**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**

**KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHẤT LƯỢNG CAO**

**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**

**KHOÁ LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHẤT LƯỢNG CAO**

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

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

**HÀ NỘI – 2025**

**TÓM TẮT**

Với sự tiến bộ vượt bậc của khoa học và công nghệ, đặc biệt là trong lĩnh vực trí tuệ nhân tạo (AI) và Internet vạn vật (IoT), việc triển khai các hệ thống tự động hóa phát hiện sớm thiên tai đang trở thành xu thế tất yếu trong công tác phòng chống và giảm nhẹ rủi ro. Trong đó, phát hiện sớm động đất là một trong những ứng dụng quan trọng và cấp thiết nhất. Bên cạnh công nghệ cảm biến hiện đại cho phép thu thập dữ liệu liên tục và chính xác từ hàng nghìn điểm quan trắc. Khi được kết hợp với AI và học máy, các hệ thống này không chỉ phát hiện rung chấn bất thường trong thời gian thực, mà còn có thể phân biệt được giữa động đất thật và nhiễu nền, đưa ra cảnh báo sớm chỉ trong vài giây điều có thể quyết định giữa sự sống và cái chết cho hàng ngàn người.

Đồ án tập trung nghiên cứu và xây dựng mô hình phát hiện động đất dựa trên dữ liệu từ cảm biến gia tốc ba trục. Trong bối cảnh Việt Nam và nhiều quốc gia khác ngày càng đối mặt với rủi ro thiên tai từ động đất, việc phát hiện sớm các tín hiệu địa chấn là một yêu cầu cấp thiết nhằm giảm thiểu thiệt hại về người và tài sản. Dữ liệu được sử dụng chủ yếu từ hai nguồn: K-NET (Nhật Bản) và Italia, bao gồm các tín hiệu động đất thật và nhiễu nền như xe cộ, gió,... Sau khi thu thập và trải qua các quá trình tiền xử lý như chuyển đổi đơn vị, loại bỏ ngoại lai, chuẩn hóa, dữ liệu được chia thành các đoạn nhỏ bằng kỹ thuật sliding window 10 giây. Tiếp theo là quá trình trích xuất đặc trưng từ tín hiệu, bao gồm các đặc trưng thời gian và tần số như IQR, ZCR, Energy, PSD,... giúp mô hình phân biệt rõ giữa tín hiệu động đất và nhiễu. Đồ án triển khai trên nhiều mô hình học máy ML cũng như DL như Random Forest, SVM, Logistic Regression,... với quy trình huấn luyện, đánh giá theo chuẩn scikit-learn, sử dụng Cross-validation (K-Fold) và các chỉ số như Accuracy, Precision, Recall, F1-Score. Mô hình Random Forest cho kết quả phân loại tốt nhất trên tập dữ liệu thử nghiệm.

Kết quả cho thấy mô hình có thể nhận diện hiệu quả tín hiệu động đất so với nhiễu với độ chính xác cao. Tuy nhiên, đề tài hiện vẫn chỉ dừng ở mức mô phỏng offline và chỉ có chức năng phân biệt giữa tín hiệu động đất và nhiễu chứ chưa áp dụng real-time và chưa tích hợp vào mạng cảnh báo sớm. Từ đó, đồ án đề xuất định hướng tiếp theo là phát triển ứng dụng thời gian thực, mở rộng dữ liệu địa phương và tăng cường học sâu để cải thiện hiệu năng.

Đây là một hướng nghiên cứu tiềm năng, kết hợp giữa kỹ thuật cảm biến và xử lý tín hiệu và trí tuệ nhân tạo nhằm nâng cao khả năng ứng phó với thiên tai tại Việt Nam.

***Từ khóa:*** *Machine Learning, DeepLearning, FFT, IQR.*

**LỜI CẢM ƠN**

Trải qua gần bốn năm hành trình học tập và nghiên cứu tại Khoa Cơ học kỹ thuật và Tự động hóa, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, em muốn dành những lời biết ơn chân thành đến Ban Giám hiệu, Ban Chủ nhiệm khoa và tất cả các thầy cô đã dành thời gian giảng dạy và truyền đạt tri thức cho em. Đặc biệt, em muốn gửi lời biết ơn sâu sắc đến thầy TS Hoàng Văn Mạnh, thầy là người đã luôn đồng hành và đưa ra nhiều định hướng quý báu trong quá trình hoàn thiện đồ án của em.

Em cũng muốn bày tỏ lòng biết ơn đến gia đình, bạn bè và những người đồng nghiệp đã luôn ủng hộ, động viên và chia sẻ cho em những trải nghiệm và kinh nghiệm quý báu.

Cuối cùng, em muốn gửi lời cảm ơn chân thành đến các bạn trong lớp K66-MT3. Đã đồng hành và để lại cho nhau nhiều bài học, kỷ niệm quý giá.

Em xin chân thành cảm ơn!

*Hà Nội, ngày 19 tháng 05 năm 2025.*

*Sinh viên*

*Thái*

*Nguyễn Ngọc Thái*

**LỜI CAM ĐOAN**

Em xin cam đoan đồ án tốt nghiệp “Nghiên cứu bài toán phát hiện động đất sử dụng dữ liệu cảm biến gia” là công trình nghiên cứu của em dưới sự hướng dẫn của TS.Hoàng Văn Mạnh. Các tài liệu mà em đã sử dụng để hoàn thành đồ án tốt nghiệp này đã được nhắc đến trong mục “Tài liệu tham khảo”. Các số liệu thử nghiệm dùng để đánh giá và trình bày trong đồ án, tất cả hoàn toàn trung thực. Em xin chịu hoàn toàn trách nhiệm về đề tài nếu có bất kỳ gian lận nào.

*Hà Nội, ngày 19 tháng 05 năm 2024*

*Người cam đoan*

*Thái*

Nguyễn Ngọc Thái

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI VÀ MỤC TIÊU NGHIÊN CỨU 8](#_Toc196125044)

[1.1. Đặt vấn đề 8](#_Toc196125045)

[*1.1.1. Tình hình Động đất trên thế giới 10*](#_Toc196125046)

[*1.1.2. Tình hình Động đất tại Việt Nam 11*](#_Toc196125047)

[*1.1.3. Vai trò của việc phát hiện sớm động đất 12*](#_Toc196125048)

[*1.1.4. Những hạn chế của hệ thống cảnh báo truyền thống 13*](#_Toc196125049)

[1.2 Mục tiêu nghiên cứu 15](#_Toc196125050)

[1.3 Phạm vi và giới hạn của đề tài 16](#_Toc196125051)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 18](#_Toc196125052)

[2.1. Giới thiệu về động đất 18](#_Toc196125053)

[2.2. Đơn vị đo động đất 19](#_Toc196125054)

[2.3. Tìm hiểu về cảm biến gia tốc 23](#_Toc196125055)

[2.4. Các mô hình học máy trong phát hiện động đất 25](#_Toc196125056)

[*2.4.1. Mô hình Logistic Regression 25*](#_Toc196125057)

[*2.4.2. Mô hình học máy Decision Tree 27*](#_Toc196125058)

[*2.4.3. Mô hình Random Forest 28*](#_Toc196125059)

[*2.4.4. Mô hình Support Vector Machines 28*](#_Toc196125060)

[*2.4.5. Mô hình ANN 28*](#_Toc196125061)

[*2.4.6. Mô hình CRNN 28*](#_Toc196125062)

[2.5. Tổng quan các nghiên cứu liên quan 28](#_Toc196125063)

[CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU 30](#_Toc196125064)

[3.1. Nguồn dữ liệu và tiền xử lý 30](#_Toc196125065)

[*3.1.1. Nguồn dữ liệu cảm biến gia tốc 30*](#_Toc196125066)

[*3.1.2. Trực quan hoá dữ liệu theo miền gia tốc thời gian 33*](#_Toc196125067)

[*3.1.3. Trực quan hoá theo miền PSD (Power Spectral Density) 36*](#_Toc196125068)

[3.2. Quá trình thu thập và xây dựng tập dữ liệu 38](#_Toc196125069)

[*3.2.1. Tiền xử lý dữ liệu 38*](#_Toc196125070)

[*3.2.2. Làm sạch dữ liệu huấn luyện 40*](#_Toc196125071)

[3.3. Trích xuất đặc trưng 45](#_Toc196125072)

[*3.3.1. IQR - Interquartile Range – Khoảng tứ phân vị 45*](#_Toc196125073)

[*3.3.2. ZC- Zero Crossing Rate 45*](#_Toc196125074)

[*3.3.3. CAV 45*](#_Toc196125075)

[*3.3.4. Dominant Frequenccy 46*](#_Toc196125076)

[*3.3.5. Energy 46*](#_Toc196125077)

[*3.3.6. Mean 46*](#_Toc196125078)

[*3.3.7. Độ lệch chuẩn (Std) 46*](#_Toc196125079)

[*3.3.8. Peak to Peak 47*](#_Toc196125080)

[*3.3.9. Skew 47*](#_Toc196125081)

[*3.3.10. Kurtosis 47*](#_Toc196125082)

[3.4. Xây dựng mô hình học máy 50](#_Toc196125083)

[3.5. Các tiêu chí đánh giá mô hình 52](#_Toc196125084)

[*3.5.1. Confusion Matrix 52*](#_Toc196125085)

[*3.5.2. Accuracy (Độ chính xác tổng thể) 53*](#_Toc196125086)

[*3.5.3. Precision (Độ chính xác của dự đoán dương tính) 54*](#_Toc196125087)

[*3.5.4. Recall (Độ bao phủ hay độ nhạy) 55*](#_Toc196125088)

[*3.5.5. F1-score 55*](#_Toc196125089)

[*3.5.6. ROC Curve và AUC 56*](#_Toc196125090)

[3.6. Công cụ và phần mềm sử dụng 57](#_Toc196125091)

[*3.6.1. Ngôn ngữ lập trình 57*](#_Toc196125092)

[*3.6.2. Môi trường làm việc 58*](#_Toc196125093)

[CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT QUÀ 59](#_Toc196125094)

[4.1. Thiết lập thí nghiệm 59](#_Toc196125095)

[4.2. Kết quả huấn luyện và đánh giá mô hình 60](#_Toc196125096)

[4.3. Dự đoán độ lớn của trận động đất 68](#_Toc196125097)

[4.4. Thiết kế giao diện người dùng 71](#_Toc196125098)

[CHƯƠNG 5: THẢO LUẬN VÀ KẾT LUẬN 74](#_Toc196125099)

[TÀI LIỆU THAM KHẢO 76](#_Toc196125100)

**MỤC LỤC HÌNH ẢNH**

**Chương 1**

[Hình 1.1. Bản đồ phân bố các trận động đất có độ lớn >4.0 trên toàn cầu năm 2015 8](#_Toc196122149)

[Hình 1.2. Số lượng động đất lớn hơn 6 Richter từ 2020-2024. 9](#_Toc196122150)

[Hình 1.3. Bản đồ động đất khu vực Đông Nam Á. 11](#_Toc196122151)

[Hình 1.4. Địa động nghi của Trung Quốc 14](#_Toc196122152)

**Chương 2**

[Hình 2.1. Hệ thống cảnh báo sớm động đất dựa trên sóng P và sóng S 23](#_Toc196122176)

[Hình 2.2. Cấu tạo của cảm biến gia tốc 23](#_Toc196122177)

[Hình 2.3. Cảm biến giá tốc MCU 6050 24](#_Toc196122178)

[Hình 2.4. Logistic Regression 26](#_Toc196122179)

[Hình 2.5. Biểu diễn Hàm Sigmoid 26](#_Toc196122180)

**Chương 3**

[Hình 3.1. Dữ liệu động đất từ trang chủ K-NET 31](#_Toc196122193)

[Hình 3.2. Dữ liệu gia tốc trả về theo hướng E-W của bộ K-NET 32](#_Toc196122194)

[Hình 3.3. Trực quan hoá dữ liệu động đất theo miền thời gian. 34](#_Toc196122195)

[Hình 3.4. Bộ dữ liệu nhiễu sử dụng trong quá trình 34](#_Toc196122196)

[Hình 3.5. Trực quan hoá dữ liệu nhiễu theo miền thời gian 35](#_Toc196122197)

[Hình 3.6. Trực quan hoá dữ liệu nhiễu theo miền tần số FFT 35](#_Toc196122198)

[Hình 3.7. Trực quan hoá dữ liệu động đất theo miền tần số FFT 36](#_Toc196122199)

[Hình 3.8. Miền PSD của một trận động đất 37](#_Toc196122200)

[Hình 3.9. Biểu đồ heatmap thể hiện mỗi quan hệ giữa các trục 38](#_Toc196122201)

[Hình 3.10. So sánh trước và sau khi loại bỏ giá trị ngoại lai 41](#_Toc196122202)

[Hình 3.11. Biểu đồ histogram sau khi loại bỏ ngoại lai theo phương pháp IQR 42](#_Toc196122203)

[Hình 3. 12. Biểu đồ histogram của nhóm dữ liệu động đất 43](#_Toc196122204)

[Hình 3.13. Các đặc trưng sau khi trích xuất của tập EQ 48](#_Toc196122205)

[Hình 3.14. Các đặc trưng sau khi trích xuất của tập noise 48](#_Toc196122206)

[Hình 3.15. Biểu đồ thể hiện tương quan giữa các đặc trưng 49](#_Toc196122207)

[Hình 3.16. Top 15 đặc trưng quan trọng được lấy từ mô hình Random Forest 50](#_Toc196122208)

[Hình 3.17. Kỹ thuật cross-validation (K-folder) 51](#_Toc196122209)

[Hình 3.18. Ma trận nhầm lẫn (Confusion matrix) 53](#_Toc196122210)

[Hình 3.19. Biểu đồ biểu diễn ROC curve 56](#_Toc196122211)

[Hình 3.20. Biểu đồ ROC thể hiện hiệu quả của các bộ phân loại 57](#_Toc196122212)

**Chương 4**

[Hình 4.1. Phân tách bộ dữ liệu 59](#_Toc196122218)

[Hình 4.2. Mô hình Logistic Regression 62](#_Toc196122219)

[Hình 4.3. Mô hình Decision Tree 63](#_Toc196122220)

[Hình 4.4. Mô hình KNN 64](#_Toc196122221)

[Hình 4.5. Mô hình Naïve Bayes 65](#_Toc196122222)

[Hình 4.6. Mô hình Random Forest 66](#_Toc196122223)

[Hình 4.7. Mô hình Support Vector Machine 67](#_Toc196122224)

[Hình 4.8. Mô hình CRNN 68](#_Toc196122225)

[Hình 4.9. Biểu đồ phân tán giữa giá trị dự đoán và giá trị thực tế 70](#_Toc196122226)

[Hình 4.10. Biểu đồ miền thể hiện giá trị thực tế và giá trị dự đoán 71](#_Toc196122227)

[Hình 4.11. Giao diện phầm mềm QT Designer 72](#_Toc196122228)

[Hình 4.12. Giao diện người dùng hệ thống 73](#_Toc196122229)

**MỤC LỤC BẢNG BIỂU**

**Chương 2**

[Bảng 2.1. Mức độ của trận động đất theo thang Richter 19](#_Toc196122134)

[Bảng 2.2. Thang đo cường độ động đất 20](#_Toc196122135)

**Chương 4**

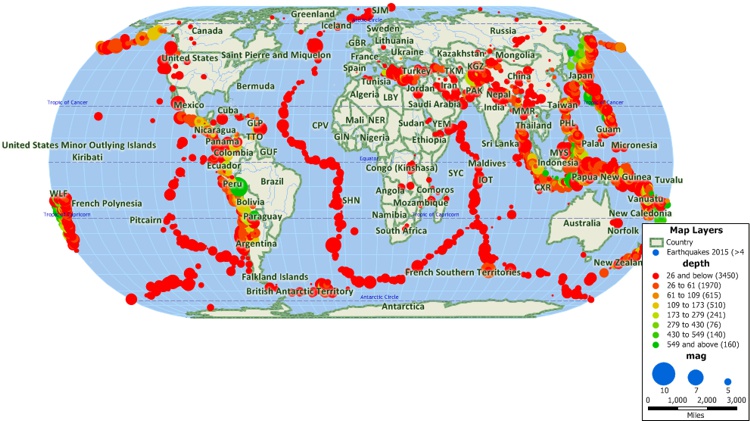
[Bảng 4.1. Bảng thống kê định lượng các metrics để đánh giá kết quả mô hình huấn luyện 61](#_Toc196122230)

[Bảng 4.2. Thống kê kết quả các mô hình phán đoán độ lớn trận động đất 69](#_Toc196122231)

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI VÀ MỤC TIÊU NGHIÊN CỨU

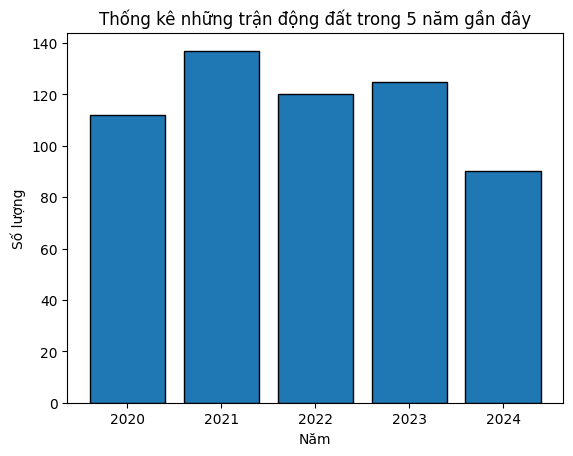
## 1.1. Đặt vấn đề

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. 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.



**Hình 1.1. Bản đồ phân bố các trận động đất có độ lớn >4.0 trên toàn cầu năm 2015**

Hình trên đây thể hiện sự phân bố các trận động đất có độ lớn trên 4.0 xảy ra trên toàn thế giới trong năm 2015. Các vị trí xảy ra động đất được đánh dấu bằng các chấm tròn, trong đó kích thước của vòng tròn biểu thị độ lớn (magnitude) của trận động đất, còn màu sắc thể hiện độ sâu (depth). Có thể thấy, phần lớn các trận động đất tập trung dọc theo các ranh giới mảng kiến tạo, đặc biệt là khu vực bao quanh Thái Bình Dương và trải dài qua Nhật Bản, Indonesia, Philippines, New Zealand, và bờ Tây châu Mỹ. Ngoài ra, các vùng như Trung Á, Iran, Thổ Nhĩ Kỳ và rìa Địa Trung Hải cũng ghi nhận nhiều hoạt động địa chấn. Ngược lại, châu Phi, Bắc Âu và phần lớn nội địa châu Á ít xuất hiện các trận động đất đáng kể.



**Hình 1.2. Số lượng động đất lớn hơn 6 Richter từ 2020-2024.**

Biểu đồ thống kê những trận động đất lớn hơn 6.0 độ Richter trong 5 năm trở lại đây. Một số trận tiêu biểu có thể nhắc đến như là:

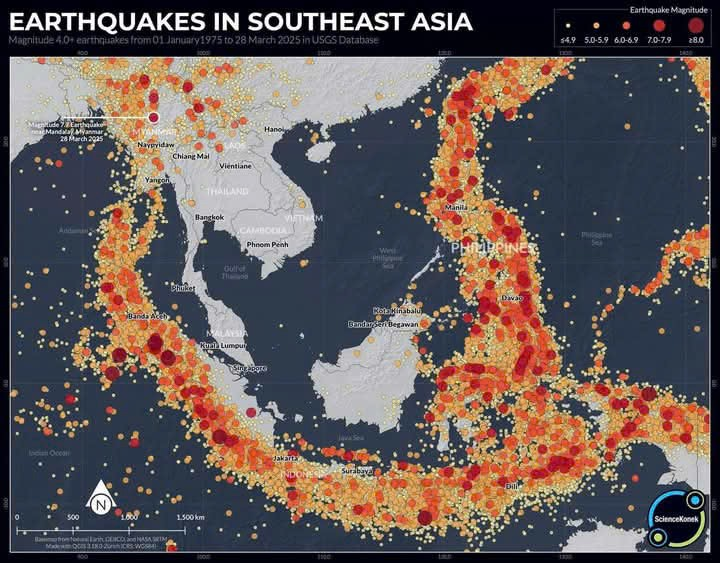
* Tại Thổ Nhĩ Kì và Syria (6/2/2023): hai trận động đất liên tiếp với cường độ 7.8 và 7.5 độ Richter đã khiến hơn 59.000 người thiệu mạng.
* Tại Nhật Bản (1/1/2024), trận động đất với cường độ 7.5 độ Richter ở đảo Noto đã kiến 84 người thiệt mạng và gây ra sóng thần nhỏ. ), trận động đất với cường độ 7.5 độ Richter ở
* Tại Myanmar (28/3/2025), trận động đất với cường độ 7.7 độ Richter ở Sagaing đã kiến 3.689 người thiệt mạng và hơn 5.000 người bị thương.

### 1.1.1. Tình hình Động đất trên thế giới

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.

Mới gần đây nhất, vào ngày 30/3/2025 tại Myanmar đã xảy ra một trận động đất mạnh 7,7 độ richter đã xảy ra tại vùng Sagaing, gần thành phố Mandalay, Myanmar. Đây là trận động đất mạnh nhất tại Myanmar kể từ năm 1912, gây ra thiệt hại nghiêm trọng về người và tài sản.

### 1.1.2. Tình hình Động đất tại Việt Nam



**Hình 1.3. Bản đồ động đất khu vực Đông Nam Á.**

Việt Nam tuy không nằm trên vành đai lửa Thái Bình Dương – khu vực có hoạt động địa chấn mạnh nhất thế giới, nhưng vẫn chịu ảnh hưởng bởi các hoạt động kiến tạo địa chất nội địa và khu vực lân cận. Các trận động đất tại Việt Nam chủ yếu liên quan đến hệ thống đứt gãy kiến tạo đang hoạt động. Đặc biệt, khu vực Tây Bắc với các đứt gãy như Điện Biên – Mường Lay, Sông Mã – Tuần Giáo – Lai Châu đã từng ghi nhận những trận động đất mạnh.

* Năm 1935, một trận động đất 6.8 độ richter xảy ra tại Điện Biên.
* Năm 1983 tại Tuần Giáo (Điện Biên) cũng trải qua trận động đất 6,7 độ richter, gây thiệt hại đáng kể về cơ sở hạ tầng và kinh tế.

Ngoài ra, những hoạt động nhân sinh cũng góp phần làm kích thích động đất. Việc xây dựng và vận hành các hồ chứa thủy điện lớn làm thay đổi áp lực nước và ứng suất trong lòng đất, có thể kích hoạt các đứt gãy tiềm ẩn. Tại huyện Kon Plông, tỉnh Kon Tum, sau khi các hồ chứa được tích nước, khu vực này đã ghi nhận nhiều trận động đất liên tiếp, cho thấy mối liên hệ giữa hoạt động nhân sinh và động đất kích thích. ​

Hơn nữa, Việt Nam cũng chịu ảnh hưởng từ các trận động đất mạnh ở khu vực lân cận. Ví dụ như trận động đất 7,7 độ richter tại Myanmar ngày 28/3/2025 đã gây rung chấn cảm nhận được tại Hà Nội và TP.HCM, dù khoảng cách đến tâm chấn lên tới hàng nghìn km. Sóng địa chấn lan truyền xa, kết hợp với nền đất yếu và chất lượng xây dựng chưa đảm bảo, dẫn đến hiện tượng rung lắc và thậm chí nứt tường tại một số tòa nhà cao tầng ở TP Hồ Chí Minh.

Từ yếu tố trên cho chúng ta thấy, dù không nằm trên vành đai lửa, Việt Nam vẫn đối mặt với nguy cơ động đất từ cả hoạt động kiến tạo tự nhiên và tác động nhân sinh. Việc nghiên cứu, theo dõi và đánh giá các đứt gãy hoạt động, cùng với quản lý chặt chẽ các hoạt động có thể kích thích động đất, là cần thiết để giảm thiểu rủi ro và thiệt hại do động đất gây ra.

### 1.1.3. Vai trò của việc phát hiện sớm động đất

Phát hiện sớm động đất đóng vai trò then chốt trong việc giảm thiểu thiệt hại về người và tài sản, đặc biệt tại những khu vực có mật độ dân cư cao và hạ tầng kém chịu lực. Khác với các thảm họa thiên nhiên khác như bão hay lũ, động đất thường xảy ra đột ngột, không có dấu hiệu báo trước, nên khoảng thời gian cảnh báo – dù chỉ vài giây đến vài chục giây – cũng có thể tạo ra sự khác biệt sinh tử.

Việc phát hiện sớm giúp cho:

* Cảnh báo kịp thời cho người dân và chính quyền để di tản khẩn cấp hoặc trú ẩn an toàn.
* Kích hoạt hệ thống tự động như ngắt điện, dừng tàu cao tốc, đóng van gas, giúp ngăn chặn cháy nổ và tai nạn dây chuyền.
* Hỗ trợ điều phối ứng cứu nhanh chóng, hiệu quả, đặc biệt ở các khu vực trọng yếu như bệnh viện, sân bay, trường học,...
* Tăng khả năng phục hồi sau thiên tai, giảm thiểu chi phí tái thiết và tổn thất kinh tế lâu dài.

Động đất là một trong những thiên tai xảy ra bất ngờ và gần như không có dấu hiệu báo trước. Chính vì thế, việc phát hiện sớm – dù chỉ trước vài giây đến vài chục giây – cũng có thể tạo ra sự khác biệt rất lớn. Trong khoảng thời gian ngắn ngủi ấy, nếu hệ thống cảnh báo được kích hoạt đúng lúc, người dân có thể kịp thời trú ẩn an toàn, các thiết bị như thang máy, tàu cao tốc hay hệ thống gas có thể được dừng hoặc ngắt tự động, giúp giảm thiểu đáng kể thiệt hại về người và tài sản.

Đặc biệt ở những khu vực đông dân cư, hoặc nơi có nhiều công trình trọng yếu như bệnh viện, trường học, sân bay… một cảnh báo sớm có thể cứu sống hàng trăm, thậm chí hàng nghìn người. Không chỉ vậy, việc phát hiện sớm còn giúp các cơ quan chức năng có thêm thời gian để điều phối ứng cứu, tổ chức lực lượng và giảm thiểu rủi ro trong giai đoạn sau động đất.

Trong thời đại công nghệ phát triển mạnh mẽ, việc ứng dụng các hệ thống cảm biến kết hợp với trí tuệ nhân tạo – đặc biệt là các mô hình học máy – đang mở ra hướng đi mới, hiệu quả hơn cho việc nhận diện sớm các tín hiệu địa chấn. Mục tiêu không chỉ là phát hiện nhanh mà còn là phát hiện đúng, hạn chế tối đa các cảnh báo giả, từ đó nâng cao độ tin cậy và khả năng phản ứng kịp thời của cộng đồng.

### 1.1.4. Những hạn chế của hệ thống cảnh báo truyền thống

Mặc dù các hệ thống cảnh báo động đất truyền thống đã đóng vai trò quan trọng trong nhiều thập kỷ qua, nhưng trên thực tế, chúng vẫn tồn tại không ít hạn chế. Một trong những điểm yếu lớn nhất là độ trễ trong việc phát hiện và phát đi cảnh báo. Các hệ thống này thường phụ thuộc vào mạng lưới cảm biến địa chấn cố định và việc xử lý tín hiệu đôi khi chưa đủ nhanh để đưa ra cảnh báo trong thời gian tối ưu – đặc biệt là với các trận động đất gần tâm chấn.

Bên cạnh đó, độ chính xác chưa cao cũng là một vấn đề. Nhiều hệ thống truyền thống khó phân biệt được giữa tín hiệu địa chấn thực sự và các nguồn nhiễu từ con người hoặc môi trường, dẫn đến tình trạng cảnh báo sai hoặc cảnh báo giả, gây hoang mang trong cộng đồng và làm suy giảm lòng tin vào hệ thống. Thậm chí, một số hệ thống còn không thể đưa ra cảnh báo sớm nếu động đất xảy ra ở vùng không có đủ cảm biến hoặc hạ tầng truyền dẫn tín hiệu chưa được đồng bộ.

Ngoài ra, việc triển khai và duy trì các hệ thống truyền thống thường tốn kém và phức tạp, đòi hỏi sự phối hợp chặt chẽ giữa nhiều cơ quan kỹ thuật và địa phương. Trong điều kiện thời tiết xấu, thiên tai liên hoàn hoặc mất điện, hệ thống có thể bị gián đoạn, làm giảm hiệu quả cảnh báo.

Chính vì những lý do đó, việc áp dụng công nghệ mới và đặc biệt là trí tuệ nhân tạo và mô hình học máy, đang được xem là một hướng tiếp cận đầy triển vọng nhằm khắc phục các điểm yếu của hệ thống truyền thống, đồng thời nâng cao tính chủ động và chính xác trong công tác phòng chống thiên tai.

Những hệ thống cảnh báo động đất tiêu biểu có thể kể đến là Hệ thống Cảnh báo Động đất Sớm (**EEW - Earthquake Early Warning**) của Nhật Bản, được vận hành bởi Cơ quan Khí tượng Nhật Bản (**JMA**). Hệ thống này sử dụng hàng nghìn trạm đo địa chấn để phát hiện sóng P (sóng đầu tiên và ít nguy hiểm) và ước lượng thời điểm sóng S (sóng gây rung lắc mạnh) sẽ đến, từ đó gửi cảnh báo sớm đến người dân qua tivi, radio và điện thoại di động. Tuy nhiên, độ trễ trong truyền tin và giới hạn trong dự báo chính xác ở vùng tâm chấn.

*Ảnh có chứa kèn đồng, đồng, kim loại, trong nhà

Nội dung do AI tạo ra có thể không chính xác.*

**Hình 1.4. Địa động nghi của Trung Quốc**

Tại Hoa Kỳ, hệ thống ShakeAlert được phát triển cho vùng bờ Tây – nơi thường xuyên xảy ra động đất. Hệ thống này kết hợp dữ liệu từ các cảm biến địa chấn do USGS (Cơ quan Khảo sát Địa chất Hoa Kỳ) quản lý và cung cấp cảnh báo qua ứng dụng di động, hệ thống giao thông và các cơ sở công cộng. ShakeAlert từng đưa ra cảnh báo trước 8–10 giây cho một số trận động đất lớn tại California, đủ để người dân kịp thời hành động. Dù vậy, ở những vùng ít cảm biến hoặc vùng có địa hình phức tạp, hệ thống vẫn có thể bị hạn chế hiệu quả.

Những ví dụ trên cho thấy, dù hệ thống truyền thống vẫn đóng vai trò nền tảng trong công tác theo dõi động đất, nhưng để đáp ứng nhu cầu cảnh báo nhanh, chính xác, phạm vi rộng, cần thiết phải tích hợp thêm các phương pháp hiện đại như học máy, học sâu và mạng cảm biến thông minh.

* 1. **Mục tiêu nghiên cứu**

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 " Phát hiện động đất sử dụng dữ liệu cảm biến gia tốc" của em 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 có thể phân biệt được đâu là dữ liệu của động đất và đâu là dữ liệu của các rung động nhiễu khác. 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.

Mục tiêu chính của nghiên cứu này là xây dựng một hệ thống có khả năng phân biệt giữa tín hiệu động đất thực sự và các tín hiệu nhiễu từ dữ liệu cảm biến gia tốc ba trục, thông qua việc ứng dụng các mô hình học máy. Hệ thống này hướng đến việc hỗ trợ phát hiện sớm các trận động đất một cách tự động, nhằm nâng cao hiệu quả cảnh báo và giảm thiểu rủi ro thiên tai.

Quá trình nghiên cứu tập trung vào các mục tiêu chính như sau:

1. Tiền xử lý và phân tích dữ liệu cảm biến gia tốc được thu thập từ các nguồn thực tế (bao gồm cả tín hiệu động đất và nhiễu).
2. Trích xuất các đặc trưng phù hợp từ tín hiệu thời gian hoặc chuyển đổi sang miền tần số, phục vụ cho quá trình huấn luyện mô hình.
3. Lựa chọn, huấn luyện và đánh giá hiệu năng của một số mô hình học máy, như Random Forest, SVM, KNN hoặc ANN, để tìm ra mô hình phù hợp nhất cho bài toán phân loại này.
4. Triển khai giao diện đơn giản (dạng phần mềm hoặc ứng dụng mô phỏng) cho phép người dùng nạp dữ liệu và nhận kết quả phân loại: “Động đất” hay “Nhiễu”.
5. Đánh giá khả năng ứng dụng thực tiễn của mô hình trong các hệ thống cảnh báo sớm động đất, đồng thời đề xuất hướng cải tiến trong tương lai.

Thông qua đó, nghiên cứu kỳ vọng đóng góp một phần nhỏ vào việc tự động hóa quá trình nhận diện tín hiệu địa chấn, giúp tăng độ chính xác và rút ngắn thời gian cảnh báo so với các phương pháp truyền thống.

* 1. **Phạm vi và giới hạn của đề tài**

Đề tài chỉ phân loại tín hiệu thành hai nhóm: Động đất (EQ) và Nhiễu (Noise), không phân tích thêm thông tin như độ lớn, tọa độ tâm chấn, thời gian xảy ra hoặc mức độ ảnh hưởng.

Chỉ sử dụng dữ liệu cảm biến gia tốc 3 trục dạng thô, không khai thác thêm các nguồn dữ liệu khác như sóng địa chấn (seismic waveform), GPS, hoặc cảm biến rung chuyên dụng.

Dữ liệu huấn luyện và kiểm thử mang tính giới hạn về số lượng và khu vực địa lý, chủ yếu thu thập được từ các nguồn mô phỏng hoặc tập dữ liệu công khai. Dữ liệu trong quá trình thực hiện chủ yếu sử dụng hai bộ dataset đó là bộ K-NET của Cơ quan Khí tượng Nhật Bản và bộ thứ hai là bộ dataset về động đất của Italia.

Một trong những giới hạn quan trọng của đề tài là chưa triển khai hệ thống phát hiện động đất theo thời gian thực (real-time), cũng như chưa tích hợp vào một mạng lưới cảnh báo sớm hoàn chỉnh. Các mô hình được xây dựng và đánh giá chủ yếu trong môi trường thử nghiệm, với dữ liệu được xử lý sẵn và thực thi dưới dạng mô phỏng. Điều này đồng nghĩa với việc kết quả nghiên cứu hiện tại chỉ mới kiểm chứng được khả năng phân loại tín hiệu trong điều kiện lý tưởng, chưa phản ánh đầy đủ hiệu quả khi áp dụng trong điều kiện thực tế, đặc biệt là trong các hệ thống yêu cầu tốc độ xử lý và tính ổn định cao như hệ thống cảnh báo sớm thiên tai.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Giới thiệu về động đất

Trái đất được tạo thành từ:

* Một lõi bên trong rắn chắc
* Lõi ngoài nóng chảy
* Lớp phủ dày và chủ yếu là rắn, chiếm khoảng 84% tổng thể tích của trái đất.
* Lớp vỏ tương đối mỏng, có độ dày thay đổi từ 5 đến 50 km.

Lớp vỏ ngoài của Trái Đất không phải là một bề mặt liên tục. Thay vào đó, nó bao gồm các phân đoạn lớn được gọi là các mảng kiến ​​tạo . Động đất xảy ra dọc theo ranh giới giữa các mảng kiến ​​tạo hoặc tại vị trí các vết nứt trong các mảng, được gọi là đứt gãy

**Khái niệm:**

Động đất là sự rung động của một khu vực trên vỏ Trái Đất dưới ảnh hưởng của những nguyên nhân nằm trong lòng Trái Đất (nội sinh) hay nguyên nhân từ bên ngoài Trái Đất (ngoại sinh), thậm chí còn do con người tạo ra (nhân sinh).

**Nguyên nhân gây ra động đất:**

Các mảng kiến ​​tạo tạo nên lớp vỏ trái đất đang chuyển động liên tục. Khi các cạnh của các mảng này trượt vào nhau trong các vùng đứt gãy, ma sát có thể làm chậm chúng lại, dẫn đến sự tích tụ áp suất trong thời gian dài. Khi lực chuyển động cuối cùng thắng được ma sát, các phần của lớp vỏ đột nhiên vỡ ra hoặc bị dịch chuyển, giải phóng áp suất bị dồn nén dưới dạng sóng địa chấn. Đây là một trận động đất tự nhiên, đôi khi được gọi là động đất kiến ​​tạo.

**Nguyên nhân nội sinh:**

* Đó có thể là do sự sụp đổ của các hang động ngầm bên dưới mặt đất, dẫn đến các vụ trượt lở đất đá tự nhiên với khối lượng lớn và dẫn đến động đất. Tuy nhiên, nguyên nhân này chỉ làm rung chuyển một vùng hẹp và chỉ chiếm khoảng 3% trên tổng số các trận động đất đã xảy ra.
* Động đất do sự phun trào của núi lửa nhưng cường độ cũng không mạnh lắm và chỉ chiếm khoảng 7% trên tổng số các trận động đất đã xảy ra.
* Động đất do kiến tạo, nghĩa là việc đứt gãy các kiến tạo, đặc biệt là ở ở rìa các mảng thạch quyển hoặc ở các đới hút chìm. Hoặc cũng có thể là hoạt động của macma xâm nhập vào vỏ Trái Đất và tạo ra những rung chuyển lớn. Nguyên nhân động đất này thường chiếm đến 90%.

**Nguyên nhân ngoại sinh:**

* Do ảnh hưởng của các thiên thạch khi di chuyển đã va chạm vào Trái Đất.

**Nguyên nhân nhân sinh**:

* Do các hoạt động của con người tạo ra sự rung lắc mạnh. Điển hình đó là các vụ thử hạt nhân, nổ nhân tạo phía dưới lòng đất hoặc tác động của áp suất cột nước ở các hồ chứa nước, hồ thủy điện.
* Trên Trái Đất, động đất là một hiện tượng tự nhiên thường xuyên xảy ra. Mỗi năm Trái Đất có khoảng 5 triệu lần động đất, trung bình mỗi ngày có khoảng gần 13.000 lần. Tuy nhiên có tới 99% trong số lần động đất này chỉ là những chấn động nhỏ mà chỉ có các thiết bị máy móc mới ghi nhận được, còn lại 1% mới gây ra ảnh hưởng hoặc tai họa cho con người.
  1. **Đơn vị đo động đất**

Để đo độ mạnh yếu của động đất, người ta dùng hai đơn vị đo là: cấp độ và cường độ.

**Cấp độ:** Biểu thị độ lớn nhỏ của động đất, người ta đo được nó thông qua năng lượng do sóng động đất giải phóng ra khi động đất và dùng thang độ Richter để biểu thị. Thang này được chia thành 9 cấp từ 1 – 9.

Thang Richter dựa vào hàm logarit cơ số là 10 để xác định biên độ tối đa các rung chấn của Trái đất. Mỗi độ của thang Richter biểu thị sự tăng giảm biên độ rung chấn theo hệ số 10 và tăng giảm về năng lượng phát sinh theo hệ số 32.

Công thức tính độ lớn động đất theo thang Richter:

Trong đó:

* M: độ lớn Richter (Magnitude)
* A: biên độ lớn nhất của sóng địa chấn (thường tính bằng micromet)
* C: hệ số hiệu chỉnh phụ thuộc vào khoảng cách từ trạm đo đến tâm chấn

Như vậy một trận động đất 5 độ Richter sẽ gây nên rung chấn mạnh gấp 10 lần và toả ra một năng lượng  gấp 32 lần độ 4, và cứ thế mà tăng theo cấp số nhân với công bội là 10 và 32. Để dễ hình dung, có thể lấy ví dụ: độ 1 Richter tương đương sức nổ của 1,5 kg thuốc nổ TNT thì của một trận động đất cấp độ Richter có sức phá hoại tương đương 6 triệu tấn thuốc nổ TNT.

**Bảng 2.1. Mức độ của trận động đất theo thang Richter**

|  |  |  |
| --- | --- | --- |
| **Cường độ** | **Mức độ tác hại** | **Độ phổ biến** |
| **< 2.0** | Động đất rất nhỏ, con người không cảm nhận được chỉ có thiết bị đo địa chấn ghi nhận. | Xảy ra hàng triệu lần mỗi năm trên toàn cầu |
| **2.0 – 2.9** | Rất nhẹ, con người hiếm khi cảm nhận được | Khoảng 1.3 triệu lần mỗi năm |
| **3.0 – 3.9** | Nhẹ, có thể cảm nhận được nhưng không gây thiệt hại. | Khoảng 130.000 trận mỗi năm |
| **4.0 – 4.9** | Trung bình, rung lắc nhẹ, có thể làm rung đồ vật trong nhà nhưng hiếm khi gây hư hại. | Khoảng 13.000 trận mỗi năm |
| **5.0 – 5.9** | Mạnh, có thể gây thiệt hại nhỏ cho các công trình yếu, toà nhà kiên cố thường không bị ảnh hưởng nhiều. | Khoảng 1.300 trận mỗi năm |
| **6.0 -6.9** | Động đất lớn, gây thiệt hại đáng kể ở khu đông dân cư, có thể làm nứt tường hay sập một số công trình | Khoảng 100 trận mỗi năm |
| **7.0 – 7.9** | Rất mạnh, gât thiệt hại nghiêm trọng, gây sạt lở đất, sóng thần nếu xảy ra dưới biển. | Khoảng 10- 20 trận mỗi năm |
| **8.0 – 8.9** | Cực kì mạnh, có thể phá huỷ hoàn toàn một thành phố lớn, gây sóng thần, ảnh hưởng rộng lớn | Khoảng 1- 2 trận mỗi năm |
| **9.0 +** | Siêu động đất, cực kì hiếm gặp, có thể thay đổi địa hình và sóng thần mạnh | Chỉ xảy ra 1 lần trong vài thập kỉ. |

**Cường độ:** Biểu thị những ảnh hưởng khác nhau do động đất gây ra trên mặt đất, thể hiện bằng thang độ **Meccali** với 12 cấp chia. Một trận động đất có cùng cấp độ nhưng ở các nơi khác nhau sẽ có cường độ khác nhau.

**Bảng 2.2. Thang đo cường độ động đất**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cấp độ** | **Mô tả cường độ** | **Mức độ tàn phá** | **Độ phổ biến** |
| **I** | Rất yếu | Không cảm nhận được, chỉ có thiết bị đo địa chấn ghi nhận. | Xảy ra hằng ngày trên thế giới. |
| **II** | Yếu | Một số người nhạy cảm có thể cảm nhận được, đặc biệt là ở tầng cao. Không gây thiệt hại. | Rất phổ biến, hằng ngày. |
| **III** | Nhẹ | Giống như rung nhẹ do xe tải lớn chạy qua, đèn treo có thể lắc nhẹ. Không thiệt hại. | Xảy ra thường xuyên. |
| **IV** | Trung bình | Nhiều người cảm nhận được, cửa sổ và đồ vật nhỏ rung lắc. Có thể đánh thức người ngủ nhẹ. | Xảy ra khá thường xuyên. |
| **V** | Khá mạnh | Đồ đạc di chuyển, cửa kính có thể vỡ, có thể gây thiệt hại nhỏ cho nhà yếu. | Xảy ra hằng năm ở một số khu vực. |
| **VI** | Mạnh | Nứt tường nhỏ, đồ vật rơi khỏi kệ, một số công trình xây dựng kém có thể hư hại nhẹ. | Xảy ra mỗi vài năm ở một khu vực nhất định |
| **VII** | Rất mạnh | Gây hư hại trung bình đến nghiêm trọng với các công trình yếu, nhà xây tốt có thể bị nứt. | Xảy ra không quá thường xuyên. |
| **VII** | Rất mạnh | Nhà kiên cố bị hư hại, tường gạch có thể sập, tượng đài và cột trụ đổ. | Xảy ra vài lần mỗi thập kỷ. |
| **IX** | Huỷ hoại | Nhiều công trình kiên cố bị phá hủy, đường ray xe lửa cong vênh, sạt lở đất xảy ra. | Hiếm, vài thập kỷ một lần. |
| **X** | Thảm khốc | Nhà cửa bị phá hủy hoàn toàn, cầu sập, mặt đất nứt toác, thay đổi địa hình. | Rất hiếm, thế kỷ một lần. |
| **XI** | Cực kì thảm khốc | Hầu hết các công trình bị phá hủy, mặt đất bị nứt nẻ mạnh, sóng thần có thể xảy ra. | Cực kỳ hiếm, hàng trăm năm mới có. |
| **XII** | Hủy diệt hoàn toàn | Mọi thứ bị phá hủy hoàn toàn, thay đổi địa hình lớn, có thể làm biến mất cả khu vực. | Hiếm gặp nhất, hàng trăm đến hàng nghìn năm. |

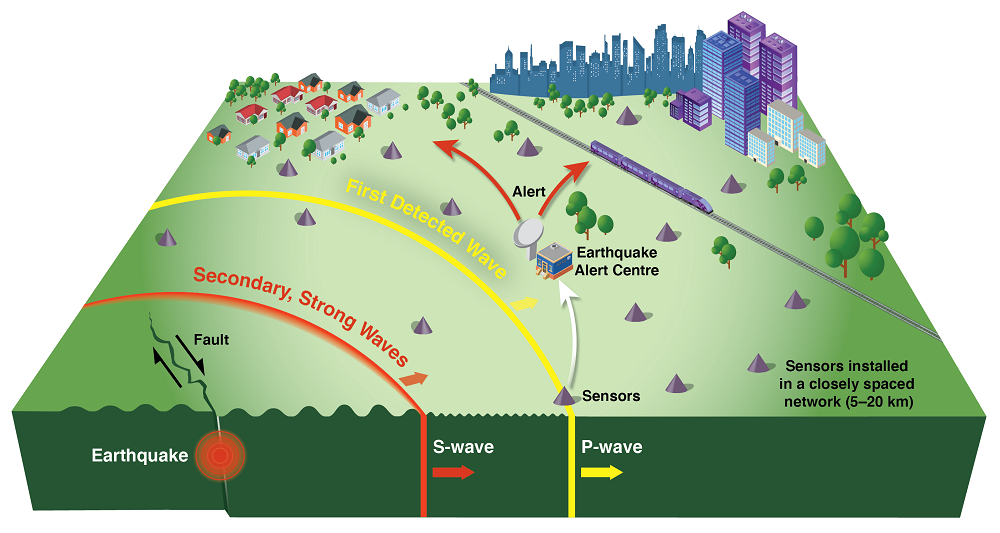
**Một số danh từ thường gọi trong động đất**

* Tâm địa chấn (còn gọi là chấn tâm): Nơi phát sinh chấn động trong lòng đất, thường ở độ sâu từ 0 - 700 km.
* Tâm động đất: Vị trí trên mặt đất, nơi tâm địa chấn truyền thẳng lên mặt đất.
* Sóng địa chấn: Sóng chấn động lan truyền trên mặt đất và trong lòng đất kể từ tâm địa chấn ra xung quanh.
* Sóng dư chấn: Sóng địa chấn như những đợt sóng lan tỏa từ tâm động đất ra xung quanh, nếu gặp các vật cản sẽ dội ngược lại thành sóng dư chấn.

**Các loại sóng địa chấn trong quá trình động đất xảy ra:**

Có một số loại sóng địa chấn khác nhau và tất cả chúng đều di chuyển theo những cách khác nhau. Hai loại sóng chính là: **sóng khối và sóng bề mặt**.

Sóng khối có thể di chuyển qua các lớp bên trong của Trái đất, nhưng sóng bề mặt chỉ có thể di chuyển dọc theo bề mặt của hành tinh giống như gợn sóng trên mặt nước. Động đất phát ra năng lượng địa chấn dưới dạng sóng khối (P và S). Khi sóng khối chạm tới bề mặt, một phần năng lượng đó được chuyển đổi thành sóng bề mặt.



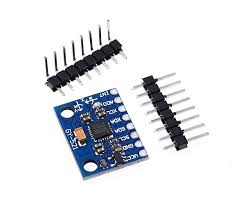
**Hình 2.1. Hệ thống cảnh báo sớm động đất dựa trên sóng P và sóng S**

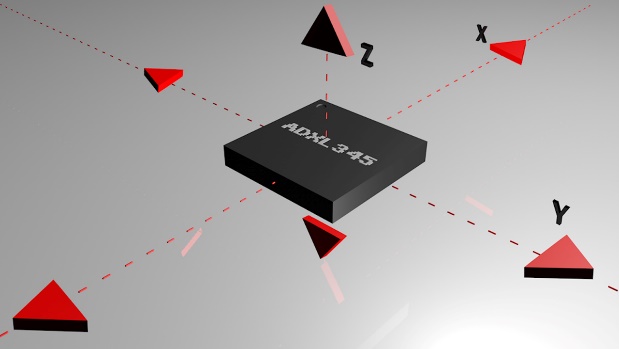
## 2.3. Tìm hiểu về cảm biến gia tốc

Cảm biến gia tốc đo rung động, hay còn gọi là gia tốc kế - Accelerometer là loại cảm biến chuyển rung động thành tín hiệu điện bằng cách sử dụng hiệu ứng áp điện.

 Dưới tác dụng của rung động (hoặc áp lực, ứng suất khác), áp điện được hình thành trong các tinh thể bên trong gia tốc kế áp điện. Hiệu ứng áp điện này là tuyến tính. Lượng điện tích sinh ra tỉ lệ với độ lớn của rung động. Vì vậy, chỉ cần một thay đổi rất nhỏ tác dụng vào cảm biến sẽ đưa ra một tín hiệu ra.

**Hình 2.2. Cấu tạo của cảm biến gia tốc**

**Nguyên lý hoạt động của cảm biến gia tốc:**



**Hình 2.3. Cảm biến giá tốc MCU 6050**

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.

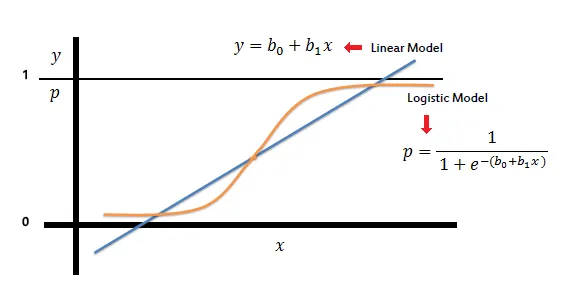
## 2.4. Các mô hình học máy trong phát hiện động đất

Tổng quan về các thuật toán học máy (như SVM, Random Forest, Neural Networks, …) ứng dụng vào bài toán.

### 2.4.1. Mô hình Logistic Regression

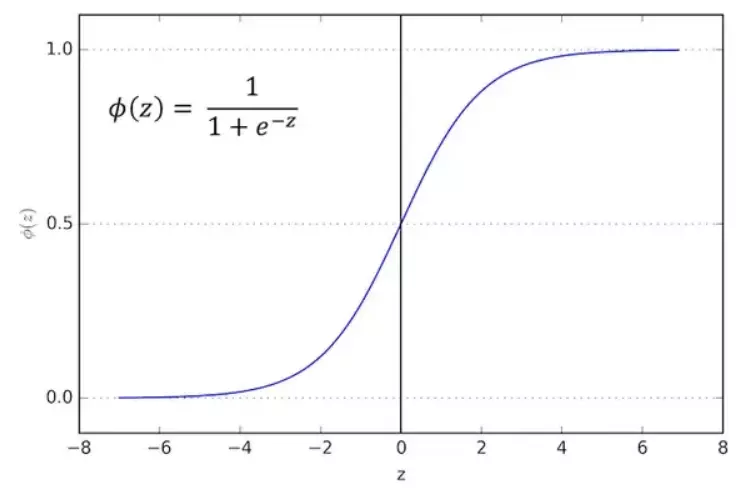
Mô hình học máy Logistic Regression là một trong những phương pháp phổ biến để thực hiện phân loại trong các bài toán học máy. Đặc biệt, nó thường được sử dụng khi muốn dự đoán xác suất của một sự kiện rơi vào một trong hai lớp (binary classification), ví dụ như dự đoán một email có phải là spam hay không. Cách hoạt động của Logistic Regression là dự đoán xác suất rơi vào một lớp dựa trên các biến đầu vào. Thay vì dự đoán giá trị cụ thể như trong hồi quy tuyến tính, Logistic Regression dự đoán xác suất sự kiện xảy ra bằng cách sử dụng một hàm sigmoid, giá trị của nó nằm trong khoảng (0, 1). Điều này làm cho nó phù hợp cho các bài toán phân loại.

Quá trình huấn luyện mô hình Logistic Regression thường được thực hiện bằng cách tối ưu hóa một hàm mất mát, chẳng hạn như Cross-Entropy Loss, để điều chỉnh các tham số sao cho mô hình dự đoán xác suất gần nhất với các nhãn thực tế. Một trong những điểm mạnh của Logistic Regression là nó có thể được hiểu và giải thích một cách tương đối đơn giản, và có thể áp dụng cho các tập dữ liệu lớn mà không gặp nhiều vấn đề về hiệu suất. Tuy nhiên, nó có thể không phù hợp cho các bài toán phức tạp hơn hoặc khi mối quan hệ giữa các biến đầu vào và kết quả không phải là tuyến tính.



**Hình 2.4. Logistic Regression**

Hồi quy Logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:



**Hình 2.5. Biểu diễn Hàm Sigmoid**

Hàm Sigmod nhận đầu vào là một giá trị z bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng [0,1]. Khi áp dụng vào mô hình Hồi quy Logistic với đầu vào là ma trận dữ liệu X và trọng số w, ta có 𝑧 = 𝑋𝑤. Việc huấn luyện của mô hình là tìm ra bộ trọng số 𝑤 sao cho đầu ra dự đoán của hàm Sigmoid gần với kết quả thực tế nhất. Để làm được điều này, ta sử dụng hàm mất mát (Loss Function) để đánh giá hiệu năng của mô hình. Mô hình càng tốt khi hàm mất mát càng nhỏ.

Hàm mất mát (Loss Function) là một hàm số được sử dụng để đo lường mức độ lỗi mà mô hình của chúng ta tạo ra khi dự đoán các kết quả từ dữ liệu đầu vào. Trong bài toán Hồi quy Logistic, chúng ta sử dụng hàm mất mát Cross-Entropy để đánh giá hiệu năng của mô hình. Hàm mất mát Cross-Entropy được định nghĩa như sau:



Trong đó:

n: số lượng mẫu dữ liệu trong tập huấn luyện.

𝑦𝑖​: giá trị thực tế của đầu ra thứ i.

𝑝𝑖 : xác suất dự đoán thuộc lớp 1 của mô hình cho đầu vào thứ i.

Hàm Cross-Entropy đo lường khoảng cách giữa hai phân phối xác suất 𝑦i và 𝑝i. Khi mô hình dự đoán chính xác, tức là nếu 𝑦i = 1 thì pi càng gần 1, và nếu 𝑦i = 0 thì pi càng gần 0, sau đó hàm mất mát sẽ tiến gần về 0. Trong quá trình huấn luyện, chúng ta tìm cách cập nhật bộ trọng số 𝑤 sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự đoán tốt nhất. Để tìm giá trị tối ưu cho bộ trọng số 𝑤, chúng ta có thể sử dụng kỹ thuật Gradient Descent. Tại mỗi bước lặp, chúng ta cập nhật 𝑤 theo phương tương ứng với đạo hàm của hàm mất mát L(w) theo w.

|  |  |
| --- | --- |
| **Ưu điểm** | **Nhược điểm** |
| Dễ hiểu, dễ cài đặt  Hiệu quả với dữ liệu tuyến tính  Có thể giải thích được ý nghĩa của các trọng số đặc trưng | Không xử lý tốt quan hệ phi tuyến (non-linear) nếu không biến đổi đặc trưng  Độ chính xác không cao nếu dữ liệu phức tạp hoặc nhiễu nhiều |

### 2.4.2. Mô hình học máy Decision Tree

### 2.4.3. Mô hình Random Forest

### 2.4.4. Mô hình Support Vector Machines

### 2.4.5. Mô hình ANN

### 2.4.6. Mô hình CRNN

## 2.5. Tổng quan các nghiên cứu liên quan

Trong những năm gần đây, việc sử dụng trí tuệ nhân tạo (AI) và học máy (Machine Learning) để phân biệt tín hiệu động đất và tín hiệu nhiễu đã thu hút sự quan tâm của nhiều nhà nghiên cứu trên thế giới. Dưới đây là một số công trình tiêu biểu:

* **Zhu & Beroza (2019) – “PhaseNet: A Deep-Neural-Network for Seismic Phase Picking”**

Phương pháp: Áp dụng mạng nơ-ron tích chập (CNN) để phân loại và xác định thời điểm xảy ra pha sóng địa chấn.

Ưu điểm: Mô hình đạt độ chính xác cao và phản hồi nhanh trong các mạng cảm biến dày đặc.

Nhược điểm: Yêu cầu lượng dữ liệu lớn và được gán nhãn kỹ lưỡng, khó triển khai trong điều kiện thiếu dữ liệu.

* **Perol et al. (2018) – “Convolutional Neural Network for Earthquake Detection”**

Phương pháp: Xây dựng mô hình CNN đơn giản nhưng hiệu quả để nhận diện các tín hiệu động đất từ dòng dữ liệu thời gian.

Ưu điểm: Khả năng tự động trích xuất đặc trưng và học biểu diễn phi tuyến tính.

Nhược điểm: Cần GPU để tăng tốc tính toán; độ tin cậy có thể giảm khi gặp nhiễu cao.

* **Xin Huang (2020) – “CrowdQuake: A Networked System of Low-Cost Sensors for Earthquake Detection via Deep Learning”**

Phương pháp: Áp dụng mô hình học sâu (Deep Learning) để phát hiện các pha sóng địa chấn một cách tổng quát, sử dụng mạng RNN và CNN.

Ưu điểm: Tổng quát hóa tốt, phát hiện được nhiều loại tín hiệu khác nhau.

Nhược điểm: Phức tạp khi triển khai trên hệ thống cảnh báo sớm do yêu cầu cấu hình cao

* **Yoon et al. (2015) – “Earthquake detection through low-frequency noise classification using Random Forest”**

Phương pháp: Dùng đặc trưng thống kê và phân loại bằng mô hình Random Forest.

Ưu điểm: Dễ triển khai, không yêu cầu GPU, hoạt động tốt với tập dữ liệu nhỏ.

Nhược điểm: Độ chính xác thấp hơn mô hình học sâu nếu dữ liệu phức tạp hoặc bị nhiễu.

**Phân tích tổng quan những nghiên cứu liên quan**

Ưu điểm chung:

* Tự động hóa: Các phương pháp học máy và học sâu giúp tự động hóa quá trình phát hiện và phân loại tín hiệu địa chấn, giảm thiểu sự can thiệp của con người.
* Độ chính xác cao: Khi được huấn luyện với dữ liệu chất lượng, các mô hình này có thể đạt độ chính xác vượt trội so với các phương pháp truyền thống.
* Xử lý thời gian thực: Một số nghiên cứu đã thành công trong việc triển khai mô hình cho phép xử lý và cảnh báo theo thời gian thực.

Nhược điểm chung:

* Yêu cầu dữ liệu lớn: Hiệu suất của các mô hình phụ thuộc nhiều vào lượng và chất lượng của dữ liệu huấn luyện.
* Tài nguyên tính toán: Các mô hình phức tạp, đặc biệt là học sâu, đòi hỏi tài nguyên tính toán mạnh mẽ, có thể là rào cản trong triển khai thực tế.
* Tính tổng quát hóa: Một số mô hình có thể hoạt động tốt trên dữ liệu huấn luyện nhưng gặp khó khăn khi áp dụng cho các khu vực địa chất khác nhau.

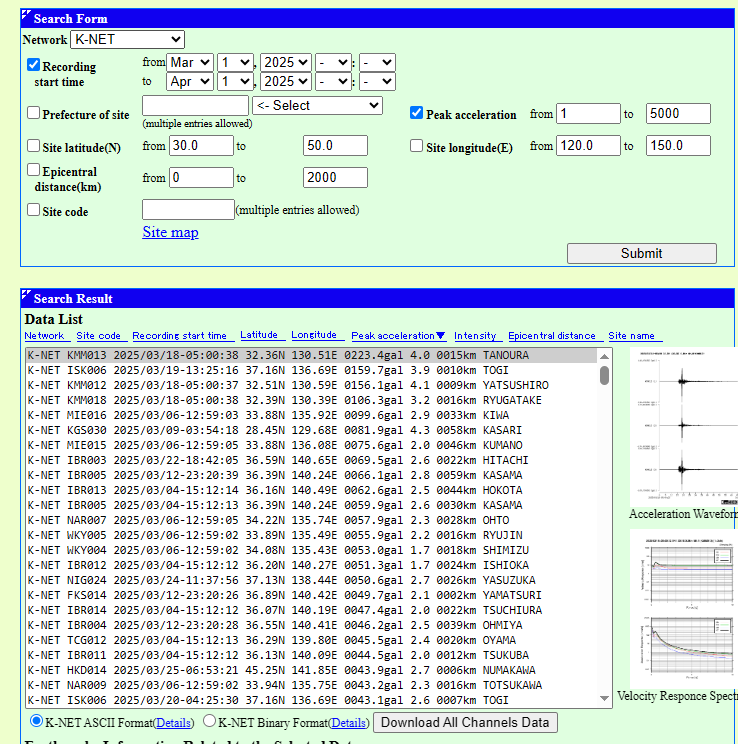
# CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

## 3.1. Nguồn dữ liệu và tiền xử lý

### 3.1.1. Nguồn dữ liệu cảm biến gia tốc

Dữ liệu được sử dụng trong nghiên cứu bao gồm tín hiệu gia tốc ba trục, được sử dụng từ bộ dữ liệu **K-NET** dữ liệu của **Kyoshin Network** - một mạng lưới quốc gia về quan trắc gia tốc nền đất mạnh ở Nhật Bản , được ghi nhận từ các cảm biến địa chấn đặt tại các trạm quan trắc cố định.

Bộ dữ liệu động đất là dữ liệu của gần 2000 trận động đất đươc trích từ năm 1996 đến năm 2021 từ trung tâm thuỷ văn Nhật Bản.

****** Dưới đây là cách truy cập là trích xuất các thông tin dữ liệu các trận động đất từ trang web chính thống của **Kyoshin Network.**

**Hình 3.1. Dữ liệu động đất từ trang chủ K-NET**

Một bản ghi dữ liệu từ trạm K-NET cho một trận động đất cụ thể thường bao gồm ba file dữ liệu riêng biệt, tương ứng với ba hướng chuyển động của mặt đất thu nhận bởi cảm biến gia tốc ba trục:

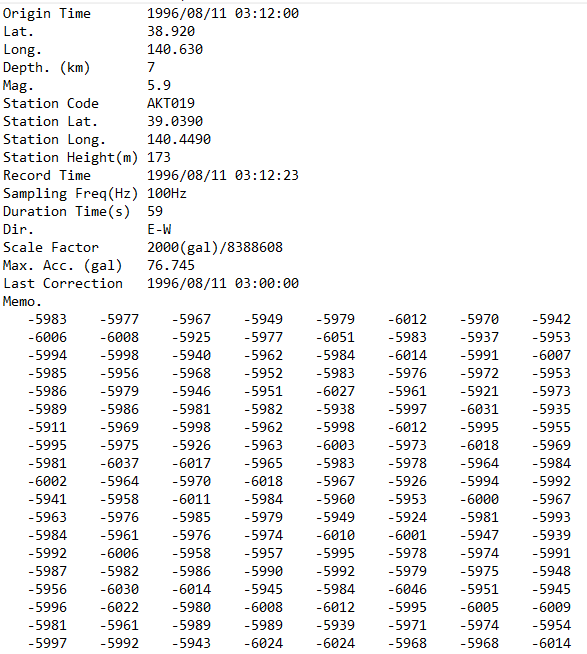
* N–S File (North–South): ghi lại gia tốc theo phương Bắc – Nam (trục ngang).
* E–W File (East–West): ghi lại gia tốc theo phương Đông – Tây (trục ngang, vuông góc với N–S).
* U–D File (Up–Down): ghi lại gia tốc theo phương thẳng đứng (trục Z).

Các tệp này thường có định dạng .txt hoặc .csv, chứa các giá trị gia tốc theo thời gian, được thu thập với tần số lấy mẫu cố định (thường là 100 Hz, tương ứng 100 mẫu/giây).

Mỗi file dữ liệu bao gồm:

* Thông tin metadata: như mã trạm, thời gian bắt đầu ghi nhận, toạ độ trạm, độ sâu chấn tiêu, cường độ trận động đất..
* Dữ liệu số: gồm các giá trị gia tốc ròng tại mỗi thời điểm, thường được lưu dưới dạng số nguyên 16-bit, cần chuyển đổi bằng hệ số scale factor để ra đơn vị đo gia tốc cm/s².

Dữ liệu từ ba file này khi kết hợp lại sẽ tạo thành chuỗi thời gian 3 chiều của tín hiệu gia tốc địa chấn, phản ánh đầy đủ chuyển động của mặt đất theo không gian 3 chiều tại thời điểm động đất xảy ra.

Hình ảnh dưới đây thể hiện ví dụ về một cấu trúc file dữ liệu của bộ K-NET:

**Hình 3.2. Dữ liệu gia tốc trả về theo hướng E-W của bộ K-NET**

Một bộ dữ liệu K-NET sẽ cung cấp cho người dùng những thông tin quan trọng như sau:

* Latitude (Lat.): Kinh độ của trận động đất
* Longtitude(Long.): Vĩ độ của trận động đất
* Depth: Thể hiện độ sâu của tâm trấn so với bề mặt, đơn vị thể hiện là km
* Mag : Thể hiện độ lớn của trận động đất theo đơn vị cường độ (Richter).
* Station code: Là mã của trạm đo
* Station Lat: vị trí của trạm đo
* Station Long: Vị trí của trạm đo
* Sampling Rate: 100Hz là tần số lấy mẫu của cảm biến, cứ 1s thu được 100 mẫu.
* Duration Time: là khoảng thời gian đo trong bao nhiêu giây.
* Direction: thể hiện hướng đo (NS: là theo hướng x, EW là theo hướng y, UD: là theo hướng z)
* Scale Factor: là hệ số chuyển đổi được sử dụng để đổi từ dữ liệu raw sang đơn vị gia tốc là gal.
* Memo: là những giá trị gia tốc trả về trong duration time.

### 3.1.2. Trực quan hoá dữ liệu theo miền gia tốc thời gian

Miền thời gian (Time Domain): là biểu diễn tín hiệu theo thời gian, trục hoành là thời gian, trục tung là gia tốc (thường là m/s² hoặc g). Dạng biểu đồ theo miền thời gian này giúp ta nhìn được dạng sóng gốc, giống như sóng địa chấn trên máy đo.

**Ý nghĩa:**

Nhìn vào biểu đồ miền Time Domain này giúp ta phát hiện được bất thường nhanh chóng.

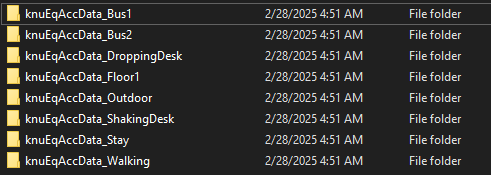
* Khi nào có rung động bất thường (động đất, va chạm…).
* Mức độ mạnh yếu (biên độ lớn hay nhỏ).
* Xác định được thời điểm bắt đầu của một trận động đất.
* Khi nào có P-wave, S-wave.
* Thời điểm bắt đầu và kết thúc của một trận động đất.

Trên đây là những hình ảnh trực quan hoá dữ liệu theo cả 3 trục dữ liệu x, y và z của một trận động động đất và một bộ dữ liệu nhiễu để cho ta thấy được sự phân bố dữ liệu.

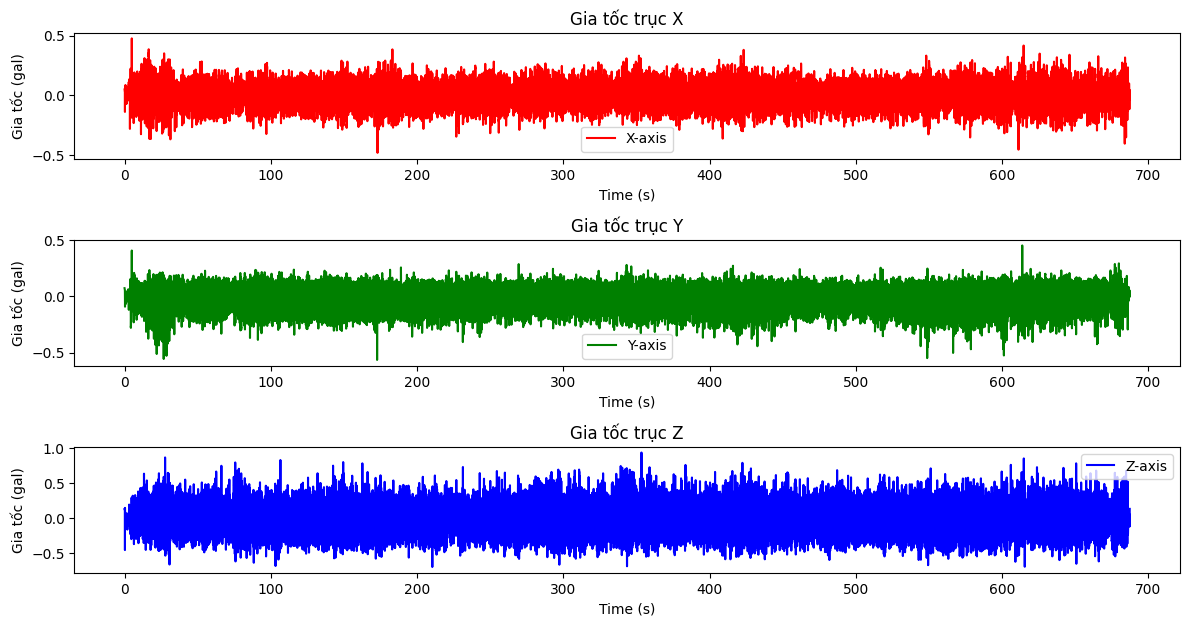


**Hình 3.3. Trực quan hoá dữ liệu động đất theo miền thời gian.**

Tập dữ liệu Noise:

 Tín hiệu nhiễu (Noise): là các đoạn tín hiệu không chứa hoạt động địa chấn, có thể là nhiễu nền từ môi trường (xe cộ, gió, hoạt động công trình, v.v.). Dữ liệu này được thu thập từ những khoảng thời gian không có sự kiện địa chấn nào xảy ra.

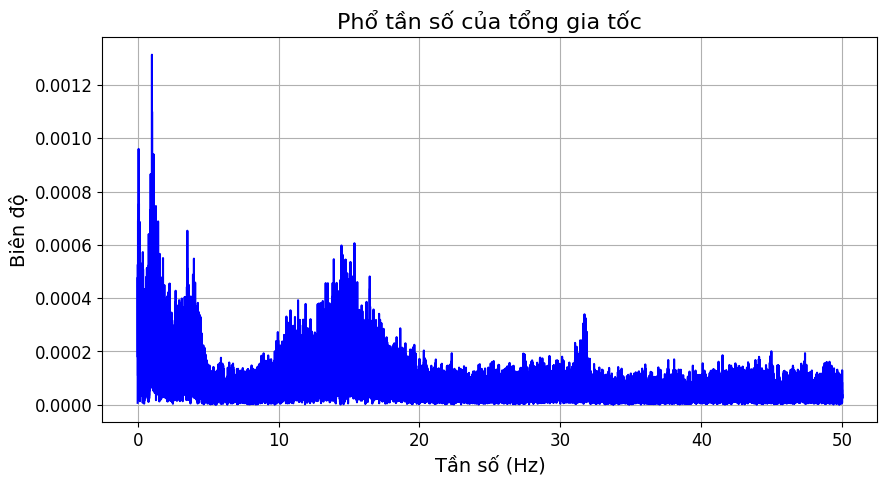
**Hình 3.4. Bộ dữ liệu nhiễu sử dụng trong quá trình**



**Hình 3.5. Trực quan hoá dữ liệu nhiễu theo miền thời gian**

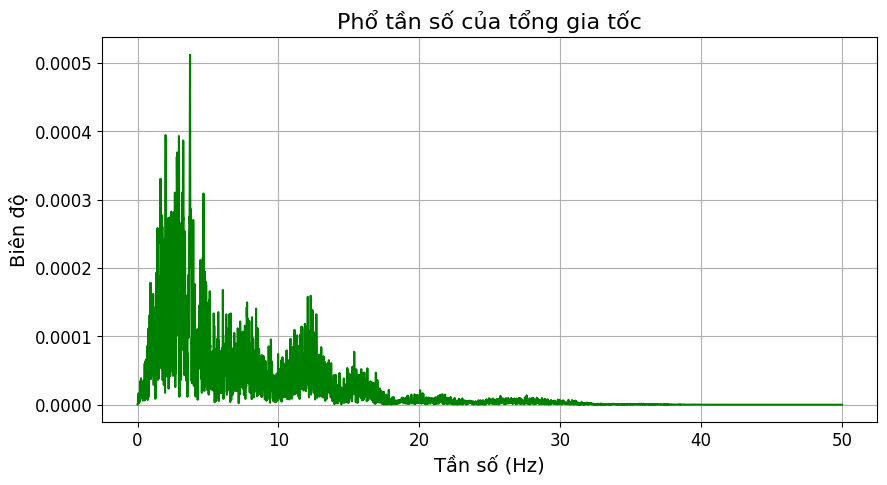
**Trực quan hoá ở miền tần số FFT**

Việc trực quan hóa dữ liệu gia tốc theo miền tần số (FFT) giúp làm rõ các đặc trưng tần số có trong tín hiệu. Tín hiệu động đất thường tập trung ở tần số thấp (dưới 10 Hz) với biên độ rõ rệt, trong khi nhiễu thường xuất hiện ở tần số cao, phân bố không đều và không có mẫu đặc trưng.

Tập dữ liệu Noise:

**Hình 3.6. Trực quan hoá dữ liệu nhiễu theo miền tần số FFT**

Tập dữ liệu động đất:



**Hình 3.7. Trực quan hoá dữ liệu động đất theo miền tần số FFT**

### 3.1.3. Trực quan hoá theo miền PSD (Power Spectral Density)

Power Spectral Density (PSD) là mật độ phổ công suất, mô tả năng lượng hoặc công suất của tín hiệu được phân bố như thế nào theo từng tần số. PSD cho biết tần số nào trong tín hiệu là mạnh nhất.

* Đối với động đất thì năng lượng sẽ tập trung trong 1-10 Hz.
* Đối với các nhiễu từ bên ngoài sẽ nằm trong dải tần số lớn hơn (Xe cộ: 20 – 60Hz).

***3.1.3.1. Năng lượng tập trung ở dải tần số thấp (1–10 Hz)***

* EQ thật thường có năng lượng mạnh tập trung ở vùng tần số thấp:
* Do bản chất sóng địa chấn (P-wave, S-wave, surface wave)
* Biểu đồ PSD thường có đỉnh rõ rệt trong khoảng 1–10 Hz

**Kết luận:** Nếu PSD có đỉnh cao trong dải 1–10 Hz rất có khả năng là EQ

***3.1.3.2. Độ lớn PSD cao hơn rõ rệt so với nhiễu nền***

* Nếu một trận động đất thường gây biên độ rung rất lớn dẫn theo mật độ năng lượng (g²/Hz) tăng vọt. Nếu PSD cao hơn nhiều lần so với PSD trong điều kiện nền bình thường có thể là dấu hiệu đáng tin cậy.

Kết luận: Nếu giá trị PSD > 1e-4 g²/Hz trong vùng tần số thấp → nghi ngờ EQ

***3.1.3.3. Dạng PSD có xu hướng "rơi nhanh" theo tần số***

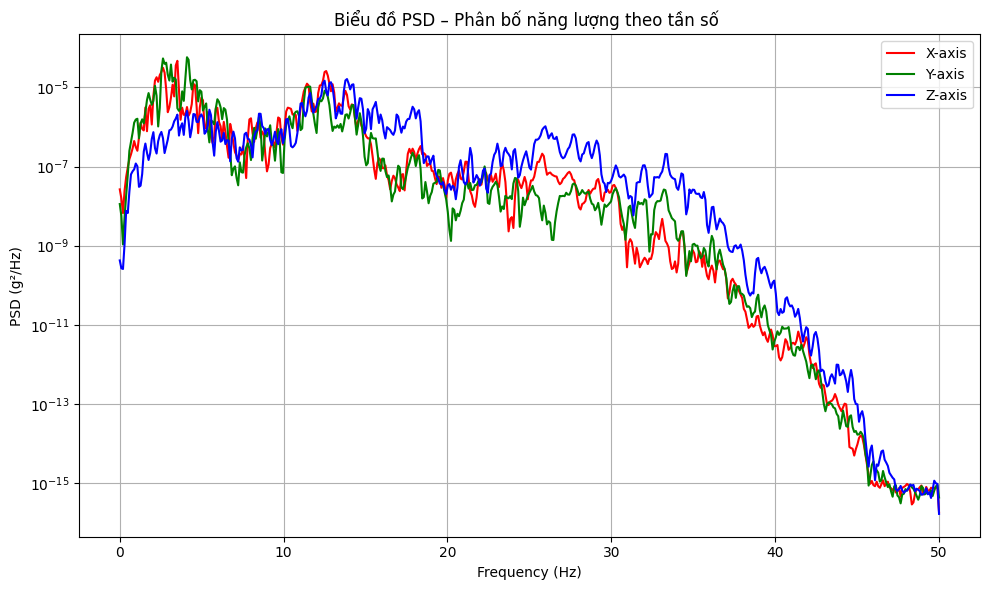
* PSD của EQ thường có dạng suy giảm mượt từ thấp lên cao.
* Biên độ lớn ở tần số thấp.
* Giảm dần khi tăng tần số.

Kết luận: Nếu đường cong PSD có dạng “gò đồi” rõ ràng rồi rơi dốc → là EQ.

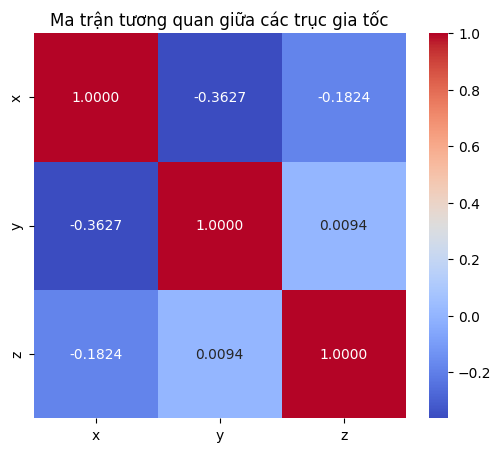
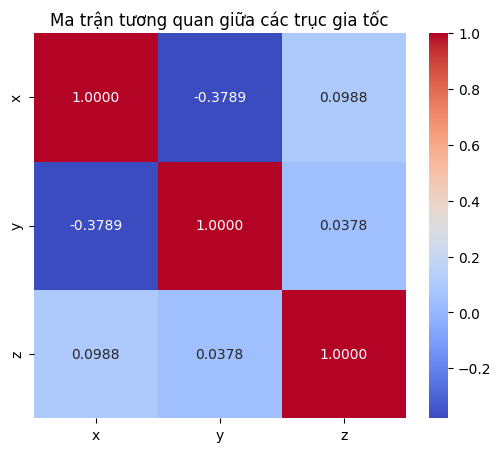
***3.1.3.4. Đặc điểm tương tự ở nhiều trục (x, y, z)***

* Sóng địa chấn tác động đến cả 3 trục nên PSD sẽ có dạng tương tự trên x, y, z.
* Nếu chỉ 1 trục có đỉnh bất thường → có thể là nhiễu do rung máy, va đập.

Kết luận: Nếu cả 3 trục cùng có đỉnh tương đồng → khả năng cao là EQ



**Hình 3.8. Miền PSD của một trận động đất**

**Biểu đồ tương quan giữa các giá trị theo 3 hướng của một trận động đất:

***Earthquake NonEarthquake***

**Hình 3.9. Biểu đồ heatmap thể hiện mỗi quan hệ giữa các trục**

## 3.2. Quá trình thu thập và xây dựng tập dữ liệu

Để tối ưu cho bộ dữ liệu và làm tăng tính chính xác, mỗi bộ dữ liệu được tổ chức và chia thành từng đoạn nhỏ (silding window) có độ dài cố định là **10 giây**, với tần số lấy mẫu của cảm biến là **100 Hz**.

Các đoạn tín hiệu được gán nhãn thủ công hoặc theo thông báo sự kiện từ hệ thống địa chấn (EQ hoặc Noise).

Tập dữ liệu động đất ban đầu ở định dạng file **mseed** sau đó chuyển đổi định dạng sang file .**csv** sau đó chuyển đổi sang dạng mảng để xử lý.

### 3.2.1. Tiền xử lý dữ liệu

Trước khi đưa vào mô hình học máy, các dữ liệu cần được trải qua bước tiền xử lý dữ liệu quan trọng, để giúp mô hình học chính xác, ổn tính, tính khả thi cao thì sẽ cần trải qua những quá trình sau đây.

***3.2.1.1 Chuyển đổi dữ liệu về đơn vị gia tốc***

Để chuyển đổi giá trị cảm biến raw (là dữ liệu dạng thô) sang đơn vị gia tốc (m/s²) hoặc gal, bạn cần biết vài thông tin cơ bản về cảm biến, đặc biệt là:

* Cảm biến đang sử dụng loại gia tốc kế nào (ví dụ: MEMS accelerometer).
* Dải đo của cảm biến (ví dụ: ±2g, ±4g, ±8g).
* Độ phân giải hoặc bit của cảm biến (ví dụ: 16-bit, 12-bit), vì điều này ảnh hưởng đến cách bạn chuyển đổi từ giá trị raw (thô) thành gia tốc.

Giả sử bạn đang sử dụng một gia tốc kế với dải đo ±g (±9.8 m/s²) và giá trị raw từ cảm biến là một số nguyên trong khoảng giá trị tối thiểu và tối đa của cảm biến. Sau đây là các bước để chuyển đổi giá trị cảm biến raw sang đơn vị gia tốc (m/s² hoặc gal):

* **Tìm hiểu về độ phân giải và dải đo của cảm biến**

Giả sử cảm biến của bạn có dải ±2g và sử dụng độ phân giải 16-bit. Để tính toán giá trị gia tốc từ dữ liệu raw, bạn cần tính toán số lệnh raw (tức là giá trị đọc được từ cảm biến) ứng với một giá trị gia tốc cụ thể.

* Dải đo ±2g có nghĩa là giá trị gia tốc tối đa là ±2g = ±19.6 m/s² (g = 9.8 m/s²).
* Với độ phân giải 16-bit, giá trị raw có thể nằm trong khoảng từ -32768 đến 32767.
* **Chuyển đổi giá trị raw thành gia tốc**

Để chuyển từ giá trị raw sang gia tốc, bạn dùng công thức:

Dải đo của cảm biến là ±2g, tức là 2g = 19.6 m/s².

Độ phân giải là 16-bit, nghĩa là các giá trị raw nằm trong khoảng từ -32768 đến 32767.

Công thức sẽ là:

Chuyển đổi đơn vị từ m/s2 sang đơn vị gal:

Đơn vị gal là đơn vị phổ biến trong các lĩnh vực như địa vật lý, đo rung động và nghiên cứu trọng lực

**1 gal = 1 cm/s2 = 0.01 m/s2**

Vậy để chuyển đổi từ đơn vị gia tốc (m/s2) sang đơn vị gal ta chỉ cần nhân với 100:

**Gia tốc (gal) = gia tốc (m/s2) \* 100**

### 3.2.2. Làm sạch dữ liệu huấn luyện

Quá trình làm sạch dữ liệu là chỉnh sửa và loại bỏ các trường dữ liệu không chính xác và chưa hoàn chỉnh, xác định và loại bỏ thông tin trùng lặp cũng như dữ liệu không liên quan và sửa lỗi định dạng, giá trị bị thiếu và lỗi chính tả hay bị mất mẫu, nhiễu lớn, hoặc có độ dài không phù hợp.

Có một vài cách có thể để xử lí trường hợp bị thiếu NaN (Not a Number) có thể kể đến như là lấy giá trị liền kề, tính giá trị trung bình trong một khoảng hay là loại bỏ luôn hàng đó.

***3.2.2.1 Loại bỏ dữ liệu ngoại lai (outliers)***

Để loại bỏ dữ liệu ngoại lai (Outliers) có rất nhiều phương pháp có thể sử dụng có thể kể đến như phương pháp **quantile, Z scores, độ lệch chuẩn** hay phương pháp khoảng tứ phân vị (**IQR**).

Loại bỏ giá trị ngoại lai làm cho mô hình của chúng ta khi huấn luyện sẽ giúp làm sạch dữ liệu từ đó làm tăng độ chính xác của mô hình, cải thiện khả năng tổng quát hoá và tránh gặp tình trạng overfitting và nhiễu không mong muốn.

Trong đề tài phân biệt dữ liệu động đất và nhiễu này thì em có sử dụng phương pháp phát hiện outlier dựa trên IQR.

Dưới đây là quá trình áp dụng phương pháp IQR vào sử lý dữ liệu:

Trước tiên chúng ta cần tính các giá trị phần tử Q1 và Q3.

* Q1 là phân vị thứ 25%
* Q3 là phân vị thứ 75%

Hay có một cách khác đó là sử dụng câu lệnh :

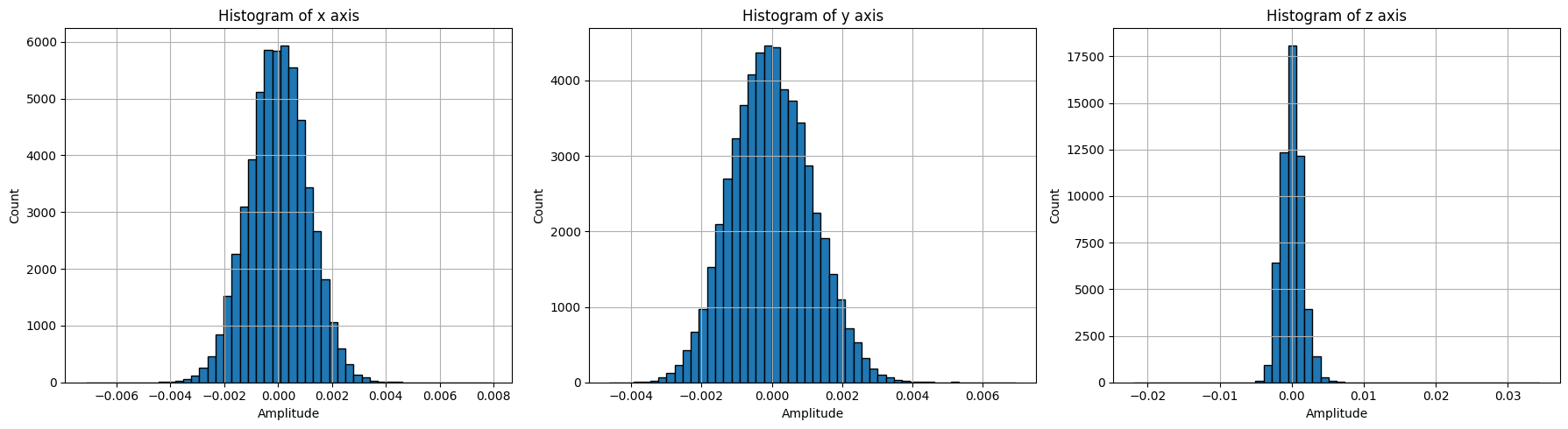
Từ đó tính được giá trị IQR bằng công thức:

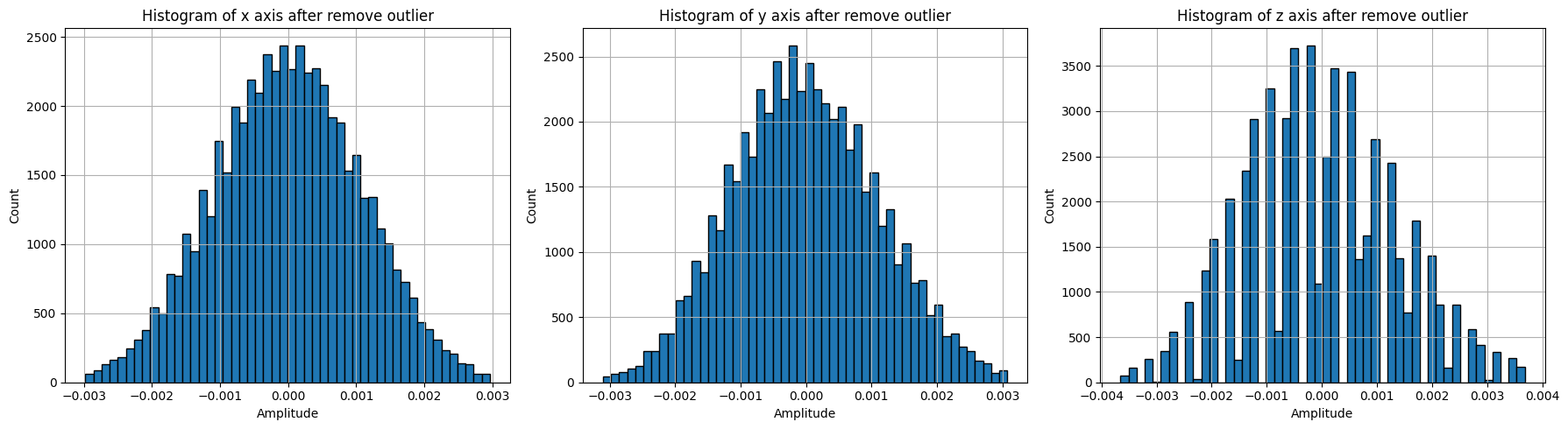
Sau đó tan cần xác định các ngưỡng threshold để loại bỏ, bằng việc sử dụng công thức tổng quát có sẵn đó là:

Những giá trị làm ngoài khoảng [Lower, upper] sẽ là những giá trị outlier.

Hình ảnh dưới đây thể hiện trực quan dữ liệu trước và sau khi đã loại bỏ những giá trị ngoại lai.

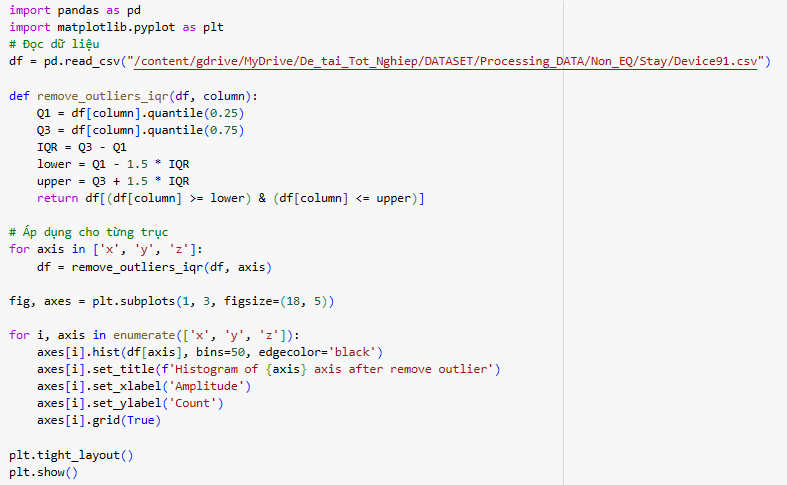
Trước khi loại bỏ OUTLIER:



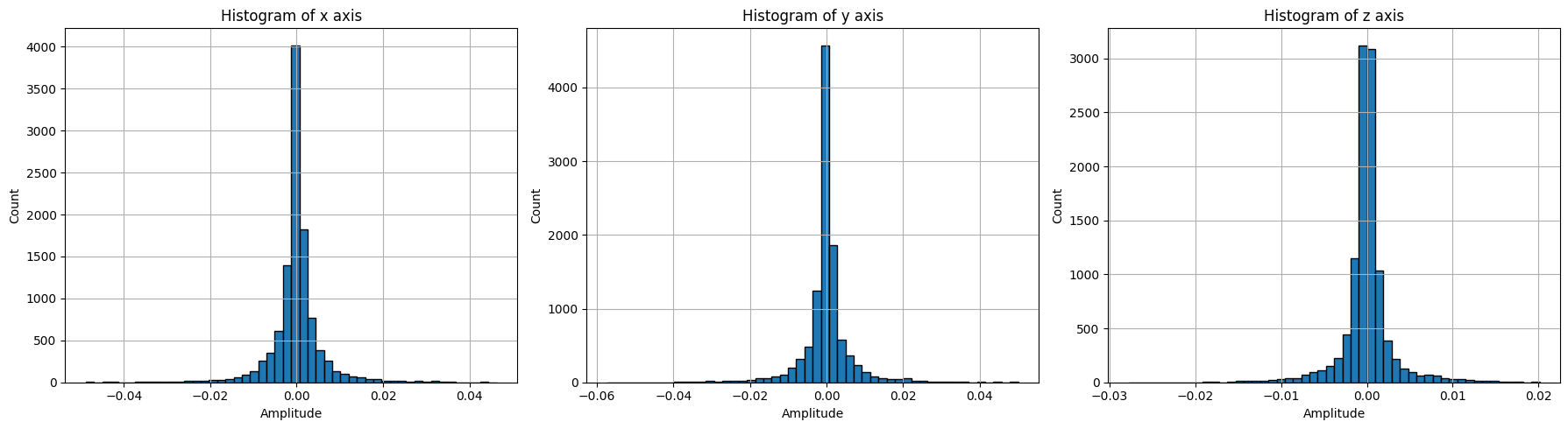
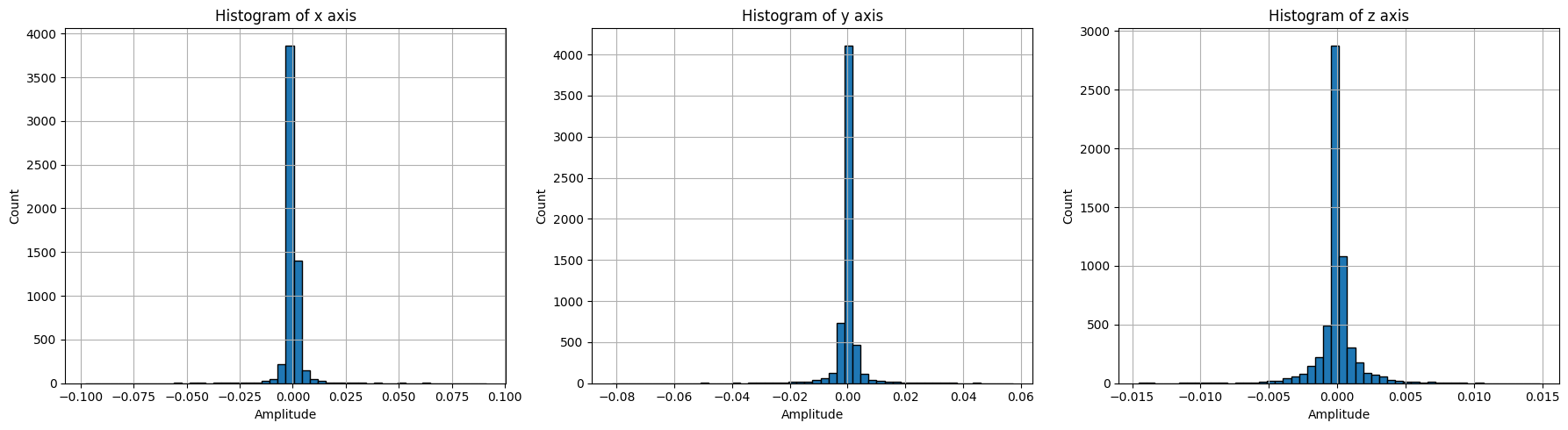
Trước khi loại bỏ OUTLIER:

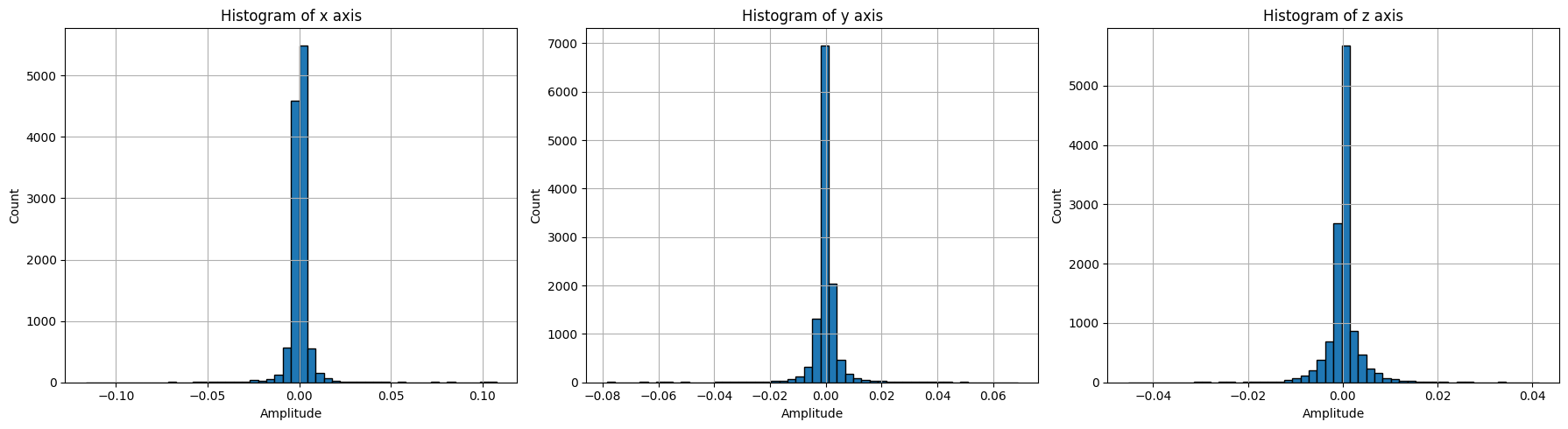
**Hình 3.10. So sánh trước và sau khi loại bỏ giá trị ngoại lai**

Code thực hiện vẽ trực quan:

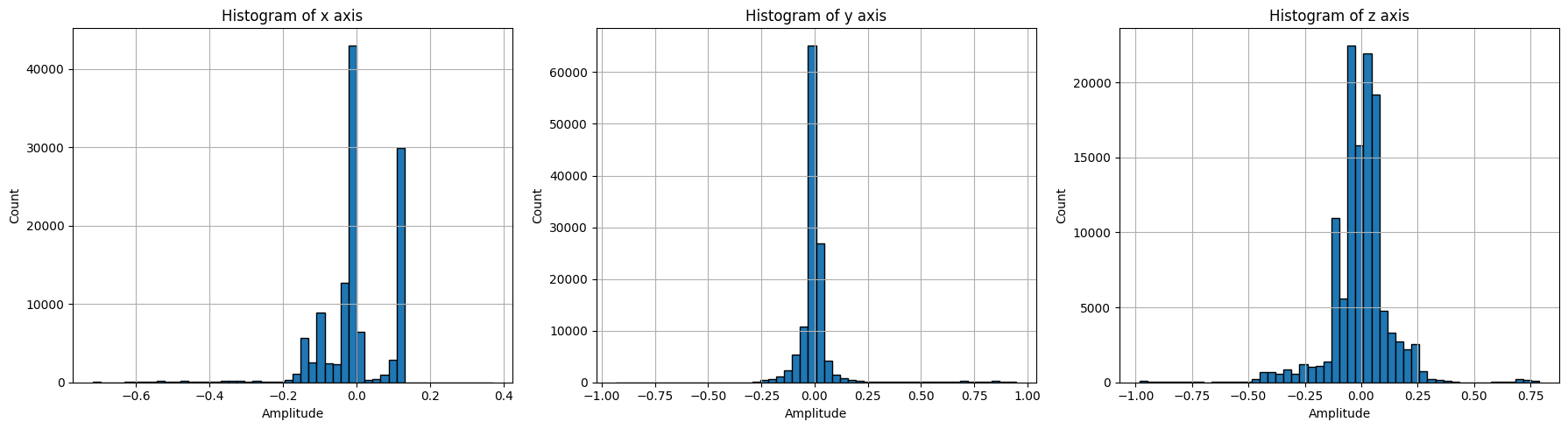


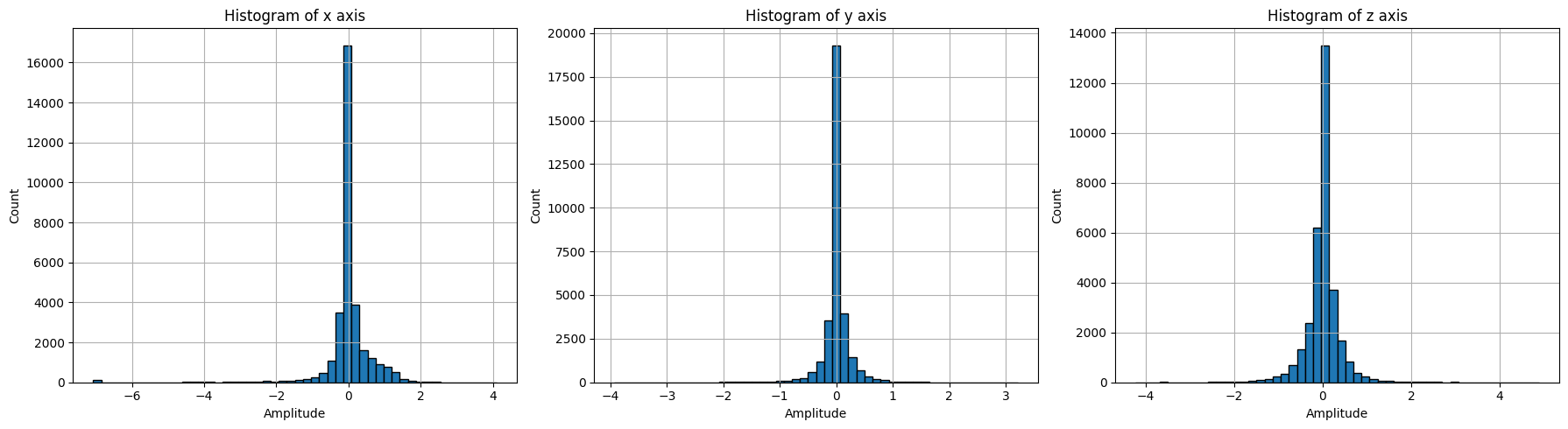
**Hình 3.11. Biểu đồ histogram sau khi loại bỏ ngoại lai theo phương pháp IQR**

 Biểu đồ histogram được sử dụng để mô tả phân bố dữ liệu gia tốc theo các trục x, y và z giúp quan sát sự phân tán và đặc điểm của tín hiệu địa chấn trên từng chiều không gian.

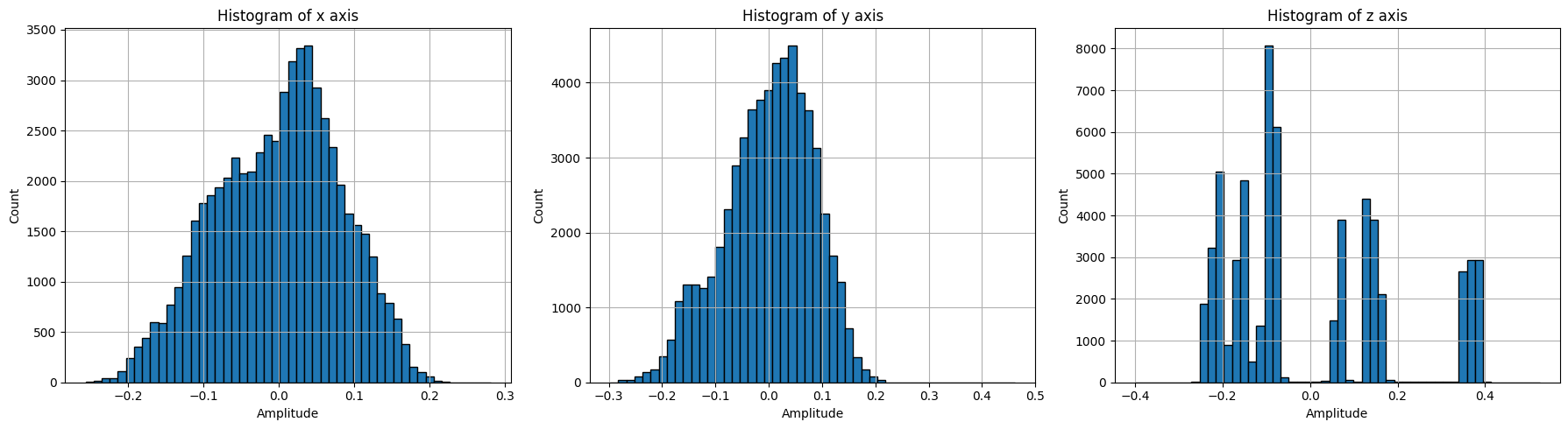
**Hình 3. 12. Biểu đồ histogram của nhóm dữ liệu động đất**

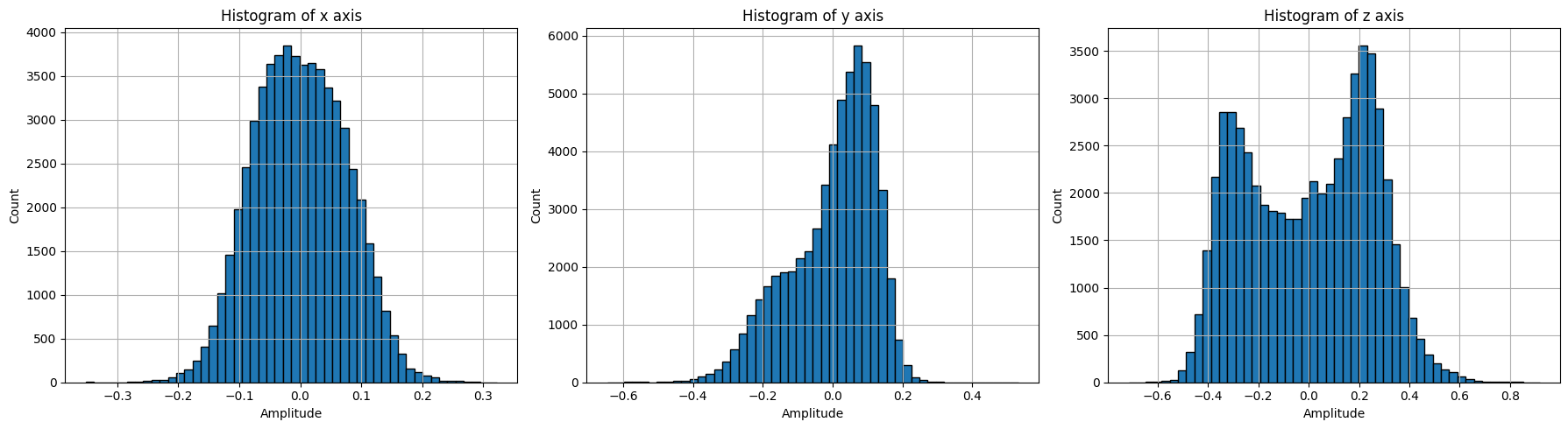
Biểu đồ histogram giúp khái quát phân bố dữ liệu trong tập dữ liệu noise.

****

****

Dữ liệu phân bố khi của data walking





***3.2.2.2. Chuẩn hóa dữ liệu***

Có một vài phương pháp thường để chuẩn hoá dữ liệu có thể nhắc tới như phương pháp Min-Max hay Z-socre (StandardScaler) giúp đưa những dữ liệu về cùng thang đo, giúp cho mô hình học tốt hơn và hội tụ nhanh hơn.

Do bộ dữ liệu sử dụng có những đặc trưng như Energy\_x, Skew\_x, Kurtosis\_z, Dominant\_freq\_z,… có thang đo dao động từ rất nhỏ 0.000001 đến rất lớn (lớn hơn 8000000).

Trong khi với những mô hình học máy như Support Vector Machine, KNN, Logistic Regression lại rất nhạy cảm với khoảng cách và phân phối dữ liệu nên việc chuẩn hoá là rất quan trọng giúp tăng hiệu quả mô hình.

Trong đó:

X là giá trị cần xét.

σ : độ lệch chuẩn

µ : Trung bình.

## 3.3. Trích xuất đặc trưng

Khi áp dụng một mô hình học máy- Machine Learning, việc quan trọng nhất đó là phải chọn ra được những đặc trưng (features) phù hợp. Đặc trưng tốt giúp cho mô hình học hiệu quả. Tuy nhiên với những đặc trưng không đủ mạnh thì cần phải chuyển sang các mô hình học sâu- nơi mà các mô hình tự học những đặc trưng từ dữ liệu.

Trong bài toán phân biệt dữ liệu động đất với dữ liệu nhiễu từ bên ngoài này, thì mỗi file dữ liệu, em có trích xuất những đặc trưng như sau:

### 3.3.1. IQR - Interquartile Range – Khoảng tứ phân vị

IQR có chức năng đo độ phân tán của dữ liệu (giữa Q3 và Q1), giúp nhận biết

biên độ biến động trung tâm. Dữ liệu nhiễu thường có IQR thấp hơn so với sự kiện động đất thực

**IQR = Q3 - Q1**

Trong đó:

Q1 (25%) là giá trị mà 25% dữ liệu nằm bên dưới.

Q3 (75%) là giá trị mà 75% dữ liệu nằm bên dưới.

IQR là vùng bao phủ 50% dữ liệu trung tâm, phản ánh mức độ biến động cốt lõi của tín hiệu.

**Ý nghĩa:**

* Nếu IQR nhỏ nghĩa là phần dữ liệu tập trung gần nhau => Tín hiệu ổn định, ít giao động.
* Nếu IQR lớn, dữ liệu phân tán rộng lớn => Tín hiệu có nhiều biến động mạnh, không ổn định.

**Trong bài toán phân biệt động đất và nhiễu:**

Tín hiệu động đất thường có IQR cao hơn vì trong giai đoạn chấn động, biên độ dao và thay đổi mạnh trong thời gian ngắn. Tín hiệu nhiễu nền như rung động từ môi trường, tiếng ồn cơ học nhẹ... thường có biên độ nhỏ và ổn định hơn nên giá trị IQR thấp hơn.

### 3.3.2. ZC- Zero Crossing Rate

Số lần tín hiệu đổi dấu. Tín hiệu động đất thường có tần suất đổi dấu khác biệt với nhiễu. Đây là đặc trưng đơn giản nhưng nhạy cảm với rung động.

* + 1. **CAV**

Tổng giá trị tuyệt đối của vận tốc (từ gia tốc tích phân). Chỉ số này liên quan đến cường độ rung động tích lũy — thường cao hơn trong động đất thật.

### 3.3.4. Dominant Frequenccy

Tần số xuất hiện năng lượng mạnh nhất. Động đất thường có phổ tần số đặc trưng, khác với nhiều dạng nhiễu môi trường (như xe chạy, gió...).

### 3.3.5. Energy

Tổng năng lượng tín hiệu (tổng bình phương biên độ). Dữ liệu động đất thật có năng lượng cao và bền vững hơn so với nhiễu.

### 3.3.6. Mean

Giá trị trung tâm của tín hiệu. Không quá đặc trưng cho động đất, nhưng giúp phát hiện thiên lệch tín hiệu (có thể do lỗi thiết bị).

### 3.3.7. Độ lệch chuẩn (Std)

Đo mức độ dao động của tín hiệu quanh trung bình. Động đất thường có Std cao hơn do rung động mạnh và rõ rệt.

Ý nghĩa:

* Nếu **độ lệch chuẩn nhỏ** nên các giá trị trong tập dữ liệu nằm gần nhau và gần với giá trị trung bình → tín hiệu **ổn định, ít dao động.**
* Nếu **độ lệch chuẩn lớn**, các giá trị phân tán rộng hơn quanh trung bình → tín hiệu **dao động mạnh, nhiều biến thiên.**

Trong bài toán phân biệt dữ liệu động đất và nhiễu này: Tín hiệu động đất thường có gia tốc biến đổi mạnh trong thời gian ngắn do đó sẽ có độ lệch chuẩn cao hơn so với tín hiệu nhiễu yếu hoặc nhiễu nền đều.

Dữ liệu nhiễu do gió, xe chạy hay nhiễu điện thường dao động nhẹ và đều nên **độ lệch chuẩn thấp hơn.**

Giả sử x1, x2, … , xn  là các số thực và xác định hàm, công thức tính độ lệch chuẩn dữ liệu:

### 3.3.8. Peak to Peak

Hiệu giữa giá trị cực đại và cực tiểu trong đoạn tín hiệu. Thể hiện độ "bất ngờ" hoặc cường độ tối đa của tín hiệu.

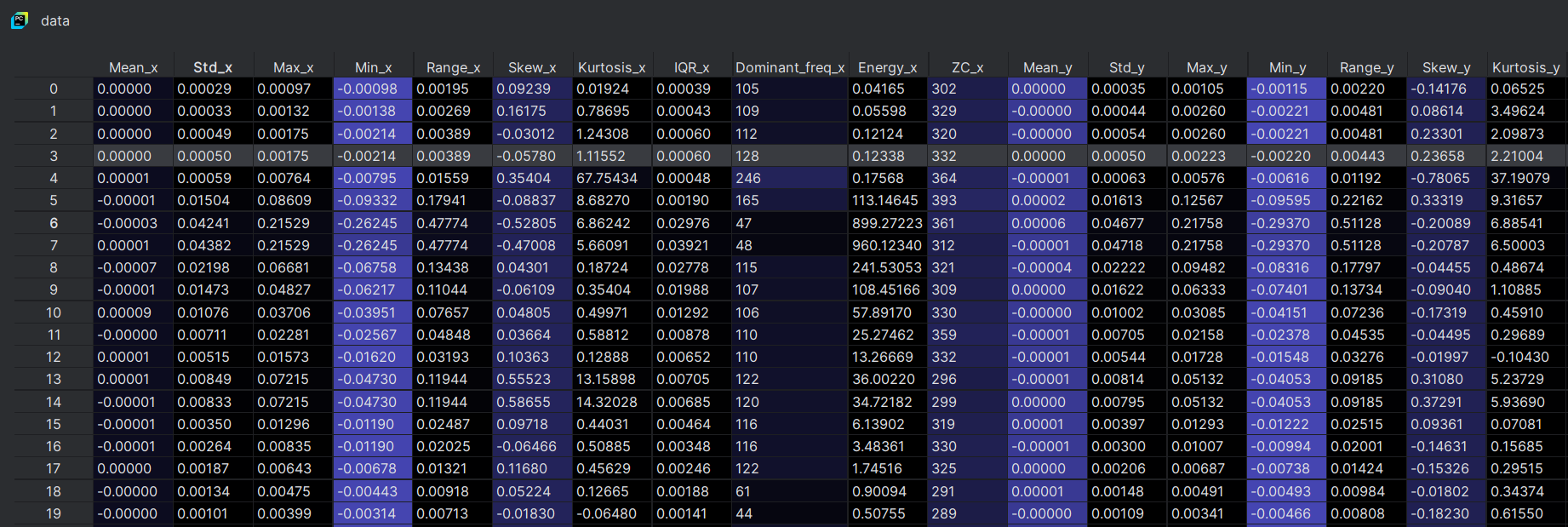
### 3.3.9. Skew

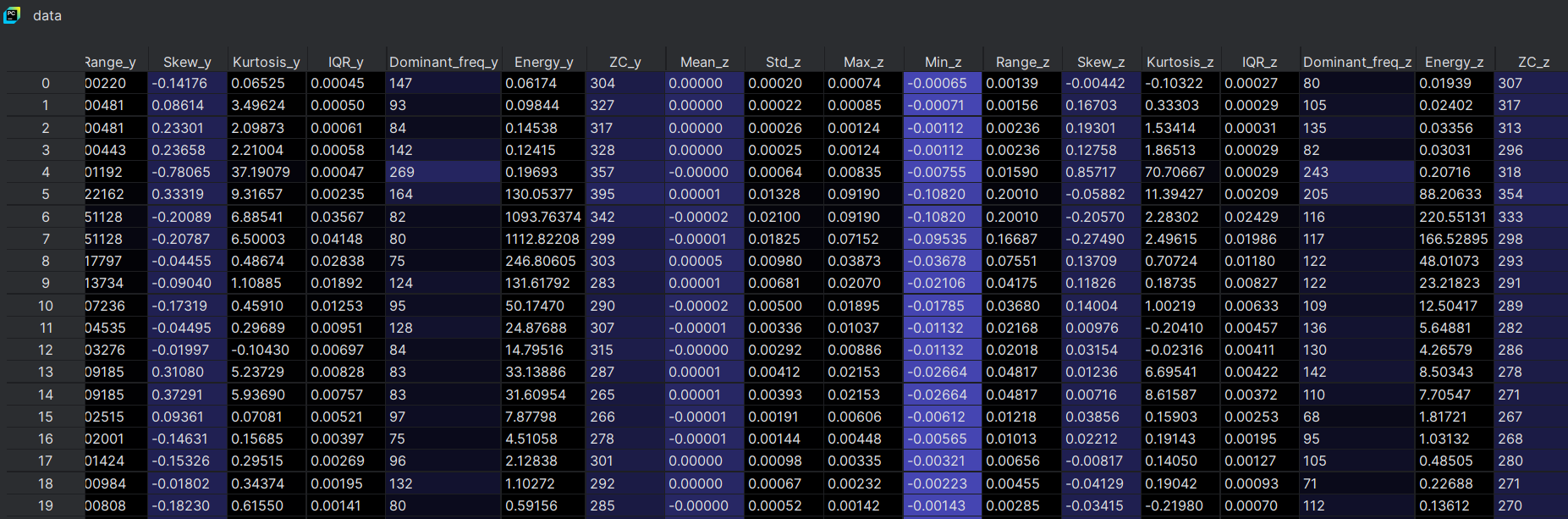
Đo tính bất đối xứng của phân phối tín hiệu. Dữ liệu động đất có thể tạo ra phân phối lệch rõ hơn so với nhiễu đều hoặc ngẫu nhiên.

### 3.3.10. Kurtosis

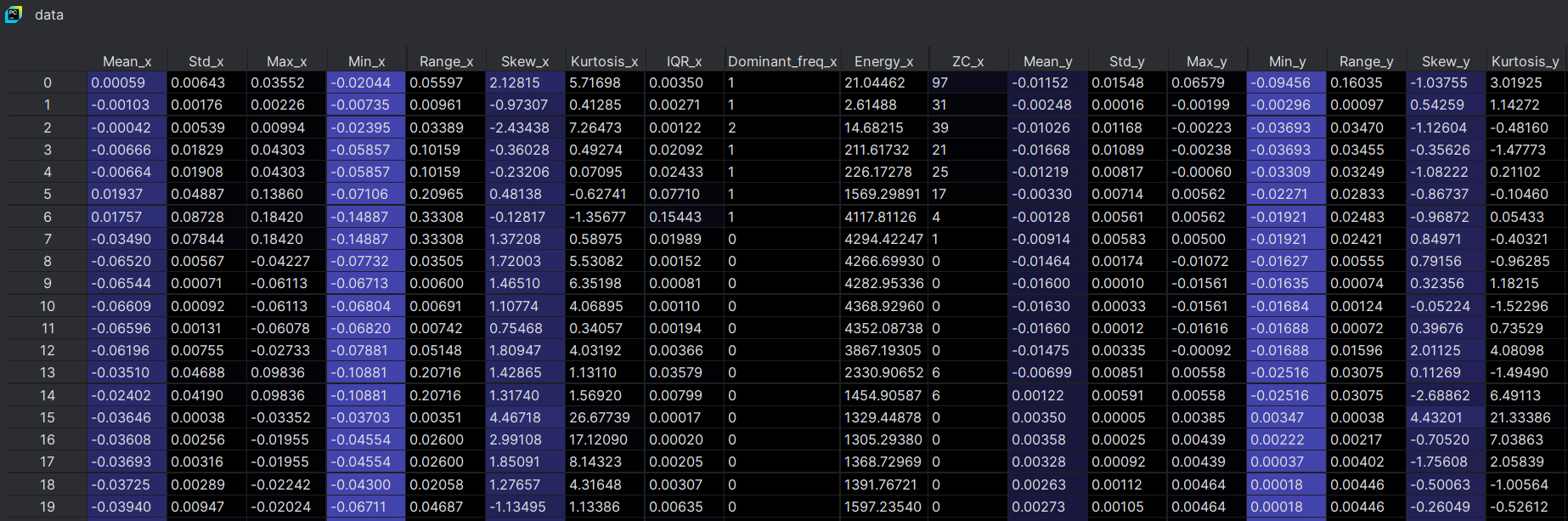
Đo độ nhọn hoặc “tập trung” của phân phối. Kurtosis cao thường chỉ ra sự kiện hiếm nhưng cường độ cao — rất phù hợp để phát hiện cú sốc trong tín hiệu động đất.

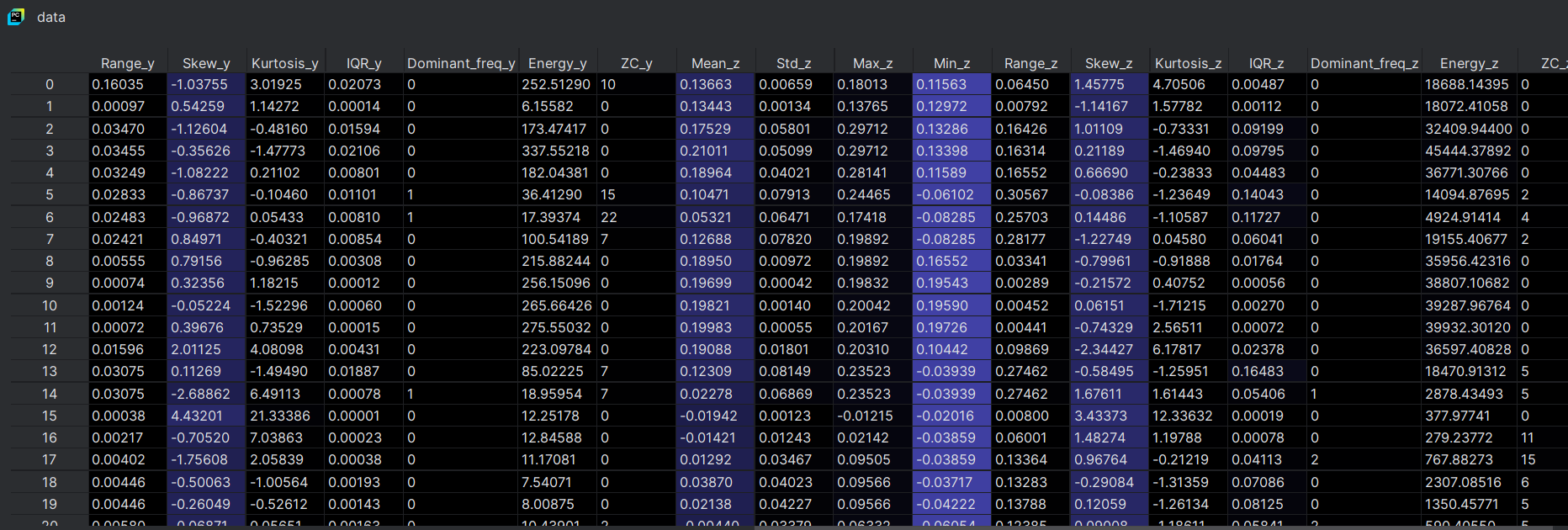
Theo gia tốc của mỗi trục x, y, z sẽ trích xuất ra được 10 đặc trưng của mỗi trục. Suy ra, theo ba trục sẽ có 3 \* 10 = 30 đặc trưng.





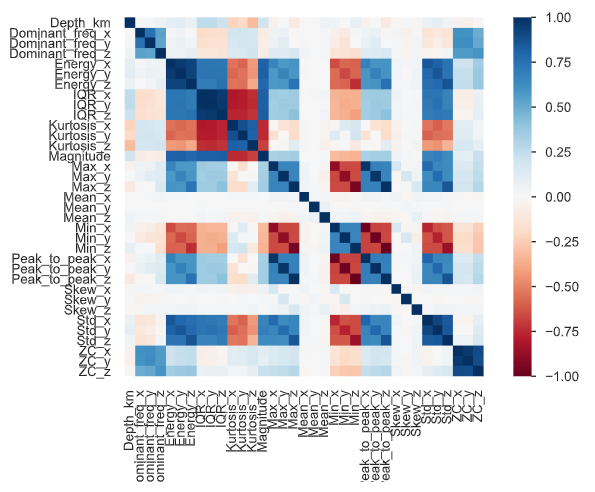
**Hình 3.13. Các đặc trưng sau khi trích xuất của tập EQ**





**Hình 3.14. Các đặc trưng sau khi trích xuất của tập noise**

Mối tương quan giữa các đặc trưng:



**Hình 3.15. Biểu đồ thể hiện tương quan giữa các đặc trưng**

Để cân bằng giữa độ chính xác và độ phức tạp, đây là cách mà em đã chọn ra những đặc trưng để đưa vào những mô hình học máy ML:

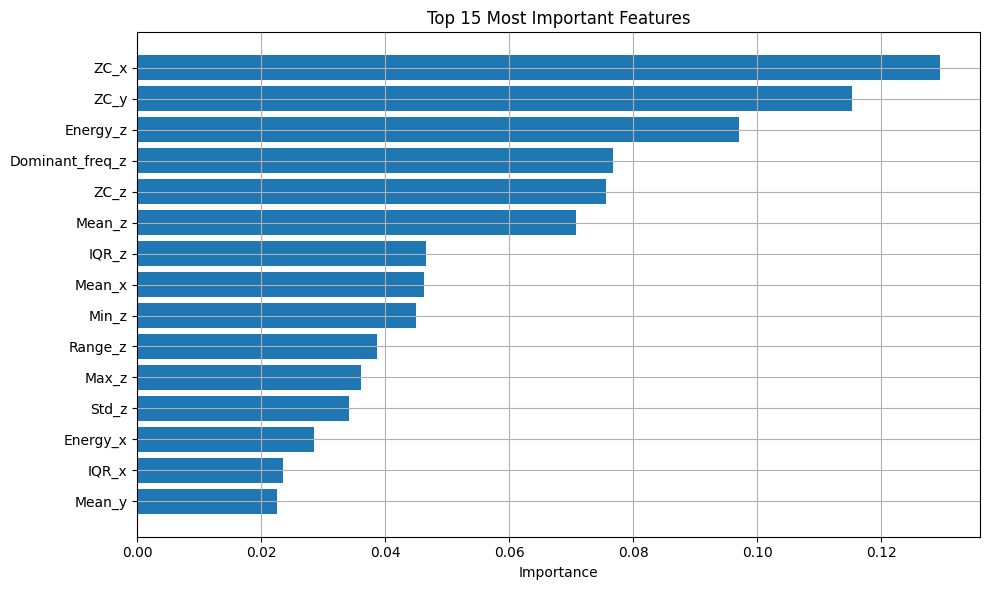
* Nếu Too few features (quá ít) dẫn đến không đủ thông tin làm cho mô hình thiếu chính xác.
* Còn nếu Too many features: dễ gây overfitting làm tăng thời gian huấn luyện, khó diễn giải, có thể chứa nhiều nhiễu (noise).

Chọn top-N đặc trưng quan trọng (như top 10, 15, 20) giúp giảm số chiều dữ liệu (dimensionality reduction) nhưng vẫn giữ lại thông tin cốt lõi.

15 đặc trưng thường là mức “đẹp” cho nhiều bài toán vừa đủ thông tin, vừa không quá phức tạp**.**

Cách lựa chọn dựa trên: Feature Importance từ mô hình **Random Forest**

* Random Forest đo lường độ giảm impurity (nhiễu) do mỗi đặc trưng mang lại khi phân chia nhánh.
* Đặc trưng nào càng “giúp ích” trong việc phân biệt EQ và Non-EQ sẽ có importance cao hơn.



**Hình 3.16. Top 15 đặc trưng quan trọng được lấy từ mô hình Random Forest**

Các đặc trưng này tạo thành vector đặc trưng để đưa vào mô hình học máy.

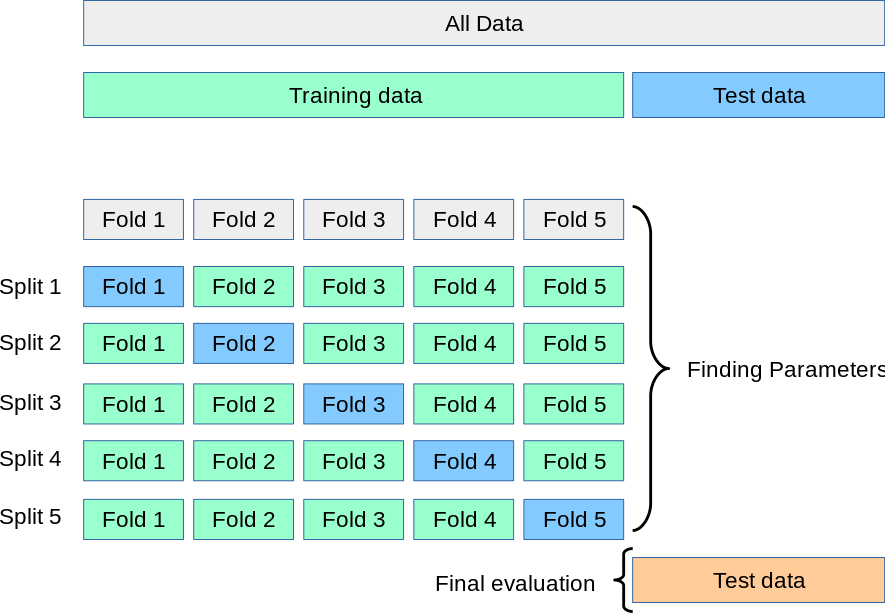
## 3.4. Xây dựng mô hình học máy

Trong phạm vi đề tài này, quá trình xây dựng mô hình học máy được thực hiện thông qua việc sử dụng các thuật toán có sẵn đã được triển khai sẵn trong thư viện **scikit-learn**. Tuy không phát triển thuật toán từ đầu, nhưng các bước thực hiện bao gồm: tiền xử lý dữ liệu, trích xuất đặc trưng từ tín hiệu cảm biến thô, lựa chọn thuật toán phù hợp (Logistic Regression, SVM, Random Forest,...), huấn luyện mô hình trên tập dữ liệu đã chuẩn bị, đánh giá hiệu quả mô hình qua các chỉ số như độ chính xác, F1-score, AUC-ROC, và cuối cùng là triển khai ứng dụng thực tế. Toàn bộ quy trình trên đáp ứng đầy đủ các giai đoạn chính trong việc xây dựng một hệ thống học máy hoàn chỉnh, theo đúng quy trình chuẩn trong lĩnh vực Machine Learning ứng dụng. Việc sử dụng các thuật toán có sẵn là cách tiếp cận phổ biến và được khuyến khích trong thực tiễn nhằm đảm bảo độ tin cậy, tính tái sử dụng và hiệu quả phát triển hệ thống.

**Phương pháp huấn luyện và điều chỉnh tham số tối ưu:**

Mô hình học máy được huấn luyện theo phương pháp giám sát (supervised learning), trong đó mỗi mẫu dữ liệu đầu vào tương ứng với một nhãn phân loại (0: nhiễu, 1: động đất). Dữ liệu được chia thành hai phần: tập huấn luyện và tập kiểm tra, đảm bảo phân phối nhãn đều bằng kỹ thuật stratified split. Quá trình huấn luyện được thực hiện thông qua các thuật toán trong thư viện scikit-learn, bao gồm: Random Forest, Logistic Regression và Support Vector Machine (SVM).

Trong quá trình huấn luyện, các tham số điều chỉnh (hyperparameters) được lựa chọn thủ công dựa trên hiệu quả thực nghiệm, hoặc sử dụng kỹ thuật **cross-validation** (K-fold) để tìm tập tham số tối ưu. Ví dụ, với mô hình Random Forest, các siêu tham số như số lượng cây (n\_estimators), độ sâu tối đa (max\_depth), và trọng số lớp (class\_weight) đã được điều chỉnh để cân bằng giữa độ chính xác và khả năng khái quát hóa. Đối với mô hình SVM, các tham số như hàm kernel (rbf), hệ số phạt (C) và class\_weight cũng được tối ưu hoá nhằm xử lý dữ liệu mất cân bằng.



**Hình 3.17. Kỹ thuật cross-validation (K-folder)**

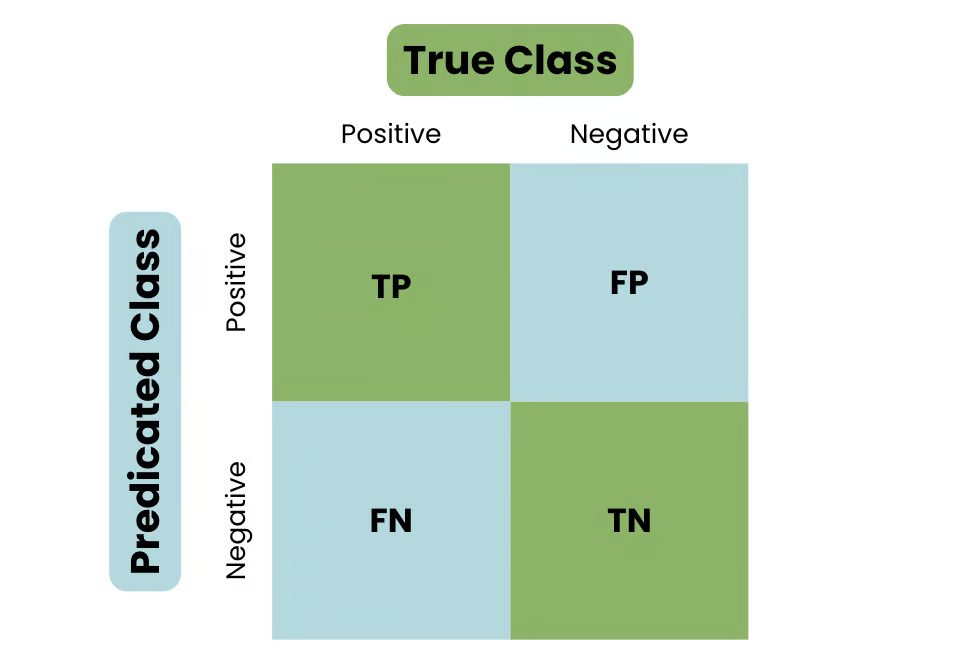
Ngoài ra, dữ liệu đầu vào được chuẩn hóa bằng StandardScaler trước khi huấn luyện để đảm bảo tính ổn định của các thuật toán như SVM và Logistic Regression. Việc đánh giá mô hình được thực hiện qua các chỉ số như Accuracy, Precision, Recall, F1-score và AUC-ROC, đảm bảo tính khách quan trong so sánh hiệu quả giữa các mô hình.

## ****3.5. Các tiêu chí đánh giá mô hình****

Trong quá trình xây dựng một mô hình Machine Learning, một phần không thể thiếu để xét xem mô hình có chất lượng tốt hay không chính là đánh giá mô hình. Đánh giá mô hình giúp chúng ta chọn lựa được các mô hình phù hợp với bài toán cụ thể. Để có thể áp dụng đúng thước đo đánh giá mô hình phù hợp, chúng ta cần hiểu bản chất, ý nghĩa cũng như các trường hợp sử dụng nó.

Để đánh giá hiệu quả của mô hình phân loại tín hiệu động đất và nhiễu, nghiên cứu sử dụng các tiêu chí phổ biến trong bài toán Classification. Một số metrics để đánh giá mô hình phân loại chính xác hay không bao gồm:

### 3.5.1. Confusion Matrix

Confusion matrix là một công cụ cơ bản nhưng vô cùng quan trọng trong việc đánh giá hiệu quả của các mô hình học máy, đặc biệt trong các bài toán phân loại. Nó cung cấp một cách trực quan và chi tiết để hiểu được hiệu suất của mô hình bằng cách so sánh giữa các nhãn dự đoán và nhãn thực tế trong tập kiểm tra. Ma trận này thường được biểu diễn dưới dạng một bảng vuông, trong đó các hàng đại diện cho nhãn thực tế và các cột đại diện cho nhãn dự đoán của mô hình.Trong trường hợp phân loại nhị phân, confusion matrix gồm bốn ô chính: True Positive (TP), False Positive (FP), True Negative (TN), và False Negative (FN). TP là số lượng trường hợp mà mô hình dự đoán đúng là dương tính và thực tế cũng là dương tính. TN là số lượng mẫu mà mô hình dự đoán đúng là âm tính và thực tế cũng là âm tính. FP là những trường hợp mô hình dự đoán nhầm là dương tính trong khi thực tế là âm tính. Ngược lại, FN là những trường hợp mô hình dự đoán nhầm là âm tính trong khi thực tế là dương tính.

**Hình 3.18. Ma trận nhầm lẫn (Confusion matrix)**

Từ confusion matrix, ta có thể tính được nhiều chỉ số quan trọng để đánh giá mô hình. Độ chính xác (accuracy) là tỷ lệ dự đoán đúng trên tổng số mẫu, thường được dùng như một chỉ số tổng quát. Tuy nhiên, trong các tập dữ liệu mất cân bằng, độ chính xác có thể gây hiểu lầm vì mô hình có thể đạt độ chính xác cao ngay cả khi không phát hiện được lớp thiểu số. Do đó, các chỉ số khác như độ nhạy (recall), độ đặc hiệu (specificity), độ chính xác (precision) và F1-score được sử dụng để đánh giá mô hình một cách toàn diện hơn.

### ****3.5.2. Accuracy (Độ chính xác tổng thể)****

Độ chính xác (accuracy) là một trong những chỉ số cơ bản và phổ biến nhất được sử dụng để đánh giá hiệu suất của mô hình học máy, đặc biệt trong các bài toán phân loại. Accuracy đo lường tỷ lệ dự đoán đúng của mô hình so với tổng số trường hợp được kiểm tra. Cụ thể, nó cho biết trong toàn bộ dữ liệu đầu vào, mô hình đã phân loại đúng bao nhiêu phần trăm mẫu.

Về mặt công thức, độ chính xác được tính bằng tổng số dự đoán đúng chia cho tổng số mẫu trong tập dữ liệu kiểm tra. Trong bài toán phân loại nhị phân, công thức accuracy được biểu diễn như sau:

Trong đó, TP (True Positive) và TN (True Negative) là số lượng các trường hợp mô hình dự đoán đúng, còn FP (False Positive) và FN (False Negative) là các trường hợp dự đoán sai. Giá trị của accuracy dao động từ 0 đến 1, với giá trị càng gần 1 thể hiện mô hình càng chính xác.

Tuy nhiên, accuracy có thể gây hiểu lầm trong những trường hợp dữ liệu mất cân bằng, tức là khi số lượng mẫu thuộc các lớp khác nhau chênh lệch lớn. Ví dụ, trong một bài toán phân loại bệnh hiếm mà chỉ có 5% bệnh nhân thực sự mắc bệnh, một mô hình luôn dự đoán "không mắc bệnh" có thể đạt độ chính xác 95%, nhưng lại hoàn toàn vô dụng trong việc phát hiện những trường hợp bệnh thực sự. Trong tình huống như vậy, các chỉ số khác như precision, recall và F1-score sẽ phản ánh rõ hơn năng lực của mô hình.

Tóm lại, mặc dù accuracy là một chỉ số dễ tính toán và dễ hiểu, việc sử dụng nó để đánh giá mô hình cần được đặt trong bối cảnh phù hợp. Đối với các bài toán có phân bố lớp không đồng đều, nên kết hợp thêm các chỉ số khác nhằm đảm bảo đánh giá toàn diện và chính xác hơn về hiệu suất của mô hình học máy.

=> Accuracy cao cho biết mô hình tổng thể hoạt động tốt, nhưng có thể bị đánh lừa nếu dữ liệu bị mất cân bằng.

### ****3.5.3. Precision (Độ chính xác của dự đoán dương tính)****

Precision, hay còn gọi là độ chính xác trong dự đoán dương tính, là một chỉ số quan trọng trong việc đánh giá hiệu suất của mô hình phân loại, đặc biệt là trong các tình huống mà sai sót dương tính giả (false positive) có thể gây ra hậu quả nghiêm trọng. Precision cho biết trong số tất cả các mẫu mà mô hình dự đoán là thuộc lớp dương, thì có bao nhiêu mẫu thực sự đúng với thực tế.

Về mặt công thức, precision được tính như sau:

Trong đó, TP (True Positive) là số lượng mẫu mà mô hình dự đoán đúng là dương tính, còn FP (False Positive) là số lượng mẫu mà mô hình dự đoán sai là dương tính trong khi thực tế là âm tính.

Precision cao thể hiện rằng mô hình có khả năng dự đoán dương tính một cách đáng tin cậy, tức là mô hình rất ít khi "nhầm lẫn" giữa các mẫu thuộc lớp âm và lớp dương.

### ****3.5.4. Recall (Độ bao phủ hay độ nhạy)****

Recall cũng là một metric quan trọng, nó đo lường tỷ lệ dự báo chính xác các trường hợp positive trên toàn bộ các mẫu thuộc nhóm positive.

Công thức của Recall như sau:

Recall cao nghĩa là mô hình không bỏ sót nhiều trận động đất.

### ****3.5.5. F1-score****

Trong các bài toán học máy phân loại nhị phân như phân biệt tín hiệu động đất (EQ) với nhiễu từ môi trường, việc lựa chọn chỉ số đánh giá phù hợp đóng vai trò then chốt để phản ánh hiệu suất thực sự của mô hình. F1-score là một chỉ số tổng hợp, được sử dụng phổ biến khi cần cân bằng giữa precision (độ chính xác của các dự đoán dương tính) và recall (khả năng phát hiện đầy đủ các trường hợp dương tính). Đây là một thước đo đặc biệt hữu ích khi bài toán có dữ liệu mất cân bằng, hoặc khi cả hai loại sai sót – bỏ sót sự kiện động đất và cảnh báo sai – đều có thể gây hậu quả nghiêm trọng.

F1-score được tính theo công thức:

Nếu một mô hình có precision cao nhưng recall thấp, hoặc ngược lại, thì F1-score sẽ phản ánh điều đó bằng cách cho ra một giá trị cân bằng – không bị "đánh lừa" bởi một chỉ số riêng lẻ quá cao.

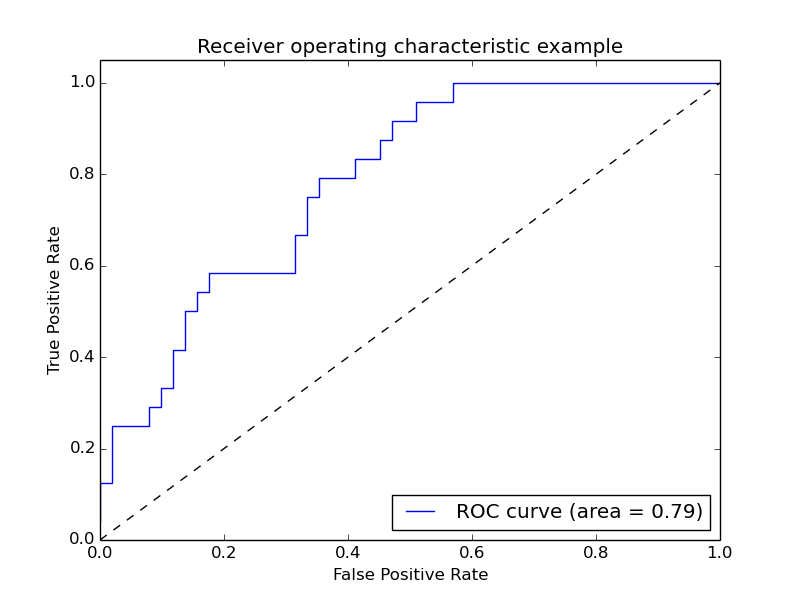
Trong ngữ cảnh phân biệt động đất, F1-score càng cao chứng tỏ mô hình vừa nhạy bén trong việc phát hiện động đất, vừa đáng tin cậy trong dự đoán.

F1-score là tiêu chí cân bằng, đặc biệt hữu ích khi dữ liệu không cân bằng.

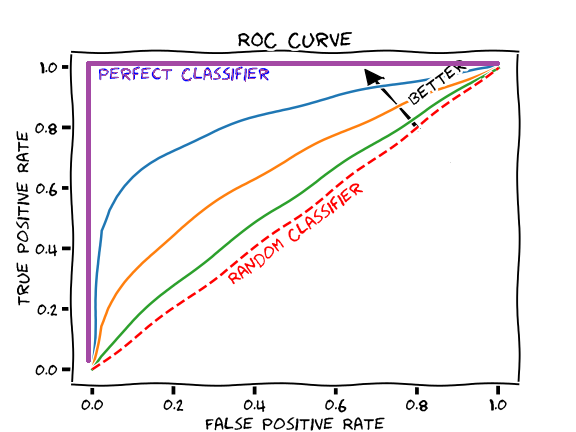
### ****3.5.6. ROC Curve và AUC****

AUC (Area Under the Curve) là một phép đo tổng hợp về hiệu suất của phân loại nhị phân trên tất cả các giá trị ngưỡng có thể có. Để hiểu rõ hơn về metric này, chúng ta sẽ tìm hiểu về một khai niệm cơ sở trước, đó là ROC Curve

ROC Curve (The receiver operating characteristic curve) là một đường cong biểu diễn hiệu suất phân loại của một mô hình phân loại tại các ngưỡng threshold. Về cơ bản, nó hiển thị True Positive Rate (TPR) so với False Positive Rate (FPR) đối với các giá trị ngưỡng khác nhau.



**Hình 3.19. Biểu đồ biểu diễn ROC curve**

 AUC là chỉ số được tính toán dựa trên đường cong ROC nhằm đánh giá khả năng phân loại của mô hình tốt như thê nào. Phần diện tích nằm dưới đường cong ROC và trên trục hoành chính là AUC, có giá trị nằm trong khoảng [0, 1].

**Hình 3.20. Biểu đồ ROC thể hiện hiệu quả của các bộ phân loại**

Khi diện tích này càng lớn, đường cong này sẽ dần tiệm cận với đường thẳng y=1 tương đương với khả năng phân loại của mô hình càng tốt. Còn khi đường cong ROC nằm sát với đường chéo đi qua hai điểm (0, 0) và (1, 1), mô hình sẽ tương đương với một phân loại ngẫu nhiên.

## ****3.6. Công cụ và phần mềm sử dụng****

Đề tài sử dụng các công cụ và phần mềm mã nguồn mở phổ biến trong lĩnh vực học máy và xử lý tín hiệu:

### ****3.6.1. Ngôn ngữ lập trình****

* **Python 3.10+**: dễ sử dụng, thư viện phong phú, phù hợp với xử lý dữ liệu và học máy.

|  |  |
| --- | --- |
| Thư viện | Chức năng |
| NumPy, Pandas | Xử lý dữ liệu và mảng số |
| Matplotlib, Seaborn | Vẽ biểu đồ, trực quan hóa kết quả |
| Scikit-learn | Huấn luyện mô hình Machine Learning: SVM, Random Forest, KNN, XGBoost,... |
| SciPy | Xử lý tín hiệu, tính năng thống kê |
| Joblib, pickle | Lưu và tải mô hình |

### ****3.6.2. Môi trường làm việc****

* **Pycharm**: chỉnh sửa mã nguồn và quản lý dự án.
* **Google Colab**: Để mô hình hóa và trực quan chạy trên cloud, sử dụng GPU miễn phí.

Quá trình huấn luyện bao gồm ba cấu hình mô hình khác nhau được đào tạo trên dữ liệu K-NET, với các mạng trước đã được xác định và sử dụng một số lượng tham số chính. Các bước thử nghiệm huấn luyện đã được triển khai trên Google Colab, một sản phẩm của Google Research, cho phép phát triển và thực thi mã Python trực tiếp thông qua trình duyệt web. Điều này rất thuận tiện khi làm việc với dữ liệu lớn và yêu cầu tài nguyên tính toán mạnh mẽ.

|  |  |  |
| --- | --- | --- |
| **CPU** | **GPU** | **TPU** |
| Intel Xeon Processor with two cores @2.30 GHz and 13GB Ram. | Up to Tesla K80 with 12 GB of GDDR5 VRAM, Intel Xeon Processor with two cores @2.30 GHz and 13GB Ram. | Cloud TPU with 180 teraflops of computation, Intel Xeon Processor with two cores @2.30 GHz and 13GB Ram. |

**Hệ điều hành**

Windows / macOS / Linux đều hỗ trợ tốt

# CHƯƠNG 4: THỰC NGHIỆM VÀ KẾT QUÀ

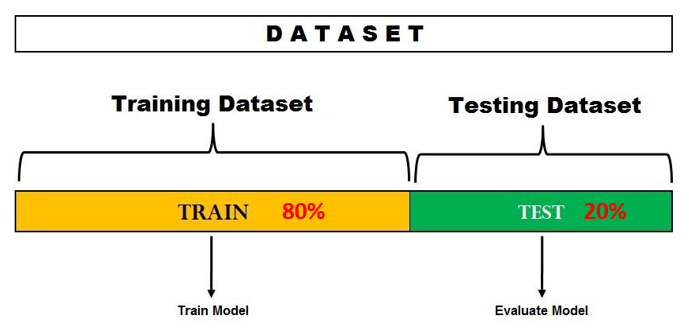
## 4.1. Thiết lập thí nghiệm

Mô tả quy trình thực nghiệm, cách chia dữ liệu và tham số mô hình.

Trong đề tài này, với bộ dữ liệu bao gồm 4237 cơn động đất và 79 dữ liệu về những tín hiệu nhiễu như gia tốc đi bộ, từ động cơ,… Sau khi biến chuyển đổi đơn vị từ giá trị thô (raw) sang đơn vị gia tốc gal thì áp dụng cửa sổ sibling\_window với during\_time bằng 10 giây và overlap bằng 50%. Tạo ra được một bộ dữ liệu mới với kích thước tập động đất (EQ) là (87177, 33) và tập nhiễu (nonEQ) là (9859, 33).

* Kích thước tập EQ: (87177, 33)
* Kích thước tập NonEQ: (9850, 33)

Gộp hai bộ dữ liệu lại với nhau bằng lệnh “**concat**” và gán nhãn tập dữ liệu EQ là 1 và tập NonEQ là 0. Sau đó chia tập dữ liệu với 80% data làm bộ train và 20% làm bộ test.



**Hình 4.1. Phân tách bộ dữ liệu**

Những đặc trưng được đưa sử dụng vào các mô hình học máy là:

* Đặc trưng ZC (Số lần gia tốc đổi dấu)
* Đặc trưng IQR (Khoảng tứ phân vị)
* Đặc trưng Std. (Độ lệch chuẩn)
* Đặc trưng Dominant\_Freq ()
* Đặc trưng RMS (độ mạnh trung bình)

## 4.2. Kết quả huấn luyện và đánh giá mô hình

Nhằm đánh giá khả năng phân biệt giữa dữ liệu EQ và Non-EQ, các mô hình học máy bao gồm Logistic Regression, Random Forest và SVM,… đã được huấn luyện trên tập dữ liệu gia tốc theo ba trục.

Trong phần này trình bày kết quả này, em sẽ đi so sánh hiệu suất các mô hình dựa trên các chỉ số đánh giá chuẩn như F1-score, Accuracy, Precision,... để chọn ra mô hình phù hợp và tối ưu nhất.

Sau khi tiến hành huấn luyện và đánh giá hiệu suất trên cùng một tập dữ liệu gia tốc nhằm phân biệt giữa dữ liệu EQ và Non-EQ, kết quả thu được từ các mô hình như Decision Tree, Random Forest, SVM, XGBoost, LightGBM, ANN và CRNN cho thấy sự khác biệt rõ rệt về khả năng học đặc trưng, độ chính xác và tính tổng quát hóa.Trên đây là những đánh giá tổng quan mang tính định tính trên các mô hình đem sử dụng:

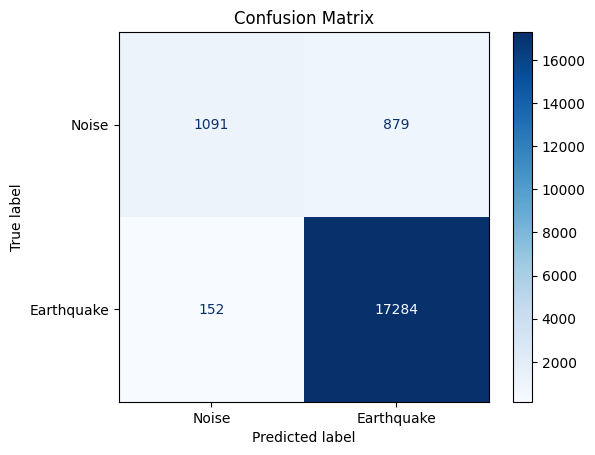
* Decision Tree là mô hình cơ bản, dễ hiểu và trực quan, tuy nhiên thường dễ bị overfitting trên dữ liệu nhiễu. Trong bài toán này, Decision Tree cho kết quả tạm ổn nhưng không vượt trội.
* Random Forest, với bản chất là tập hợp nhiều cây quyết định và áp dụng bagging, cho kết quả ổn định và tốt hơn rõ rệt so với mô hình đơn lẻ, nhờ khả năng giảm phương sai và tránh overfitting.
* XGBoost là thuật toán boosting tiên tiến, cho hiệu suất cao nhất trong nhóm mô hình truyền thống. XGBoost học tốt hơn trong các dữ liệu nhiều nhiễu. Cả hai đều có khả năng học đặc trưng phức tạp, mang lại F1-score và độ chính xác cao, rất phù hợp cho bài toán phân biệt tín hiệu rung động EQ/Non-EQ.
* Support Vector Machines (SVM) đạt kết quả tốt khi dữ liệu có ranh giới phân lớp rõ ràng, đặc biệt hiệu quả nếu không gian đầu vào đã được chuẩn hóa tốt. Tuy nhiên, hiệu suất có thể bị ảnh hưởng nếu dữ liệu nhiều nhiễu hoặc không tuyến tính rõ ràng. SVM hoạt động khá tốt trong bài toán này, nhưng có thể tốn thời gian tính toán hơn nếu kích thước dữ liệu lớn.
* Artificial Neural Network (ANN) thể hiện khả năng học đặc trưng phi tuyến từ dữ liệu thô, tuy nhiên hiệu suất còn phụ thuộc vào thiết kế mạng và dữ liệu đầu vào đã được xử lý đặc trưng hay chưa. ANN hoạt động tốt nhưng chưa vượt qua được XGBoost hay LightGBM nếu không có tối ưu cấu trúc mạng.
* CRNN (Convolutional Recurrent Neural Network) – kết hợp giữa CNN và RNN – cho thấy kết quả vượt trội, đặc biệt khi dữ liệu là dạng chuỗi thời gian như tín hiệu gia tốc. CNN giúp trích xuất đặc trưng không gian (pattern rung động), RNN (hoặc LSTM) học được tính thời gian của tín hiệu. Đây là mô hình có hiệu quả cao nhất về F1-score và recall trong số tất cả, nhưng cũng đòi hỏi nhiều tài nguyên tính toán hơn.

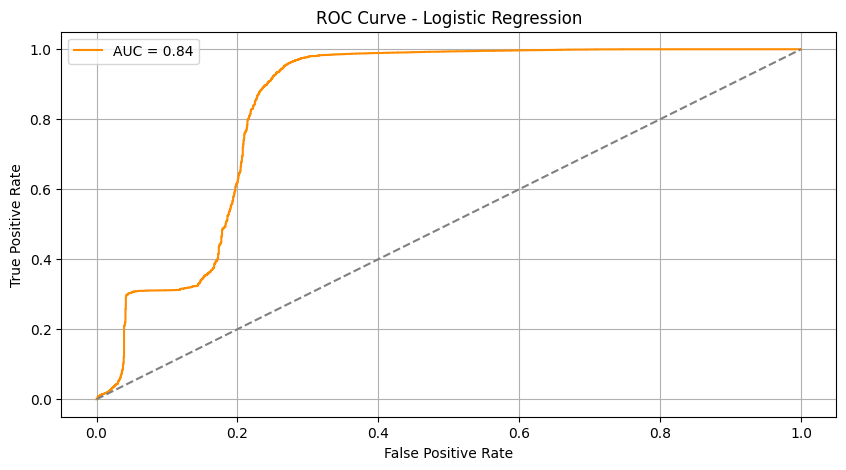
Sau đây là bảng thống kê định lượng các metrics để đánh giá kết quả mô hình huấn luyện.

**Bảng 4.1. Bảng thống kê định lượng các metrics để đánh giá kết quả mô hình huấn luyện**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Accuracy** | **Precision** | **Recall** | **F1** |
| **Decision Tree** | 0.9279 | 0.85 | 0.7 | 0.75 |
| **KNN** | 0.91 | 0.98 | 0.97 | 0.98 |
| **Naive Bayes** | 0.923 | 0.79 | 0.79 | 0.79 |
| **Logistic Regression** | 0.9468 | 0.91 | 0.77 | 0.83 |
| **Random Forest** | **0.9967** | **0.99** | **0.99** | **0.99** |
| **SVM** | 0.9883 | 0.97 | 0.96 | 0.96 |
| **XGBoost** | 0.9928 | 0.98 | 0.98 | 0.98 |
| **ANN** | 0.976 | 0.8935 | 0.8646 | 0.975 |
| **CRNN** | 0.9989 | 1.0 | 0.99 | 0.99 |

*Kết quả huấn luyện của các mô hình:*

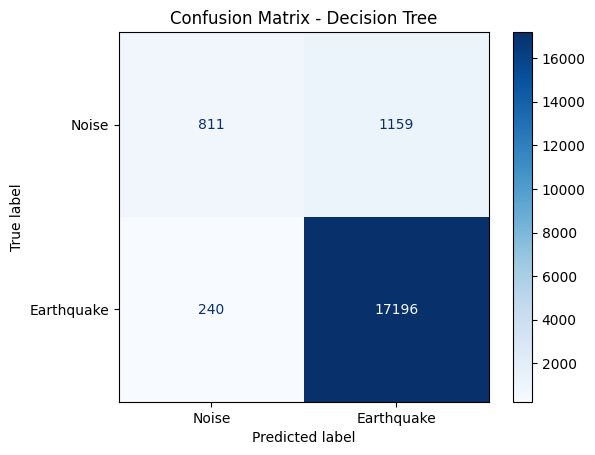
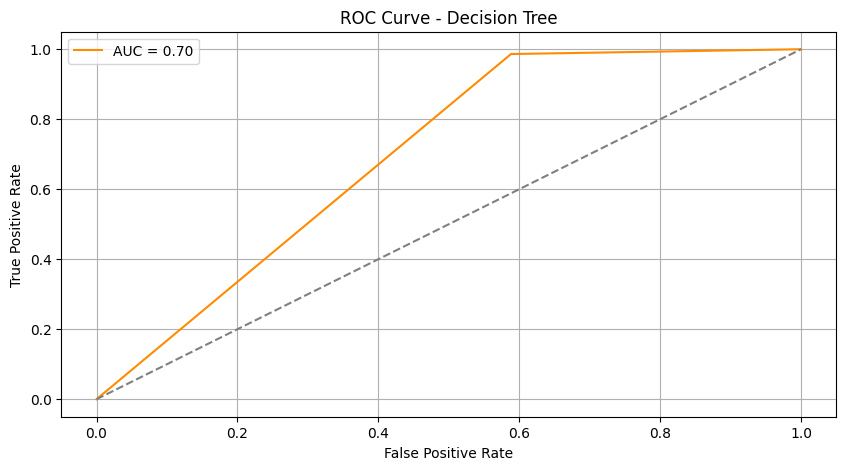
**Mô hình Logistic Regression:**



**Hình 4.2. Mô hình Logistic Regression**

Nhận xét:

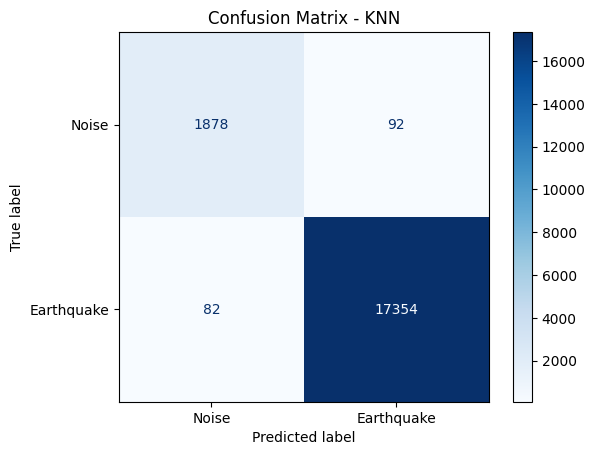
Mô hình Logistic Regression đạt độ chính xác cao (Accuracy 94.68%) và độ chính xác khi dự đoán đúng (Precision 91%), tuy nhiên Recall chỉ đạt 77%, cho thấy mô hình còn bỏ sót nhiều trường hợp động đất. Do đó, mặc dù đơn giản và dễ triển khai, mô hình này chưa thực sự phù hợp với yêu cầu cảnh báo sớm cần độ nhạy cao.

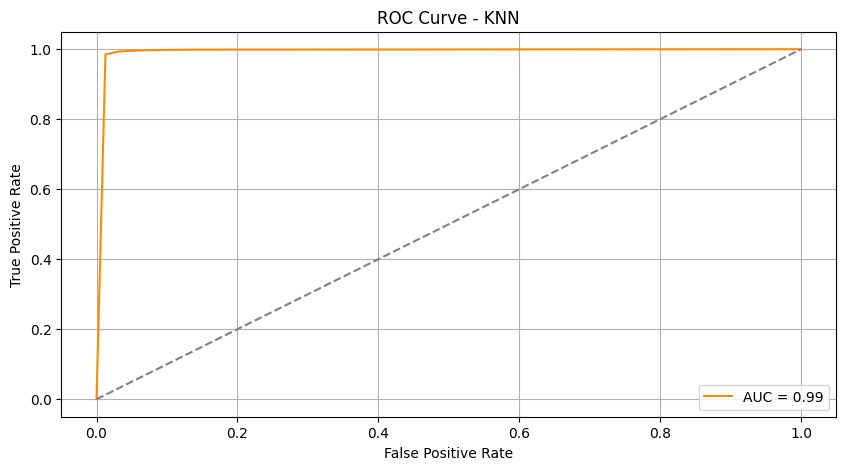
**Mô hình Decision Tree:**

**Hình 4.3. Mô hình Decision Tree**

Nhận xét:

Mô hình Decision Tree có độ chính xác (Accuracy) là 92.79%, nhưng Recall chỉ đạt 70%, thấp nhất trong số các mô hình, cho thấy khả năng phát hiện đúng các trận động đất còn hạn chế. F1-score đạt 0.75, phản ánh hiệu suất tổng thể chưa cao.

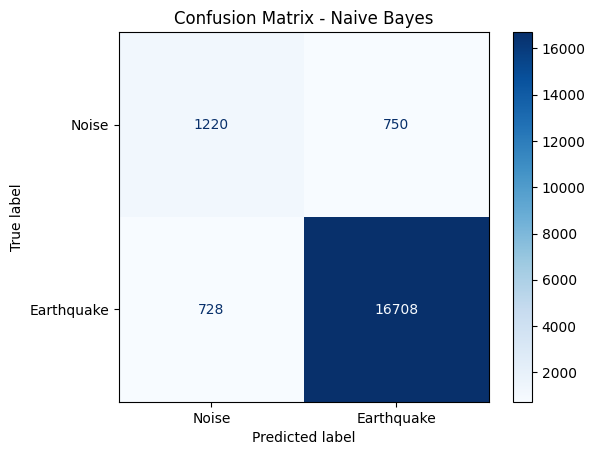
**Mô hình KNN:**

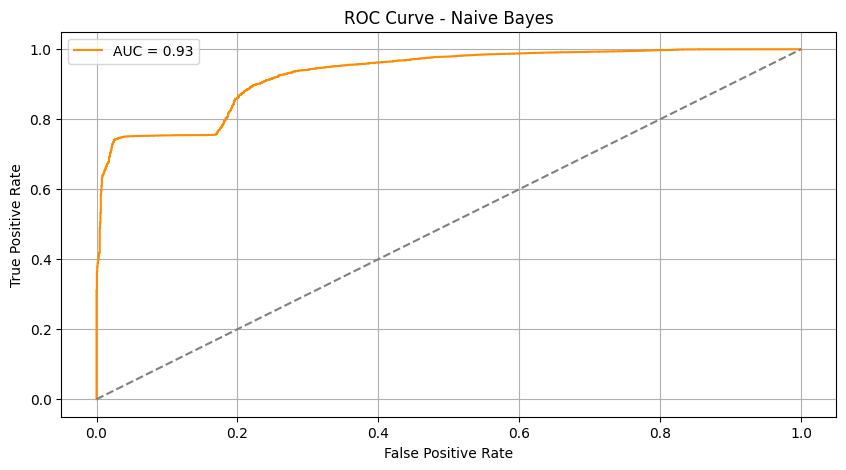


**Hình 4.4. Mô hình KNN**

Nhận xét:

Mô hình KNN đạt hiệu suất khá tốt với Accuracy 91%, Precision 98% và Recall 97%, cho thấy mô hình dự đoán rất chính xác và ít bỏ sót các trận động đất. F1-score đạt 0.98, phản ánh sự cân bằng tốt giữa Precision và Recall. Tuy nhiên, KNN có thể gặp khó khăn về tốc độ khi xử lý dữ liệu lớn.

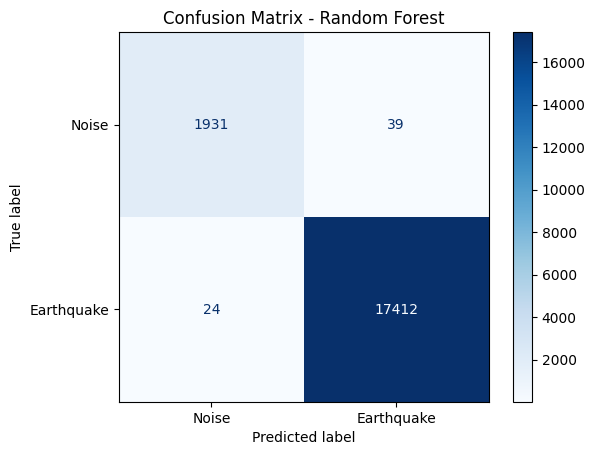
**Mô hình Naïve Bayes:**

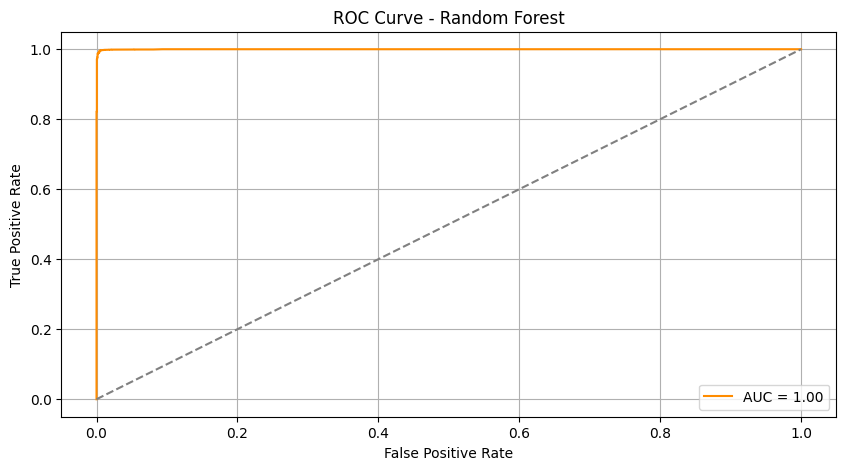


**Hình 4.5. Mô hình Naïve Bayes**

Nhận xét:

Mô hình Naive Bayes đạt Accuracy 92.3%, nhưng Precision và Recall đều chỉ đạt 79%, cho thấy hiệu suất phân loại ở mức trung bình. F1-score cũng là 0.79, phản ánh mô hình không quá nổi bật trong việc phân biệt giữa động đất và nhiễu.

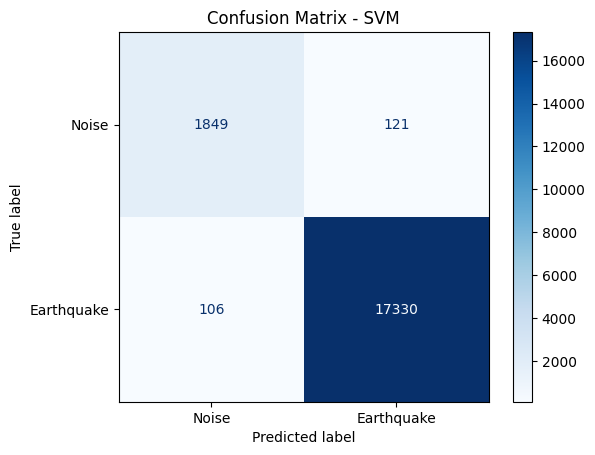
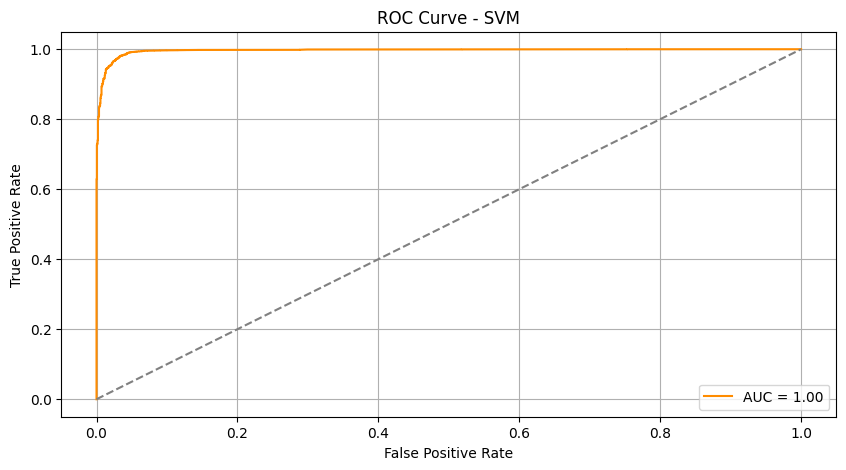
**Mô hình Random Forest:**



**Hình 4.6. Mô hình Random Forest**

Nhận xét:

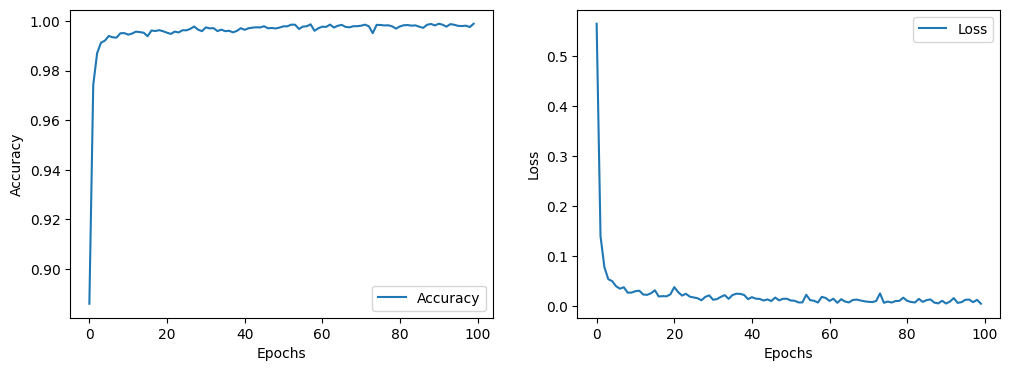
Mô hình Random Forest đạt hiệu suất rất cao với Accuracy 99.67%, Precision và Recall đều đạt 99%, cho thấy khả năng phân loại gần như tuyệt đối và rất ít bỏ sót động đất. F1-score cũng đạt 0.99, phản ánh sự ổn định và chính xác toàn diện của mô hình. Với khả năng xử lý tốt dữ liệu phi tuyến và chống overfitting.

**Mô hình Support Vector Machine**

**Hình 4.7. Mô hình Support Vector Machine**

Nhận xét:

**Mô hình CRNN**



**Hình 4.8. Mô hình CRNN**

## 4.3. Dự đoán độ lớn của trận động đất

Sau khi phân biệt được liệu dữ liệu đưa vào đó có phải là dữ liệu của động đất không thì em có thêm một chức năng nữa đó là xác định được độ lớn của trận EQ (Magnitude). Để xác định độ lớn của trận động đất (Magnitude), bạn có thể sử dụng phương pháp tính toán từ dữ liệu gia tốc thu được từ cảm biến MPU6050. Một trong những phương pháp phổ biến là sử dụng công thức:

Trong đó:

M là độ lớn của trận động đất

A là biên độ sóng động đất đo được từ cảm biến.

A0 là biên độ chuẩn. giá trị phổ biến cho A0 là 0.0001 cm, đây là giá trị chuẩn được sử dụng trong nhiều hệ thống đo động đất.

Tuy nhiên, trong bài toán này em đã thử nghiệm và áp dụng thêm bài toán phán đoán độ lớn của trận động đất sử dụng các mô hình Regression. Dưới đây là kết quả thống kê các mô hình đã áp dụng.

Để đánh giá mô hình Regression hoạt động tốt hay không thì người ta sử dụng những metrics có khả năng làm việc với các giá trị liên tục như MSE, MAE hay R2.

**MAE: Mean Absolute Error** là một metric đánh giá mô hình bằng cách tính trung bình các giá trị tuyệt đối sai số giữa giá trị thực tế và giá tị dự đoán. Công thức để tính MAE đuợc tính như sau:

**MSE (Mean Square Error)** là một metric phổ biến nhất trong các bài toán hồi quy. Về cơ bản, nó tính trung bình của bình phương sai số giữa giá trị thực tế và giá trị dự đoán. Công thức để tính MSE đuợc tính như sau:

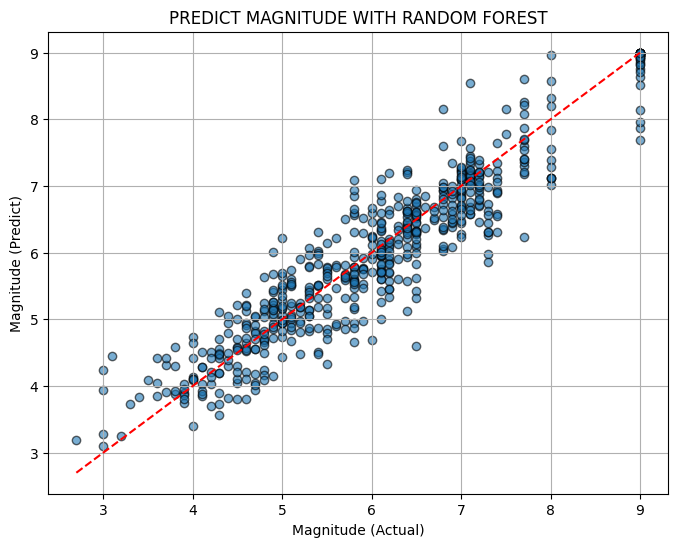
**R2( R-squared)** là một metric dùng trong bài toán hồi quy. R² đo lường mức độ mà mô hình giải thích được phương sai của dữ liệu thực tế. Công thức:

**Bảng 4.2. Thống kê kết quả các mô hình phán đoán độ lớn trận động đất**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mô hình** | **MAE** | **MSE** | **R2** |
| **Logistic Regression** | 0.7816 | 0.7955 | 0.6319 |
| **Random Forest** | **0.3577** | **0.477** | **0.8678** |
| **SVM** | 0.7816 | 0.9541 | 0.4705 |

Nhận thấy mô hình Random Forest đưa ra kết quả tốt nhất cho bài toán dự đoán độ lớn của trận động đất khi có metric R2  gần 1.0 nhất.

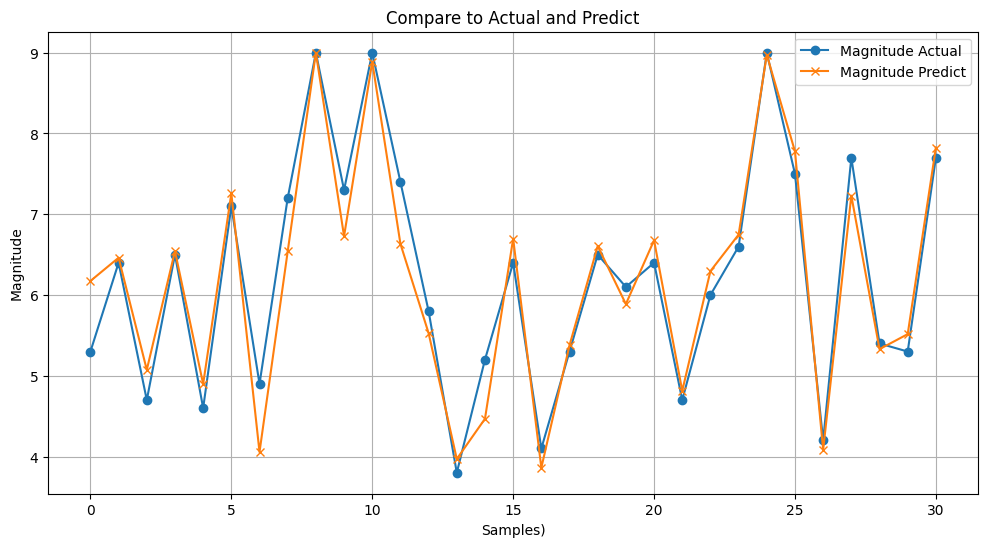
Biểu đồ trên là một scatter plot (biểu đồ phân tán) dùng để so sánh giá trị dự đoán và giá trị thực tế của độ lớn trận động đất (Magnitude) từ mô hình Random Forest.



**Hình 4.9. Biểu đồ phân tán giữa giá trị dự đoán và giá trị thực tế**

**Nhận xét:**

* Nhìn chung, các điểm phân bố khá sát đường đỏ, cho ta thấy mô hình Random Forest của bạn dự đoán khá tốt độ lớn động đất.
* Tuy nhiên vẫn có một số điểm lệch khá xa, nhất là ở vùng độ lớn lớn hơn 7.
* Mô hình chưa học tốt các trận động đất lớn, có thể vì số mẫu ít hoặc bị underfitting.



**Hình 4.10. Biểu đồ miền thể hiện giá trị thực tế và giá trị dự đoán**

## 4.4. Thiết kế giao diện người dùng

Thiết kết giao diện người dùng

Qt Designer là một công cụ mạnh mẽ giúp xây dựng giao diện người dùng đồ họa cho ứng dụng sử dụng framework Qt. Với tính năng kéo và thả thông minh, người dùng có thể dễ dàng bố trí và tùy chỉnh các thành phần như nút, trường văn bản, hộp tổ hợp và nhiều hơn nữa mà không cần phải viết mã từ đầu. Điều này giúp tiết kiệm thời gian và công sức trong quá trình phát triển phần mềm, đặc biệt là với những người không có kinh nghiệm lập trình giao diện người dùng.

A screenshot of a computer

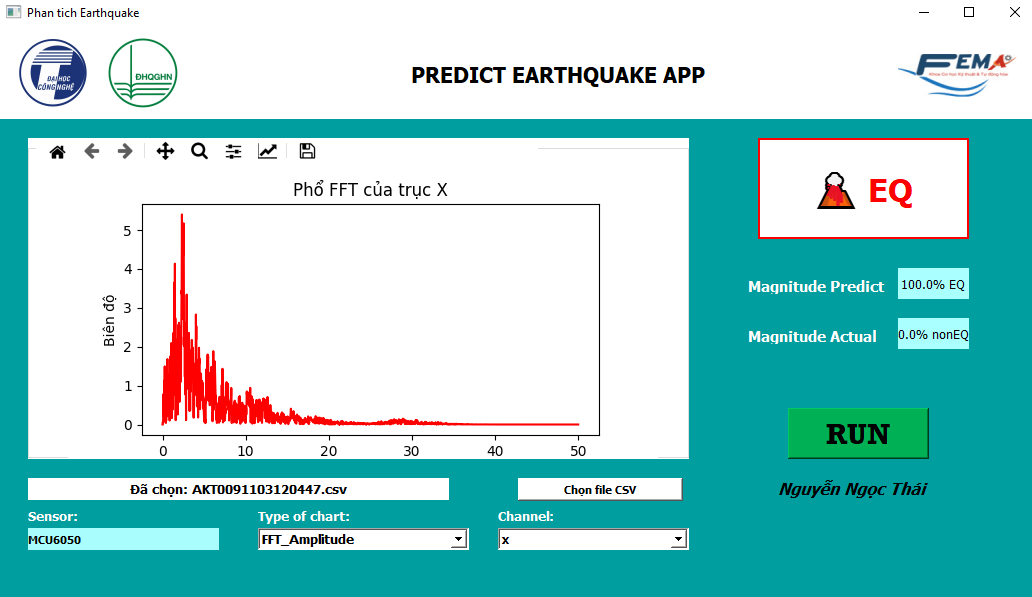
Description automatically generated

**Hình 4.11. Giao diện phầm mềm QT Designer**

Qt Designer tạo ra các tệp có định dạng .ui, đó là định dạng XML đặc biệt để lưu trữ cấu trúc của các widget theo dạng cây. Các tệp này có thể được tải vào trong quá trình chạy của ứng dụng hoặc biên dịch sang mã ngôn ngữ như C++ hoặc Python. Điều này cho phép người dùng dễ dàng sử dụng lại và chỉnh sửa giao diện người dùng của ứng dụng, đồng thời phân tách rõ ràng quá trình thiết kế giao diện và lập trình logic.

Trong phạm vi đồ án này, em đã sử dụng công cụ QT Designer để có thể dễ dàng hơn trong quá trình thiết kế giao diện, đồng thời công cụ này giúp cho quá trình quản lý các dòng lệnh cũng đơn giản và dễ hiểu hơn không chỉ cho người lập trình mà còn cho người dùng. Để dễ dàng sử dụng đối với những người công nhân thì giao diện cần phải dễ hiểu, dễ sử dụng và chỉ hiển thị những thông tin cần thiết. Hiểu được vấn đề đó em đã thiết kế giao diện như hình dưới đây.

Giao diện đã được tích hợp cửa sổ hiện thị quá trình xử lý, nhận diện lỗi trong thời gian thực từ file dữ liệu truyền vào. Đồng thời người dùng cũng có thay đổi kiểu biểu đồ muốn hiện theo từng miền thời gian như Domain-time hay miền tần số FFT.. Ngoài ra giao diện còn hiện ra độ lớn của trận động đất bao gồm giá trị dự đoán và thực tế, để mô hình tiến hành so sánh, đánh giá với giá trị nhận diện được từ đó đưa ra kết quả phân loại và hiển thị ra màn hình. Trong tương lai giao diện người dùng sẽ tích hợp thêm nhiều tính năng đặc biệt là tính năng phân quyền cho người dùng, chỉ người nào được cấp quyền mới có thay đổi được cái giá trị cài đặt để đảm bảo tính bảo mật của mô hình.



**Hình 4.12. Giao diện người dùng hệ thống**

# CHƯƠNG 5: THẢO LUẬN VÀ KẾT LUẬN

Trong khuôn khổ đồ án tốt nghiệp, đề tài "Nghiên cứu bài toán phát hiện động đất sử dụng dữ liệu cảm biến gia tốc" đã tiến hành xây dựng một hệ thống nhận diện tín hiệu địa chấn thông qua việc ứng dụng các mô hình học máy trên dữ liệu cảm biến ba trục. Dữ liệu được xử lý, trích xuất đặc trưng và huấn luyện với nhiều thuật toán khác nhau như Logistic Regression, Random Forest, SVM, XGBoost, ANN và CRNN nhằm mục đích đánh giá khả năng phân loại giữa tín hiệu động đất (EQ) và tín hiệu nhiễu (Noise).

Kết quả thực nghiệm cho thấy các mô hình truyền thống như Random Forest, SVM hay XGBoost đều đạt hiệu năng cao với độ chính xác và F1-score vượt trội. Đặc biệt, mô hình học sâu CRNN thể hiện khả năng học đặc trưng không gian – thời gian hiệu quả nhất, với độ chính xác gần như tuyệt đối và khả năng tổng quát hóa tốt hơn trên tập dữ liệu kiểm thử. Điều này khẳng định rằng với dạng dữ liệu chuỗi thời gian như tín hiệu cảm biến địa chấn, các mô hình có khả năng khai thác đồng thời thông tin cục bộ và thứ tự thời gian sẽ đem lại hiệu quả vượt trội.

Ngoài ra, đồ án cũng đã thiết kế giao diện người dùng giúp minh họa trực quan quá trình nạp dữ liệu và phân loại, mở ra hướng tiếp cận gần hơn với ứng dụng thực tế. Hệ thống này có thể được triển khai như một công cụ hỗ trợ trong các mạng lưới cảnh báo sớm động đất quy mô nhỏ, tiết kiệm chi phí nhưng vẫn đảm bảo độ tin cậy trong việc phát hiện rung động bất thường.

Mặc dù kết quả đạt được là khả quan, đề tài vẫn còn một số hạn chế nhất định. Dữ liệu sử dụng trong huấn luyện và kiểm thử chủ yếu được lấy từ các nguồn công khai với số lượng và phạm vi địa lý còn hạn chế. Việc triển khai hệ thống phát hiện động đất theo thời gian thực vẫn chưa được hiện thực hóa trong phạm vi nghiên cứu này. Hơn nữa, hệ thống hiện tại mới chỉ dừng lại ở bước phân loại tín hiệu, chưa tích hợp chức năng định vị chấn tâm hay ước lượng độ lớn của trận động đất một cách tự động và chính xác.

Trong tương lai, để nâng cao tính ứng dụng và độ tin cậy của hệ thống, đề tài có thể được mở rộng theo các hướng sau. Trước hết là tích hợp hệ thống phát hiện theo thời gian thực với khả năng thu thập và xử lý tín hiệu liên tục từ các cảm biến đặt tại hiện trường, nhằm phục vụ mục tiêu cảnh báo sớm. Thứ hai, mở rộng việc thu thập dữ liệu từ nhiều loại cảm biến khác như địa chấn kế, cảm biến âm thanh hoặc GPS để tăng tính đa chiều và giảm thiểu sai số do nhiễu đơn kênh. Thứ ba, triển khai thí điểm hệ thống tại một số địa phương có nguy cơ động đất hoặc hoạt động xây dựng mạnh để đánh giá hiệu quả trong môi trường thực tế, từ đó điều chỉnh mô hình phù hợp hơn với điều kiện hoạt động thực tế ở Việt Nam.

Tóm lại, nghiên cứu này không chỉ góp phần khẳng định tiềm năng của cảm biến gia tốc trong việc phát hiện sớm động đất mà còn minh chứng cho khả năng ứng dụng của các mô hình học máy trong bài toán phân tích và cảnh báo thiên tai. Đây là một bước khởi đầu quan trọng để hướng tới việc xây dựng một hệ thống giám sát và cảnh báo động đất thông minh, chi phí thấp và có khả năng mở rộng trong tương lai.

# TÀI LIỆU THAM KHẢO

**Tiếng Việt:**

1. baochinhphu.vn. “Tìm Hiểu về Động Đất.” *Baochinhphu.vn*, 22 Apr. 2011, baochinhphu.vn/tim-hieu-ve-dong-dat-10298355.htm. Accessed 18 Apr. 2025.
2. Động đất là gì? Nguyên nhân và hậu quả của động đất. “Động Đất Là Gì? Nguyên Nhân và Hậu Quả Của Động Đất.” *Phuongnam24h.com*, 2023, phuongnam24h.com/dong-dat-la-gi.html. Accessed 18 Apr. 2025.

**Tiếng Anh:**

1. Abbadia, Jessica. “Understanding Earthquake Magnitudes and Their Impact.” *Mind the Graph Blog*, 27 Feb. 2023, mindthegraph.com/blog/earthquake-magnitude/.
2. Michigan Tech. “Where Do Earthquakes Happen? | UPSeis.” *Michigan Technological University*, 2023, www.mtu.edu/geo/community/seismology/learn/earthquake-location/.
3. Panchuk, Karla. “12.2 Seismic Waves and Measuring Earthquakes.” *Opentextbc.ca*, 20 Aug. 2021, opentextbc.ca/physicalgeologyh5p/chapter/seismic-waves-and-measuring-earthquakes/.
4. “What Are Accelerometer Sensors? Explain How They Work, What They Measure, and How They Are Used - Sensing System - Epson.” *Epson.com*, 2024, global.epson.com/products\_and\_drivers/sensing\_system/what\_are\_accelerometers/.