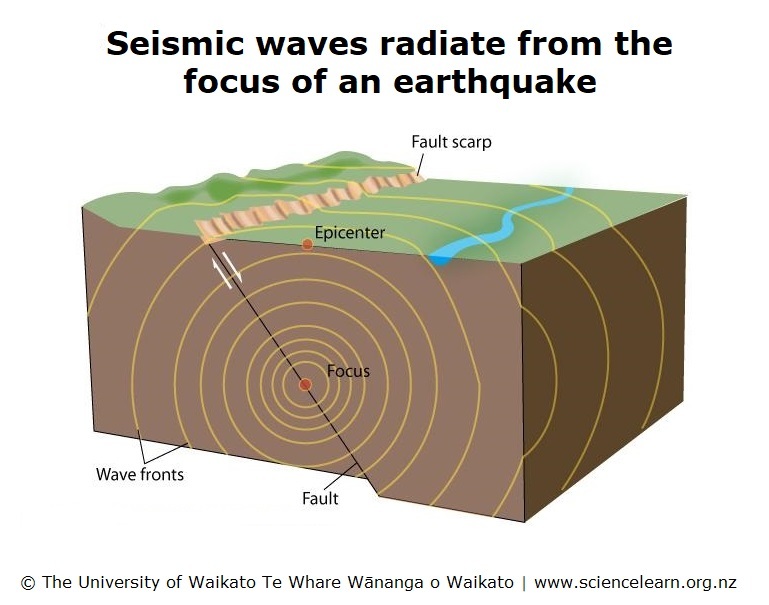
* **Các đặc trưng từ trục Z** (như Energy\_z, ZC\_z, Peak\_z,...) thường nằm trong **top các đặc trưng quan trọng nhất**.
* Nhưng đôi khi **kết hợp các trục X/Y/Z** lại cho kết quả tốt hơn nhờ tính bù trừ.

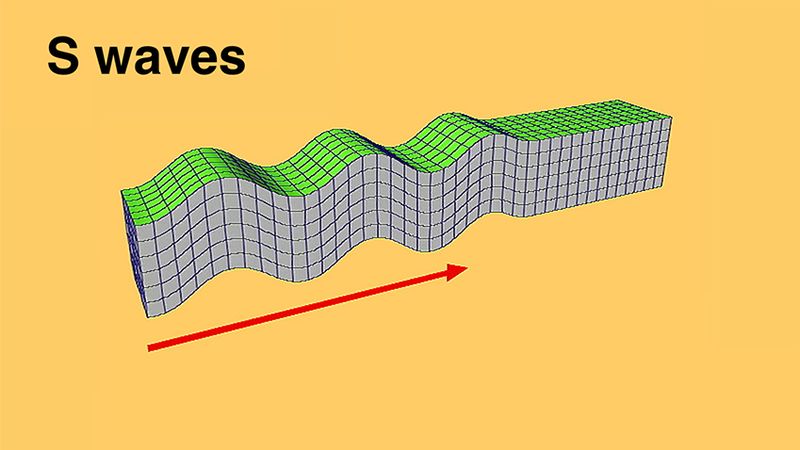
### ✅ Kết luận chuyên gia:

* Nếu **chỉ chọn 1 trục** để phát hiện động đất: **chọn trục Z**.
* Nếu có thể dùng kết hợp → **tốt nhất dùng cả 3**, nhưng có thể **ưu tiên đặc trưng từ trục Z** trong feature selection.

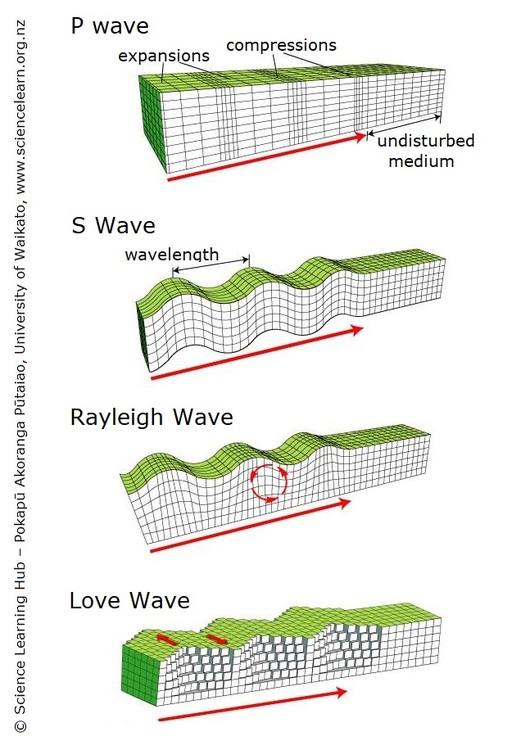
### ****Top 5 đặc trưng quan trọng nhất:****

1. **Dominant\_freq\_y**  
   → Tần số trội của trục Y  
   👉 Động đất thường có tần số đặc trưng rõ rệt, trong khi noise thì phân bố tần số ngẫu nhiên hơn.
2. **Energy\_z**  
   → Tổng năng lượng tín hiệu trục Z  
   👉 Động đất thường tạo ra năng lượng lớn, nhất là theo trục thẳng đứng (Z).
3. **ZC\_z** (Zero-Crossing count)  
   → Số lần tín hiệu trục Z đổi dấu  
   👉 Phản ánh mức độ dao động – động đất thường có dạng dao động đặc trưng hơn noise.
4. **ZC\_y**  
   → Số lần đổi dấu trên trục Y  
   👉 Giúp phân biệt sự ngẫu nhiên của noise so với dao động hệ thống từ động đất.
5. **IQR\_z** (Interquartile Range)  
   → Độ phân tán trung bình của tín hiệu trục Z  
   👉 Cho thấy biên độ dao động – động đất thường tạo ra tín hiệu với mức phân tán lớn hơn.





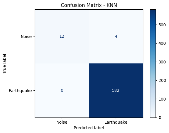
https://www.britannica.com/video/P-wave/-2820



ĐÁNH GIÁ MÔ HÌNH:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mô hình | Accuracy | Precision | Recall | F1 | Confution matrix | ROC |
| Decision Tree | 0.9816 | 0.9831 | 0.9983 | 0.9906 |  | 0.69  C:\Users\Admin\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7963ED7C.tmp |
| KNN | 0.9933 | 0.9932 | 1 | 0.9966 | **C:\Users\Admin\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\2389BCEA.tmp** | C:\Users\Admin\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\4F1B3488.tmp |
| Naive Bayes |  |  |  |  |  |  |
| Logistic Regression |  |  |  |  |  |  |
| Random Forest |  |  |  |  |  |  |
| SVM |  |  |  |  |  |  |
| XGBoost |  |  |  |  |  |  |



****

## 💡 Ví dụ phân tích một số đặc trưng trong top của bạn:

| **Đặc trưng** | **Giải thích** | **Vì sao quan trọng với EQ/Non-EQ** |
| --- | --- | --- |
| ZC\_x, ZC\_y, ZC\_z | Zero Crossing Rate – số lần tín hiệu đổi dấu | Tín hiệu EQ có tần suất dao động cao hơn, nhiều điểm đổi dấu hơn |
| Energy\_z | Tổng năng lượng tín hiệu (∑x²) | EQ thường có năng lượng rung động cao rõ rệt |
| Dominant\_freq\_z | Tần số mạnh nhất trong tín hiệu | EQ có đặc trưng tần số khác biệt rõ so với nhiễu |
| IQR\_z | Độ phân tán tín hiệu | Tín hiệu EQ thường “rộng hơn” so với nhiễu tĩnh |

=> Những đặc trưng này không chỉ **có ý nghĩa vật lý** mà còn được **mô hình tự động xác định là nổi bật nhất** trong việc phân loại.

## 📌 Tổng kết lý do chọn 15 đặc trưng:

1. **Tự động chọn theo độ quan trọng thống kê từ mô hình học máy**
2. **Giảm chiều dữ liệu** → nhanh hơn, tránh overfitting
3. **Giải thích được bằng trực giác tín hiệu EQ** (năng lượng, tần số, độ dao động)
4. **Thực nghiệm cho thấy mô hình với top-15 đặc trưng thường có hiệu suất cao gần bằng toàn bộ**