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 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. 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.

Khoá luận tập trung nghiên cứu và xây dựng mô hình phát hiện động đất dựa trên dữ liệu từ cảm biến gia tốc ba trục. 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 quá trình tiền xử lý và quá trình trích xuất đặc trưng từ tín hiệu. Khoá luận 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 để chọn ra mô hình đánh giá tốt nhất.

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, khoá luận 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, Earthquake.*

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 khoá luận 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 .... tháng ...năm 2025.

Sinh viên

Thái

Nguyễn Ngọc Thái

LỜI CAM ĐOAN

Em xin cam đoan khoá luận 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 khoá luận 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 khoá luận , 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 .... tháng ...năm 2025

Người cam đoan

Thái

Nguyễn Ngọc Thái

MỤC LỤC

MỤC LỤC HÌNH ẢNH

MỤC LỤC BẢNG BIỂU

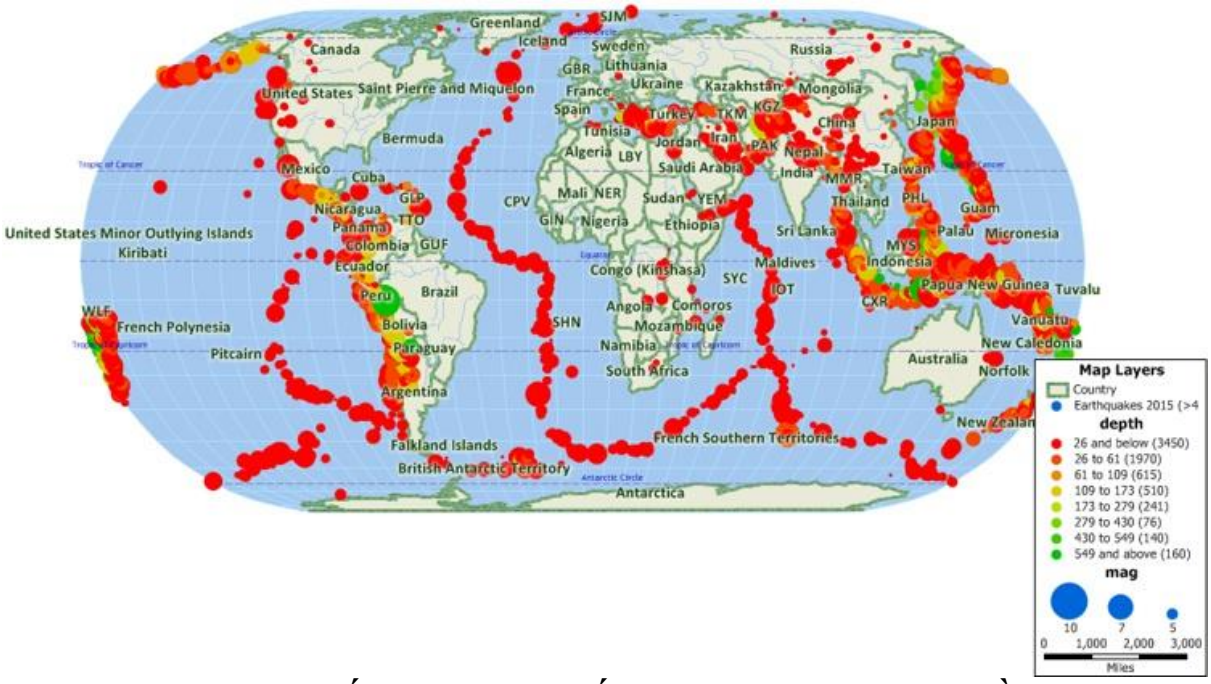
DANH SÁCH CHỮ VIẾT TẮT

STT	Chữ viết tắt	Chữ viết đầy đủ/ Giải thích nghĩa
1	CNN	Convolutional Neural Network: Mạng neural tích chập
2	ML	Meachine Learning
3	DL	Deep Learning
4	RL	Random Forest
5	PSD	Power Spectral Density – mật độ phổ công suất.
6	FFT	Fast Fourier Transform

MỞ ĐẦU

Tính cấp thiết đề tài

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 i.1. 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. 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.

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.

Bảng i.1. Thống kê những trận động đất từ 2014- 2024

M_w	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
8.0–9.9	1	1	0	1	1	1	0	3	0	0	0
7.0–7.9	11	18	16	6	16	9	9	16	11	19	10
6.0–6.9	143	127	131	104	118	135	111	141	117	125	90
5.0–5.9	1.580	1.413	1.550	1.447	1.671	1,484	1.315	2.046	1.603	1.373	1,267
4.0–4.9	15.817	13.777	13.700	10.544	12.782	11.897	12.135	14.643	13.707	12.172	11,200
Tổng cộng	17.552	15.336	15.397	12.102	14.589	13.530	13.572	16.849	15.438	13.689	23,456

(Nguồn: https://vi.wikipedia.org/wiki).

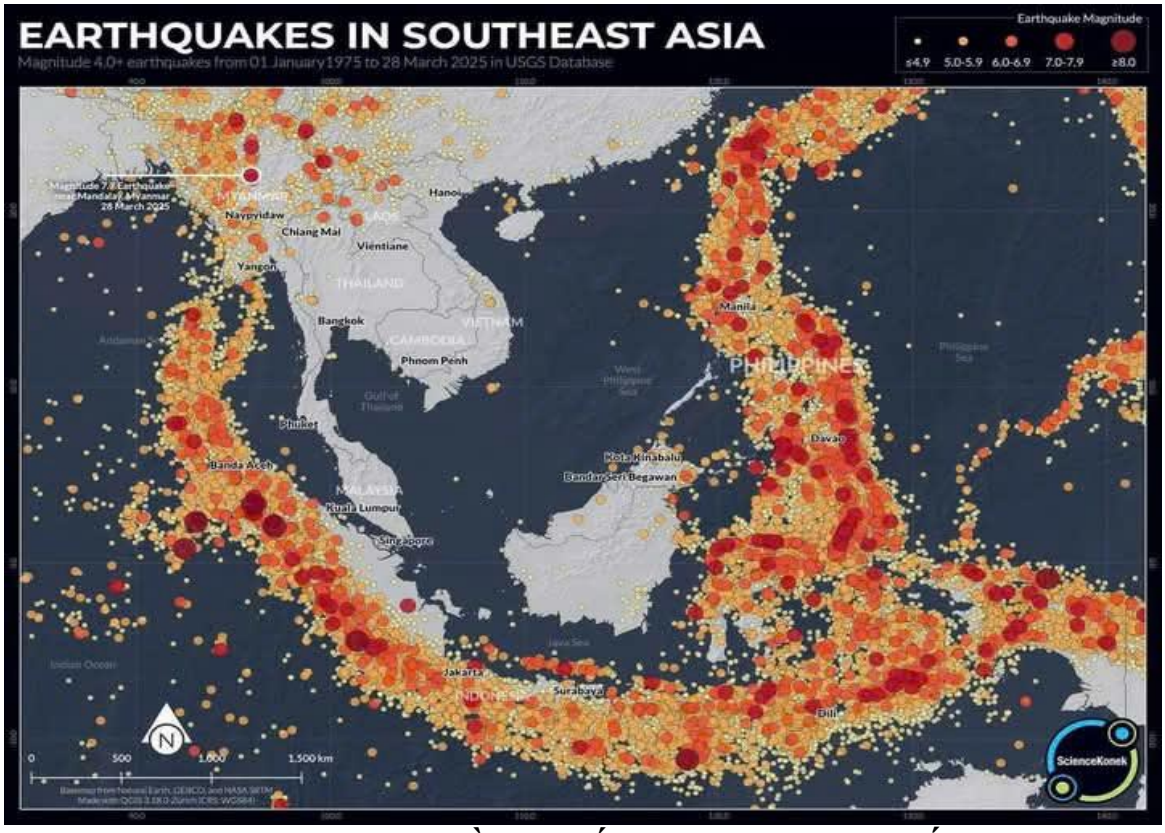
Tình hình Động đất tại Việt Nam

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.

Theo báo điện tử Tiền Phong, đưa tin vào ngày 29/03/2025 về "Những trận động đất lớn nhất ở Việt Nam thế kỷ qua [3] " có thể kể đến như sau:

- Ngày 1/11/1935, 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.
- Gần đây nhất, vào ngày 28/7/2024, liên tục nhiều trận động đất 5 độ đã gây ra rung chấn cho khắp Tây Nguyên và nhiều tỉnh miền Trung.



Hình i.2. Bản đồ động đất khu vực Đông Nam Á.

(Nguồn: Cơ quan Khảo sát Địa chất Hoa Kỳ (USGS)

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.

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.

Ý nghĩa khoa học và thực tiễn

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 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.

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.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.

Đối tượng và phương pháp nghiên cứu

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

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.

Nội dung nghiên cứu

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

1. Thu thập và tiền xử lý dữ liệu: Chuyển đổi đơn vị đo, loại bỏ nhiễu và giá trị ngoại lai, chuẩn hóa dữ liệu và cân bằng dữ liệu để đảm bảo đầu vào phù hợp với các mô hình học máy.
2. Trích xuất đặc trưng: Tính toán các đặc trưng thống kê và phổ tần số như: năng lượng, IQR, ZC, PSD, kurtosis, skew,... từ tín hiệu gia tốc theo từng trục.
3. Xây dựng và huấn luyện mô hình: Sử dụng các thuật toán học máy như Random Forest, SVM, và Logistic Regression. Mô hình được huấn luyện với tập dữ liệu được chia theo kỹ thuật cross-validation để đảm bảo độ chính xác và tính tổng quát.
4. Đánh giá hiệu quả mô hình: Dựa trên các tiêu chí như Accuracy, Precision, Recall, F1-score và AUC-ROC để so sánh hiệu suất phân loại động đất và nhiễu.
5. 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".

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.

Bố cục đề tài

Mở đầu

Chương 1: Cơ sở lí thuyết

Chương 2: Phương pháp nghiên cứu

Chương 3: Thực nghiệm và kết quả

Chương 4: Thảo luận và kết luận

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. 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.

1.1.1. 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).

1.1.2. 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.

1.1.3. 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:

$$M = \log_{10} A + C \qquad (1.1.3)$$

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

Bảng 1.1. Mức độ của trận động đất theo thang Richter

Độ lớn (M)	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.

(*Nguồn: Thư viện Pháp luật*) [1]

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 1.2 .Thang đo cường độ động đất.

Cường 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.

(*Nguồn: Thư viện Pháp luật*) [1]

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.

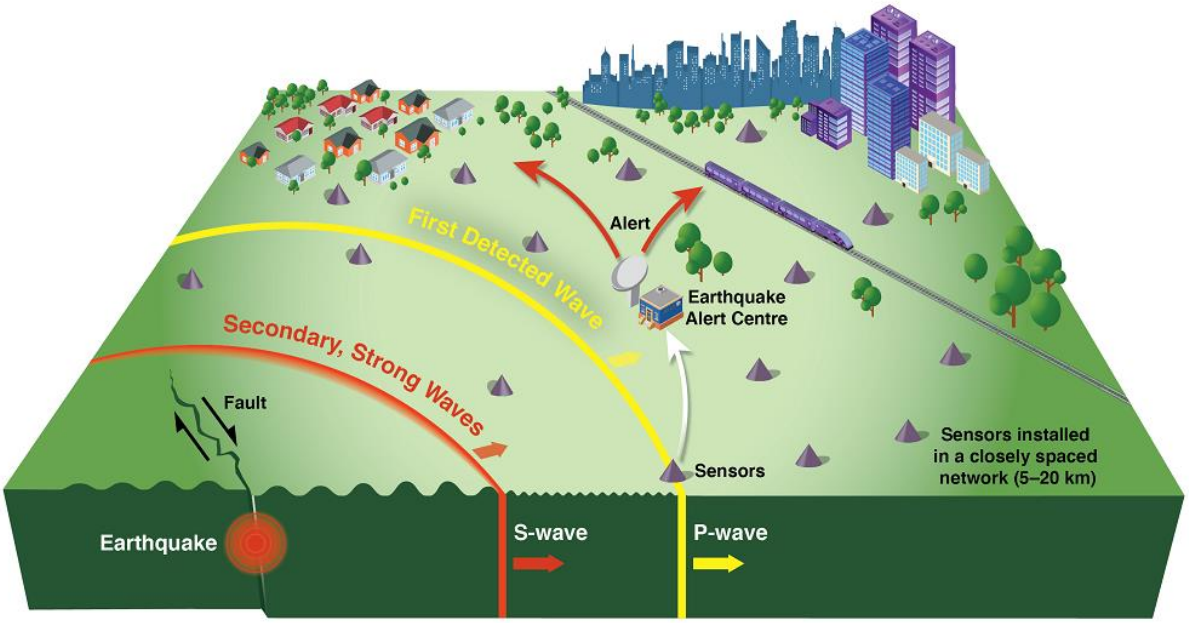
1.1.4. Các loại sóng địa chấn trong quá trình động đất xảy ra

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.

Sóng P còn được gọi là sóng nén hoặc sóng chính. Nó di chuyển nhanh hơn sóng S. Do đó, sóng P sẽ được ghi lại đầu tiên tại máy ghi địa chấn. Chuyển động của sóng P tương tự như sóng âm. Nó đẩy và kéo các tảng đá mà nó đi qua, theo hướng di chuyển. Sóng P có thể di chuyển qua chất rắn cũng như chất lỏng. Sóng S là sóng thứ cấp hoặc sóng ngang. Sóng S di chuyển chậm hơn sóng P. Sóng S chỉ di chuyển qua chất rắn. Sóng S di chuyển các hạt của trái đất theo chuyển động lên xuống. Sóng S có tính phá

hủy nhiều hơn so với sóng P. Khi các sóng địa chấn di chuyển qua thể tích của trái đất ra xa tâm chấn của trận động đất, khoảng cách giữa sóng P và sóng S tăng lên. Thời gian khác biệt này giữa sóng P và sóng S sẽ là thời gian cảnh báo có thể được đưa ra cho con người.



Hình 1.1. Hệ thống cảnh báo sớm động đất dựa trên sóng P và sóng S

1.2. Các mô hình học máy áp dụng

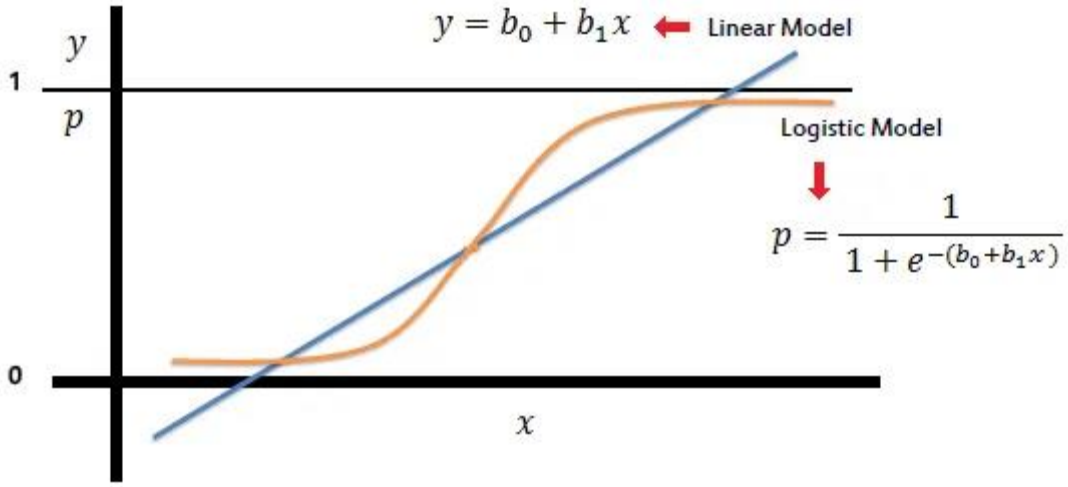
Tổng quan về các thuật toán học máy (như SVM, Random Forest, Neural Networks, …) ứng dụng vào bài toán.

1.2.1. 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). 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

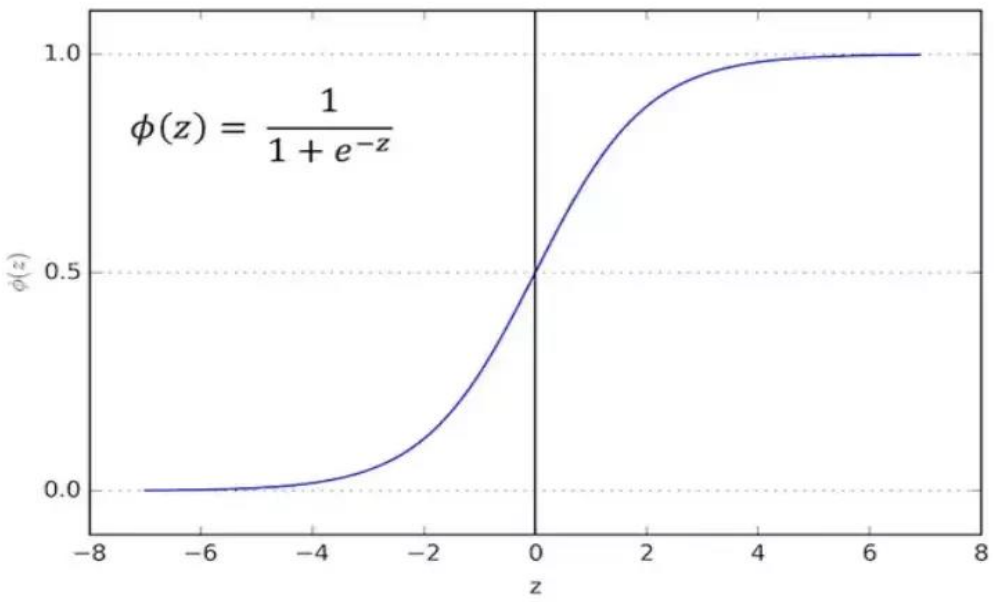
những điểm mạnh của Logistic Regression là nó có thể được hiểu và giải thích một cách tương đối đơn giản, và có thể áp dụng cho các tập dữ liệu lớn mà không gặp nhiều vấn đề về hiệu suất. Tuy nhiên, nó có thể không phù hợp cho các bài toán phức tạp hơn hoặc khi mối quan hệ giữa các biến đầu vào và kết quả không phải là tuyến tính



Hình 1.2. Logistic Regression

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

$$S(z) = \frac{1}{1+e^{-z}} (1.2.1.1)$$



Hình 1.3. 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:

$$L(w) = \frac{-1}{n}\sum_{i=1}^{n}[yi \times \log(pi) + (i - yi) \times \log(1 - pi)] \qquad (1.2.1.2)$$

Trong đó:

- n: số lượng mẫu dữ liệu trong tập huấn luyện.
- y_i: giá trị thực tế của đầu ra thứ i.
- p_i : xác suất dự đoán thuộc lớp 1 của mô hình cho đầu vào thứ i.

Khi mô hình dự đoán chính xác, tức là nếu $y_i = 1$ thì p_i càng gần 1, và nếu $y_i = 0$ thì p_i 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.

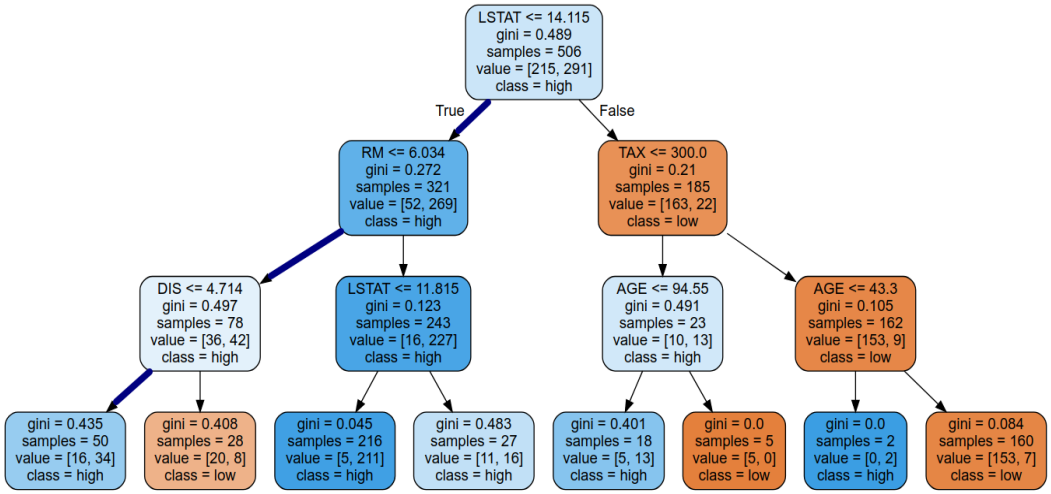
Bảng 1.2.1: Ưu và nhược điểm của mô hình Logistic Regression.

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

1.2.2. Mô hình học máy Decision Tree

Decision Tree là một mô hình học máy thuộc nhóm supervised learning, được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression). Mô hình hoạt động bằng cách mô phỏng quá trình ra quyết định của con người: từ câu hỏi lớn ban đầu, dần chia nhỏ thành các quyết định nhánh cụ thể hơn, đến khi đưa ra kết luận cuối cùng tại "lá" của cây.

Cấu trúc của Decision Tree bao gồm ba thành phần chính đó là nút gốc, nút nhánh và nút lá.



Hình 1.4. Mô hình Decision Tree.

Mô hình cây quyết định sử dụng các chỉ số như Gini impurity, Entropy hoặc Mean Squared Error (cho bài toán hồi quy) để lựa chọn cách chia tách dữ liệu tại mỗi nút, nhằm tối ưu khả năng phân biệt các lớp hoặc giảm sai số dự đoán.

Bảng 1.2.2: Ưu và nhược điểm của mô hình Decision Tree.

Ưu điểm	Nhược điểm
Dễ hiểu, dễ cài đặt. Không cần chuẩn hoá dữ liệu . Có thể xử lý dữ liệu cả dạng số và dạng phân loại.	Dễ bị overfitting với dữ liệu huấn luyện.

1.2.3. Mô hình Random Forest

Random Forest là một mô hình học máy thuộc nhóm ensemble learning, được phát triển để cải thiện hiệu suất và độ chính xác của Decision Tree. Thay vì sử dụng một cây đơn lẻ, Random Forest xây dựng một tập hợp nhiều cây quyết định và đưa ra dự đoán dựa trên bình chọn (đa số) hoặc trung bình kết quả từ các cây đó.

Nguyên lý hoạt động:

Từ tập dữ liệu gốc, Random Forest tạo ra nhiều tập dữ liệu con thông qua bootstrap sampling. Với mỗi cây, chỉ một tập con ngẫu nhiên của các đặc trưng (features) được sử dụng để chia tách tại mỗi nút, điều này giúp tăng tính ngẫu nhiên và giảm độ tương quan giữa các cây.

Kết quả đầu ra:

- Đối với bài toán phân loại: Mô hình sẽ chọn lớp được đa số cây dự đoán.
- Đối với bài toán hồi quy: tính trung bình dự đoán từ các cây.

Bảng 1.2.3: Ưu và nhược điểm mô hình Random Forest.

Ưu điểm	Nhược điểm
Khắc phục được hiện tượng overfitting gặp ở Decision Tree. Có độ chính xác cao, đặc biệt khi dữ liệu có nhiễu. Trích xuất được feature importance.	Kém trực quan hơn so với cây đơn. Có thể tốn tài nguyên tính toán.

1.2.4. Mô hình Support Vector Machines

Support Vector Machine là một mô hình học máy mạnh mẽ, thường được sử dụng cho các bài toán phân loại và đôi khi là hồi quy. Mục tiêu chính của SVM là tìm ra siêu phẳng tốt nhất để phân tách các lớp dữ liệu trong không gian đặc trưng.

Trong không gian hai chiều, SVM tìm một đường thẳng sao cho khoảng cách từ đường đó đến các điểm gần nhất của hai lớp là lớn nhất – đây gọi là biên tối đa (maximum margin). Những điểm dữ liệu nằm gần siêu phẳng nhất và quyết định vị trí của nó được gọi là support vectors.

Với các bài toán không thể phân tách tuyến tính, SVM sử dụng hàm kernel để ánh xạ dữ liệu vào không gian cao hơn, nơi mà việc phân tách tuyến tính có thể thực hiện được.

Các loại kernel thường dùng:

- Linear kernel: dữ liệu phân tách tuyến tính.
- Polynomial kernel: cho biên quyết định cong.
- RBF (Radial Basis Function) kernel: phù hợp với phân lớp phi tuyến mạnh.

Bảng 1.2.4: Ưu và nhược điểm của mô hình SVM.

Ưu điểm	Nhược điểm
Hiệu quả cao với dữ liệu có kích thước đặc trưng lớn. Tốt với các bài toán phân lớp biên rõ ràng. Hạn chế overfitting tốt.	Hiệu quả giảm khi dữ liệu có nhiều nhiễu hoặc không phân tách rõ ràng. Chậm với dữ liệu lớn do tính toán kernel phức tạp. Khó hiểu và khó triển khai cho người mới bắt đầu.

1.3. 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:

Lei Chen1, Christopher Kadlec (2018) – "Improving earthquake prediction accuracy in Los Angeles with machine learning"

Phương pháp: Áp dụng những mô hình học máy 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.

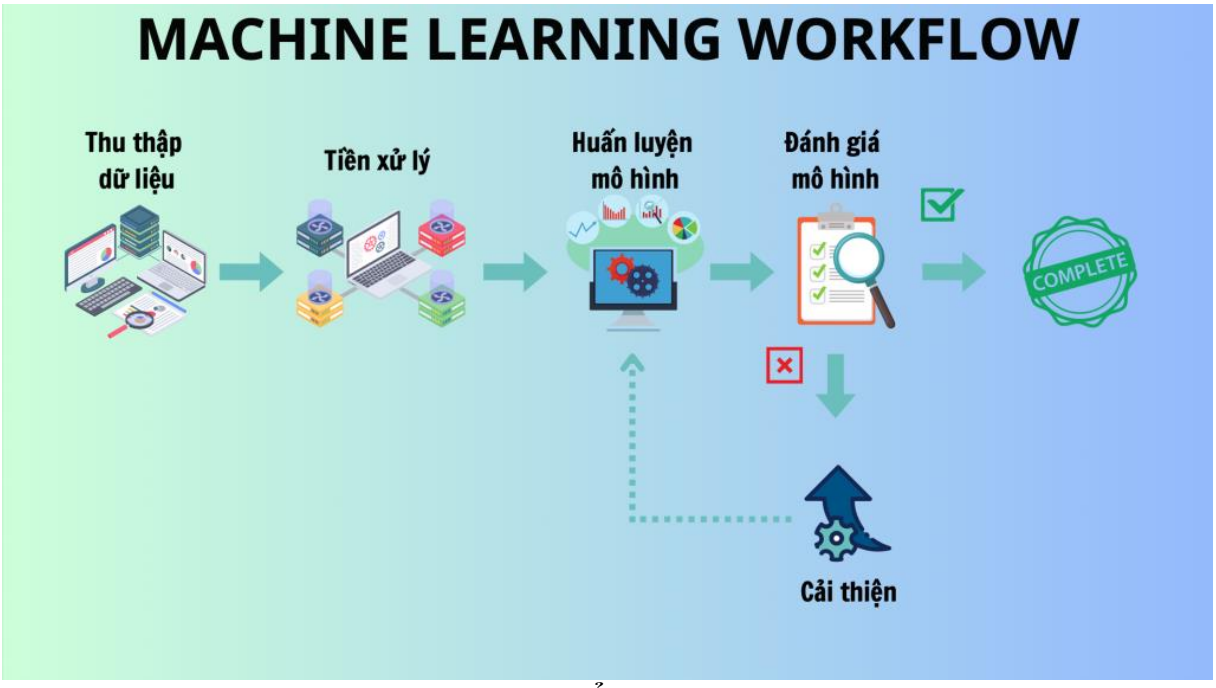
Nhược điểm chung:

- 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 2: PHƯƠNG PHÁP NGHIÊN CỨU

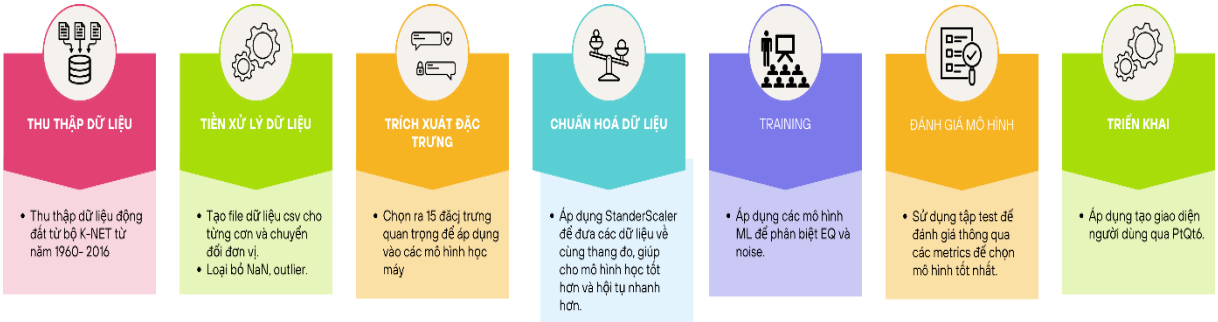
2.1. Quy trình nghiên cứu tổng thể

Dưới đây là quy trình để xây dựng một mô hình học máy ML.



Hình 2.1. Các bước để xây dựng mô hình học máy.

Trong dự án này, em đã áp dụng một mô hình học máy nhằm phân loại tín hiệu cảm biến gia tốc thành hai nhóm: tín hiệu động đất (EQ) và nhiễu (Non-EQ) và dưới đây là những bước mà em đã triển khai.



Hình 2.2. Quá trình triển khai dự án

Quá trình 1: Thu thập dữ liệu.

Dữ liệu được thu thập từ hệ thống K-NET (Kyoshin Network – Nhật Bản), bao gồm các file tín hiệu gia tốc ba trục (X, Y, Z) với tần suất lấy mẫu 100 Hz. Dữ liệu được chia thành hai nhóm.

Quá trình 2: Tiền xử lý dữ liệu.

Sau khi thu thập, dữ liệu được tiền xử lý thông qua các bước loại bỏ các dòng chứa giá trị thiếu (NaN), cắt thành các đoạn tín hiệu ngắn theo cửa sổ sibling_window với thời gian cố định để kiểm tra trực tiếp sự khác biệt giữa EQ và Noise.

Quá trình 3: Trích xuất đặc trưng.

Thực hiện trích xuất đặc trưng từ từng đoạn tín hiệu. Các đặc trưng bao gồm các chỉ số thống kê như giá trị trung bình, độ lệch chuẩn, max, min, độ lệch và độ nhọn, IQR, ZC,… tính riêng trên từng trục x, y và z.

Quá trình 4: Chuẩn hoá dữ liệu.

Thực hiện áp dụng phương pháp Z-score (StandardScaler) giúp đưa những dữ liệu về cùng thang đo, giúp cho mô hình học tốt hơn và hội tụ nhanh hơn.

Quá trình 5: Áp dụng mô hình học máy

Áp dụng những mô hình học máy ML có sẵn trong thư viện sklearn như Random Forest, SVM, Logistic Regression,… để phân biệt giữa tín hiệu EQ và noise.

Quá trình 6: Đánh giá mô hình học máy.

Sau bước huấn luyện mô hình, chúng tôi thử nghiệm với nhiều thuật toán khác nhau. Qua quá trình đánh giá dựa trên những metrics như accuracy, F1-score và chỉ số AUC của đường cong ROC để chọn ra mô hình phân loại tốt nhất.

Quá trình 7: Triển khai mô hình.

Cuối cùng, mô hình được tích hợp vào một giao diện người dùng đơn giản được xây dựng bằng PyQt6. Giao diện này cho phép người dùng tải lên một file tín hiệu bất kỳ, thực hiện dự đoán và hiển thị kết quả phân loại là động đất hay nhiễu.

2.2. Thu thập và mô tả dữ liệu

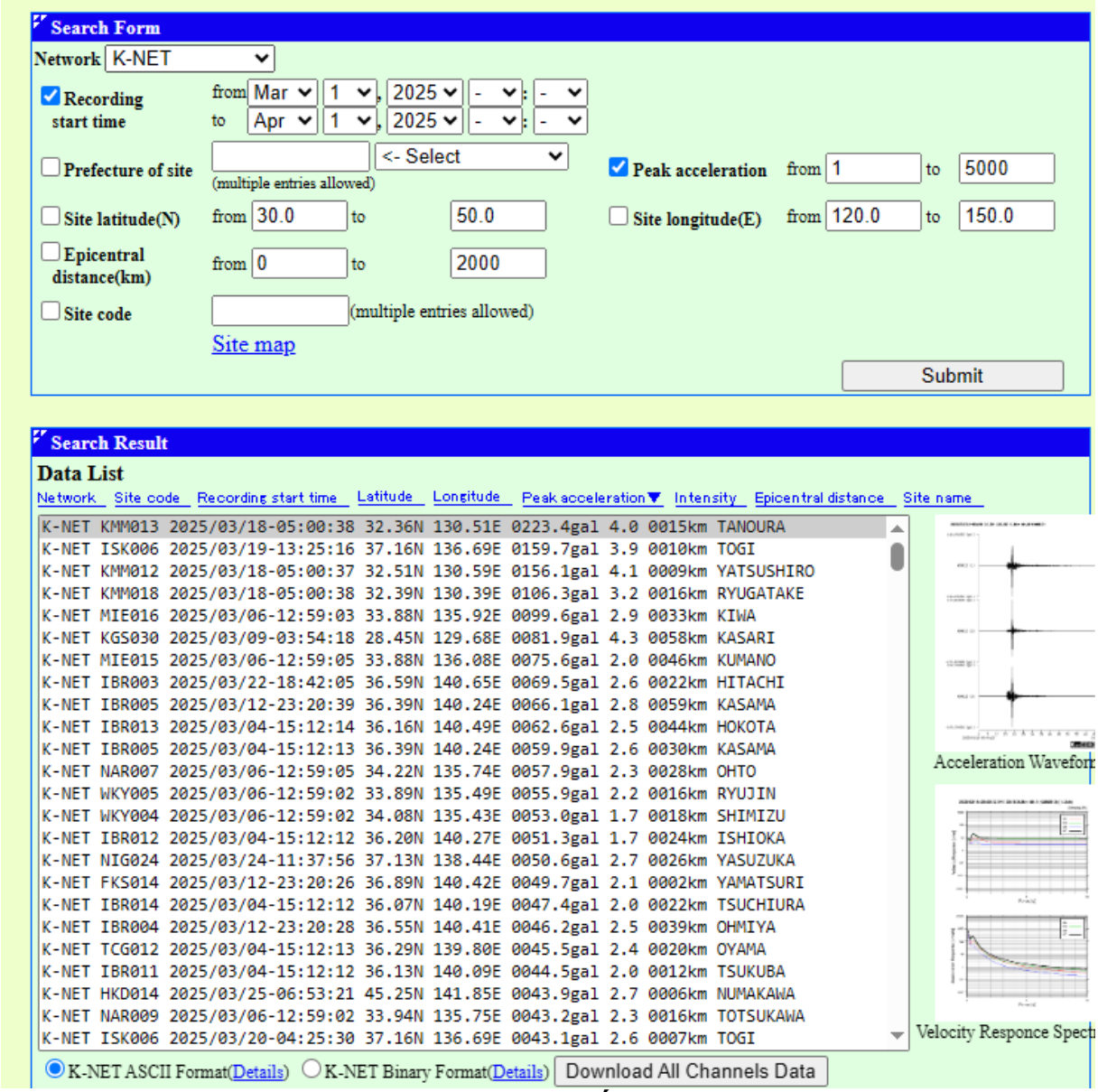
2.2.1. Nguồn dữ liệu

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** [10] - 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 2.3. Dữ liệu động đất từ trang chủ K-NET.

(Nguồn: https://www.kyoshin.bosai.go.jp/kyoshin/data/)[10]

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/s^2.

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.

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.

2.2.2. Mô tả dữ liệu

Sau khi tổng hợp thông tin các dữ liệu theo ba hướng tạo thành một file dữ liệu csv bao gồm có các cột x, y, z và timestamp. Dữ liệu bao gồm khoảng **2320** dữ liệu động đất và **80** dữ liệu nhiễu từ các hoạt động bên ngoài như động cơ, đi bộ, xe bus,…

2.2.2.1. Trực quan theo miền thời gian gia tốc (Domain Time)

Miền thời gian (Time Domain) là quá trình biểu diễn tín hiệu theo thời gian, trục hoành là thời gian, trục tung là gia tốc (thường là m/s^2 hoặc gal). 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…).

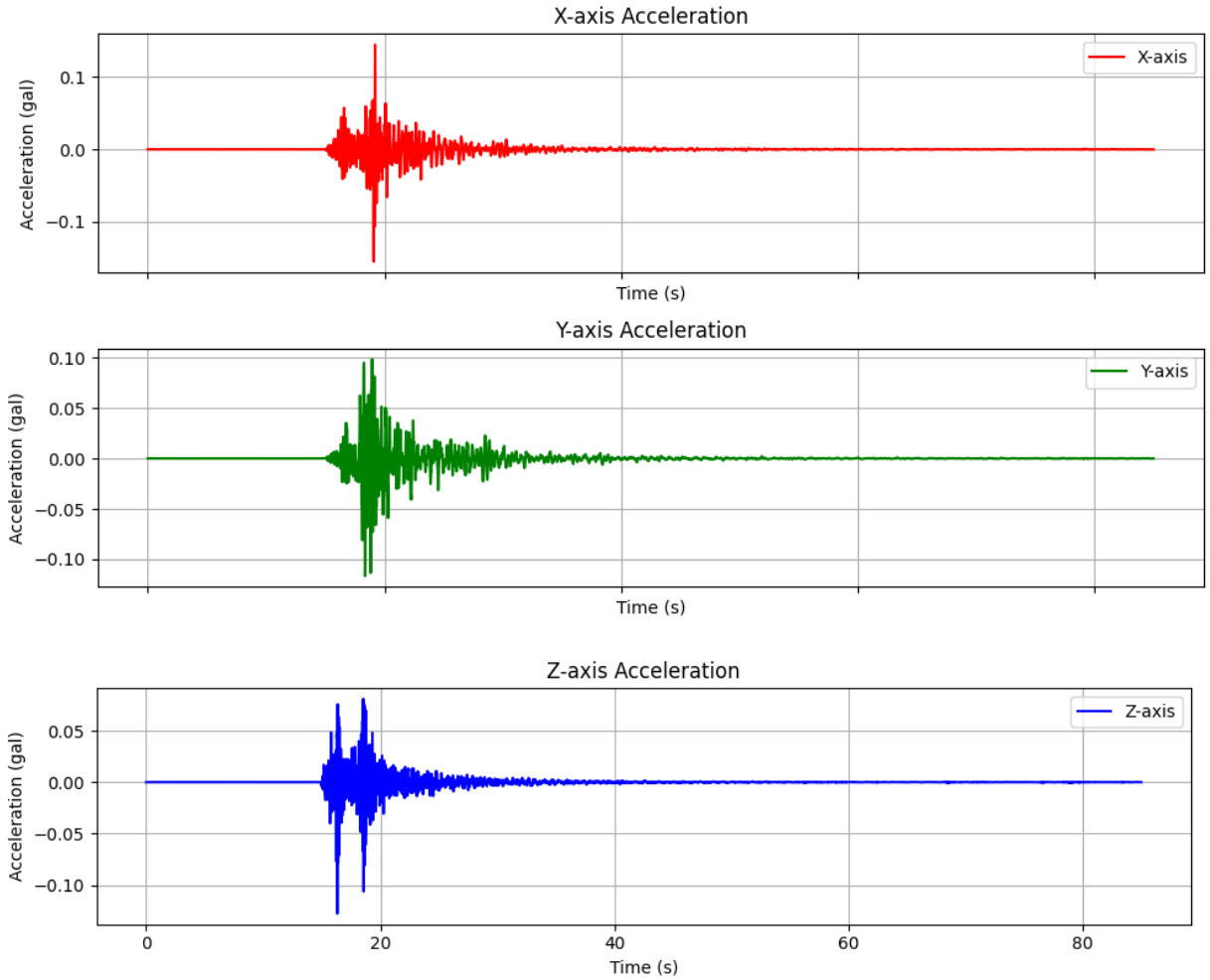
Mức độ mạnh yếu (biên độ lớn hay nhỏ).Xác định được thời điểm bắt đầu của một trận động đất.

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.

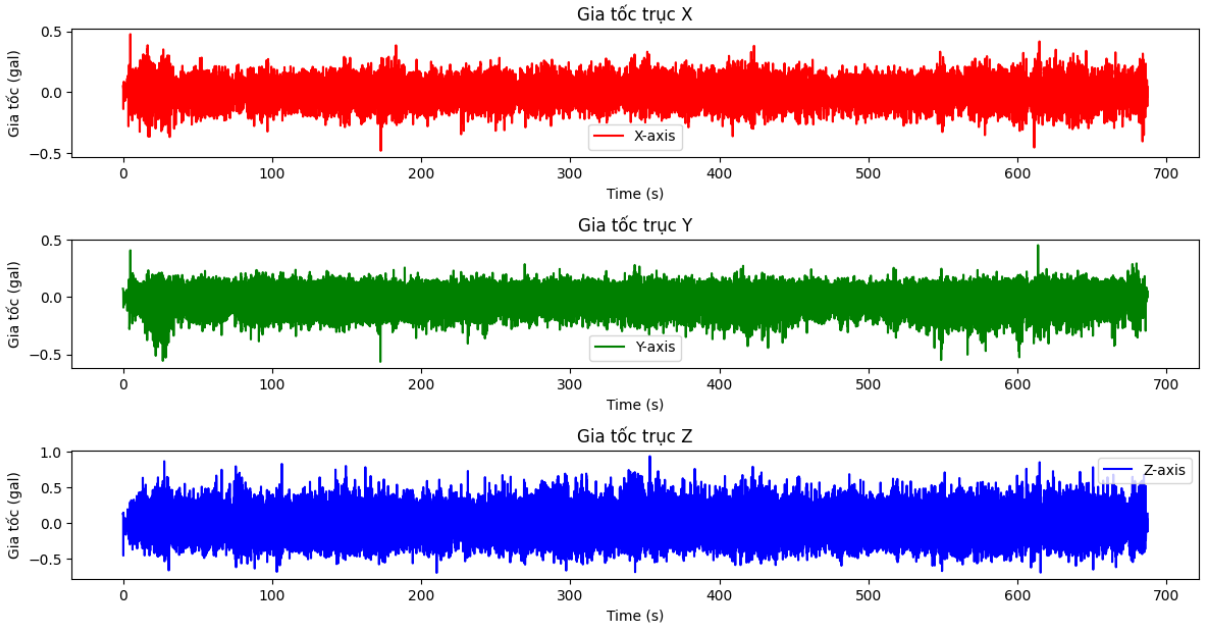
Trực quan của tập dữ liệu động đất được thể hiện theo ba trục x, y và z.



Hình 2. 4. Trực quan hoá dữ liệu động đất theo miền thời gian.

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.

Trực quan của tập dữ liệu nhiễu theo ba trục toạ độ x, y và z.

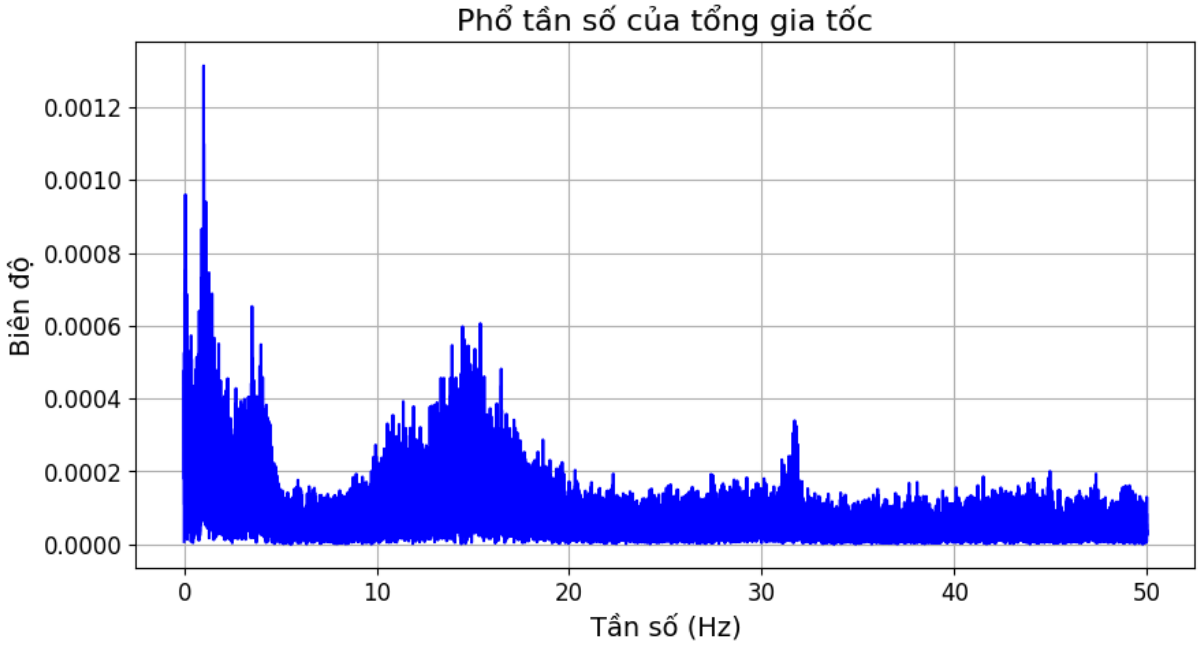


Hình 2.5. Trực quan hoá dữ liệu nhiễu theo miền thời gian.

2.2.2.2. Trực quan theo miền tần số (FFT)

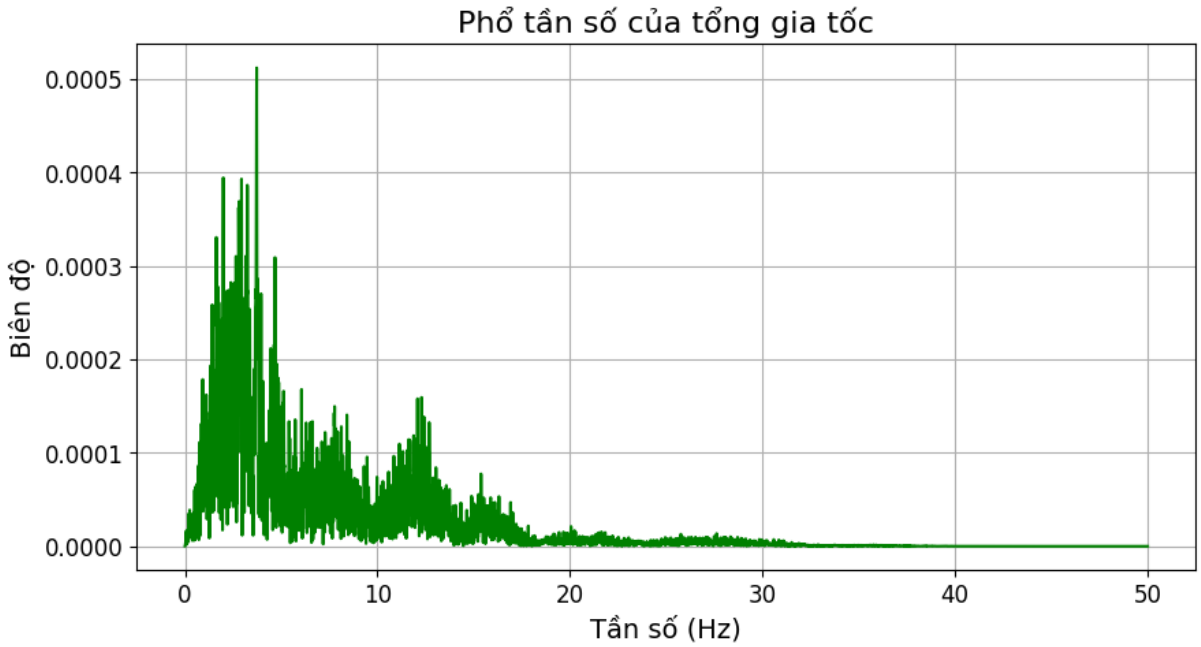
Việc trực quan hóa dữ liệu gia tốc theo miền tần số (FFT) giúp làm rõ các đặc trưng tần số có trong tín hiệu. Tín hiệu động đất thường tập trung ở tần số thấp (dưới 10 Hz) với biên độ rõ rệt, trong khi nhiễu thường xuất hiện ở tần số cao, phân bố không đều và không có mẫu đặc trưng.

Tập dữ liệu Noise:



Hình 2.6. Trực quan hoá dữ liệu nhiễu theo miền tần số FFT

Tập dữ liệu động đất:



Hình 2.7. Trực quan hoá dữ liệu động đất theo miền tần số FFT

2.2.2.3. Trực quan 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 (ví dụ xe cộ: từ 20 đến 60Hz).

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

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 (g^2/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

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 và giảm dần khi tăng tần số.

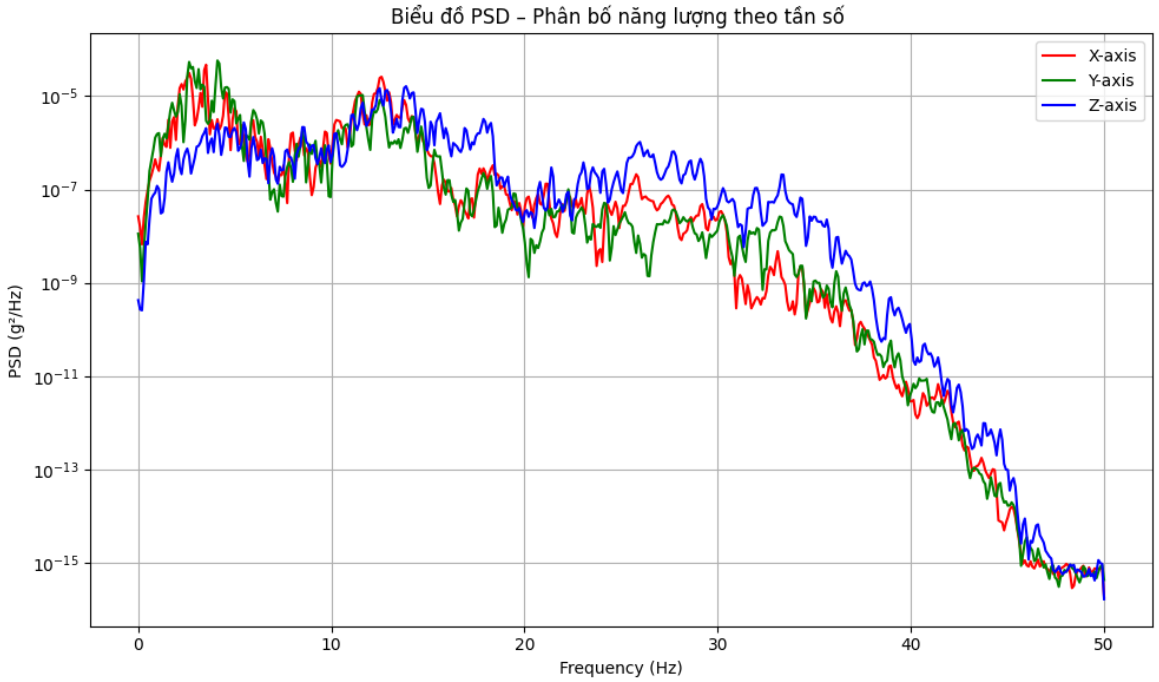
Kết luận: Nếu đường cong PSD có dạng "gò đồi" rõ ràng rồi rơi dốc → là EQ.

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

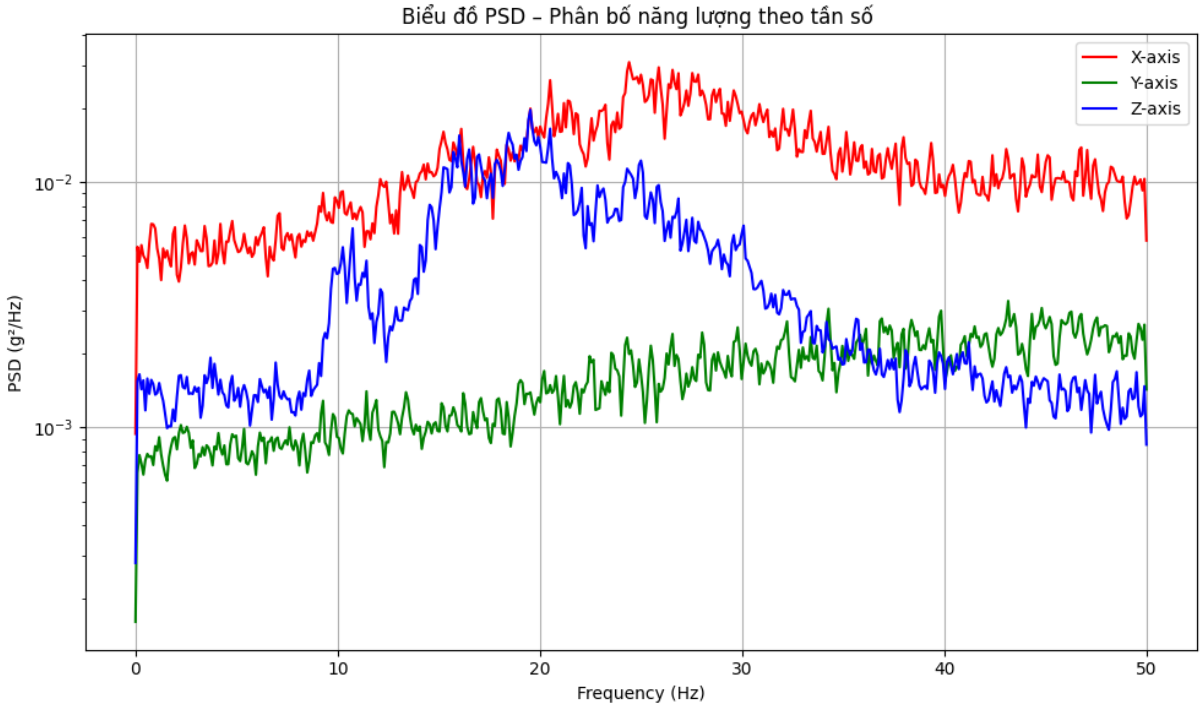
Sóng địa chấn tác động đến cả 3 trục nên PSD sẽ có dạng tương tự trên x, y, z.

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 2.8. Miền PSD của một trận động đất.



Hình 2.9. Miền PSD của một trận nhiễu

2.3. Tiền xử lý dữ liệu

2.3.1. Chuyển đổi 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.

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

$$Gia\ tốc\ \left(\frac{m}{s^2}\right) = \frac{Raw\ value}{Resolution} \times Dải\ đo\ \times 9.8055$$

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:

$$Gia\ tốc\ \left(\frac{m}{s^2}\right) = \frac{Raw\ value}{32768} \times 2 \times 9.8055$$

Chuyển đổi đơn vị từ m/s^2 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

$$1\ gal = 1\ cm/s^2 = 0.01\ m/s^2$$

Vậy để chuyển đổi từ đơn vị gia tốc (m/s^2) sang đơn vị gal ta chỉ cần nhân với 100:

$$Gia\ tốc\ (gal) = gia\ tốc\ (m/s^2) * 100$$

2.3.2. Làm sạch dữ liệu huấn luyện

2.3.2.1. Loại bỏ giá trị NaN

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.

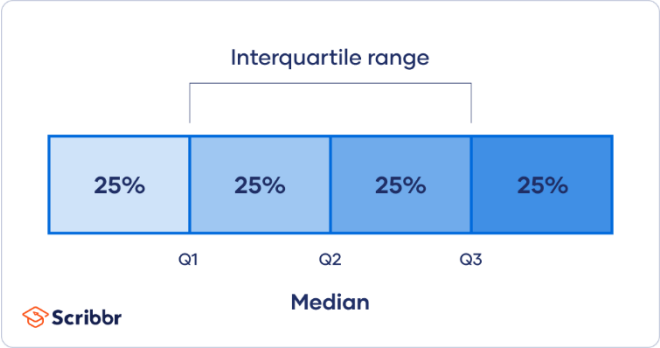
2.3.2.2. Loại bỏ dữ liệu ngoại lai (outliers)

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:



Hình 2.10. Phương pháp IQR

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 : $df.describer()$

Từ đó tính được giá trị IQR bằng công thức:

$$IQR = Q3 - Q1$$

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

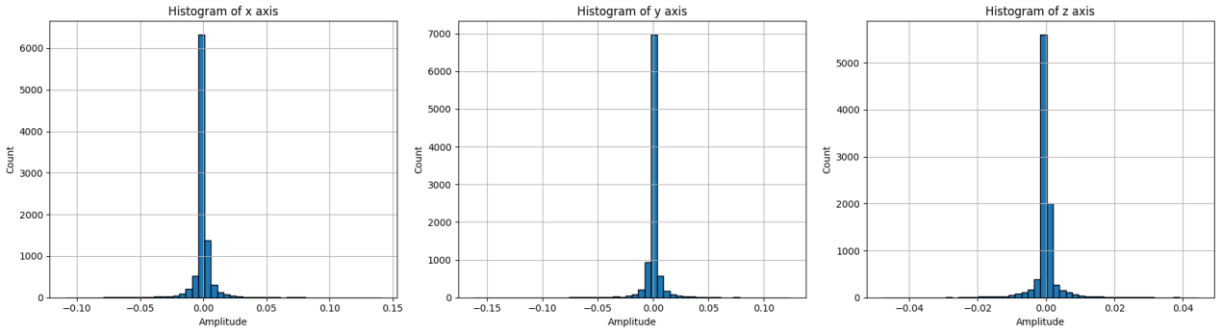
$$Lower\_threshold = Q1 - 1.5 \times IQR$$

$$Upper\_threshold = Q3 + 1.5 \times IQR$$

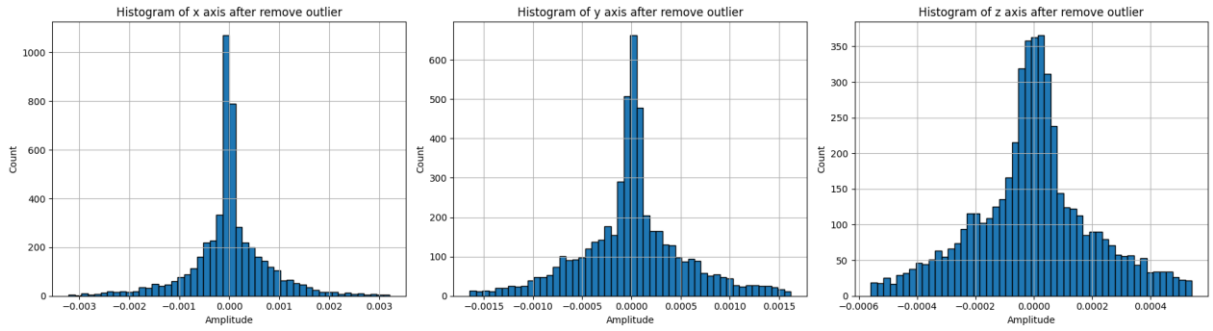
Những giá trị làm ngoài khoảng [Lower, upper] sẽ là những giá trị outlier.

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:

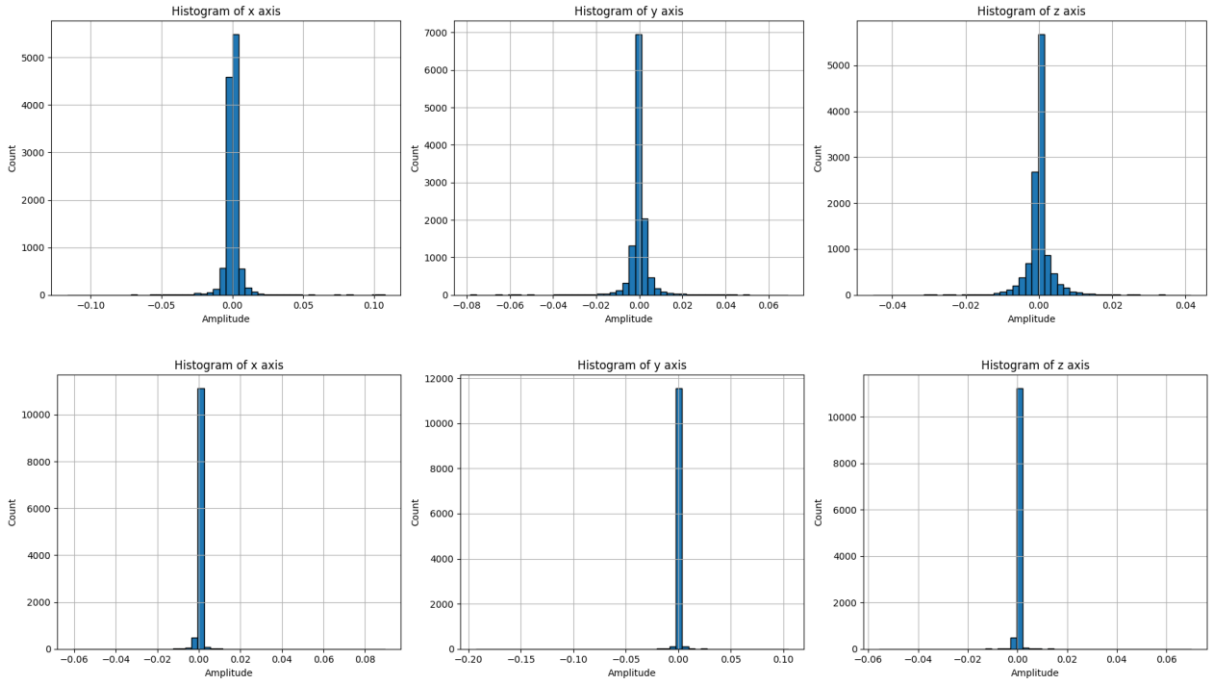


Sau khi loại bỏ outlier:



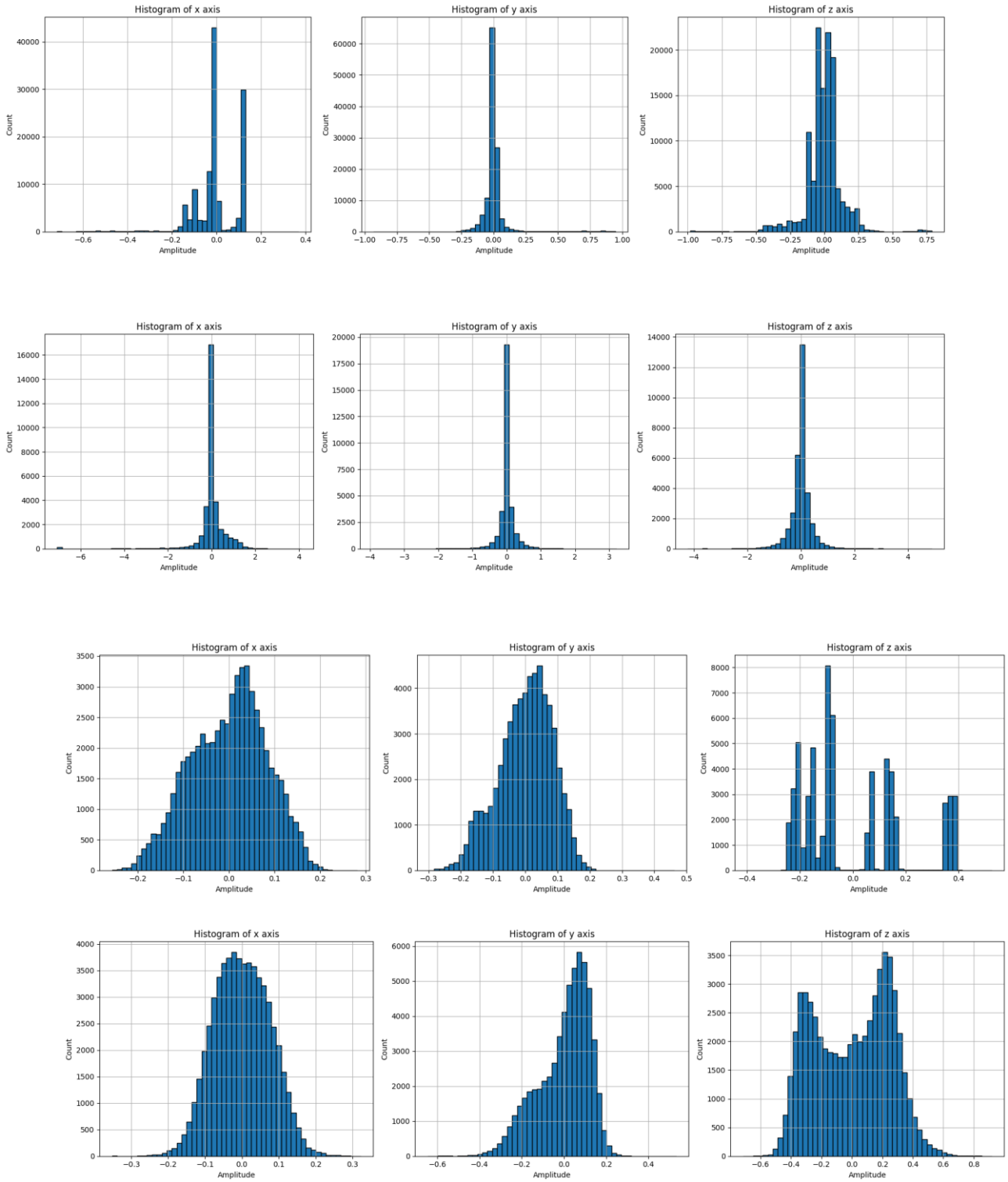
Hình 2.11. So sánh trước và sau khi loại bỏ giá trị ngoại lai.

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 2.12. Biểu đồ histogram của nhóm dữ liệu động đất

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



Hình 2.13. Biểu đồ histogram của nhóm dữ liệu nhiễu

2.3.3. Chuẩn hoá dữ liệu

Có một vài phương pháp thường để chuẩn hoá dữ liệu có thể nhắc tới như phương pháp Min-Max hay Z-score (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 đến rất lớn.

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.

$$Z = \frac{X - \mu}{\sigma}$$

Trong đó:

X là giá trị cần xét.

σ : độ lệch chuẩn

μ : Trung bình.

2.4. Trích xuất đặc trưng dữ liệu

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:

2.4.1. IQR- Interquartile Range

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.

2.4.2. Zero Crossing Rate – ZC

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.

2.4.3. 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...).

2.4.4. 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.

2.4.5. 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).

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

Giả sử $x_1, x_2, \dots, x_n$ là các số thực và xác định hàm, công thức tính độ lệch chuẩn dữ liệu:

$$\sigma(r) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(xi - r)}$$

2.4.7. 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.

2.4.8. Skew

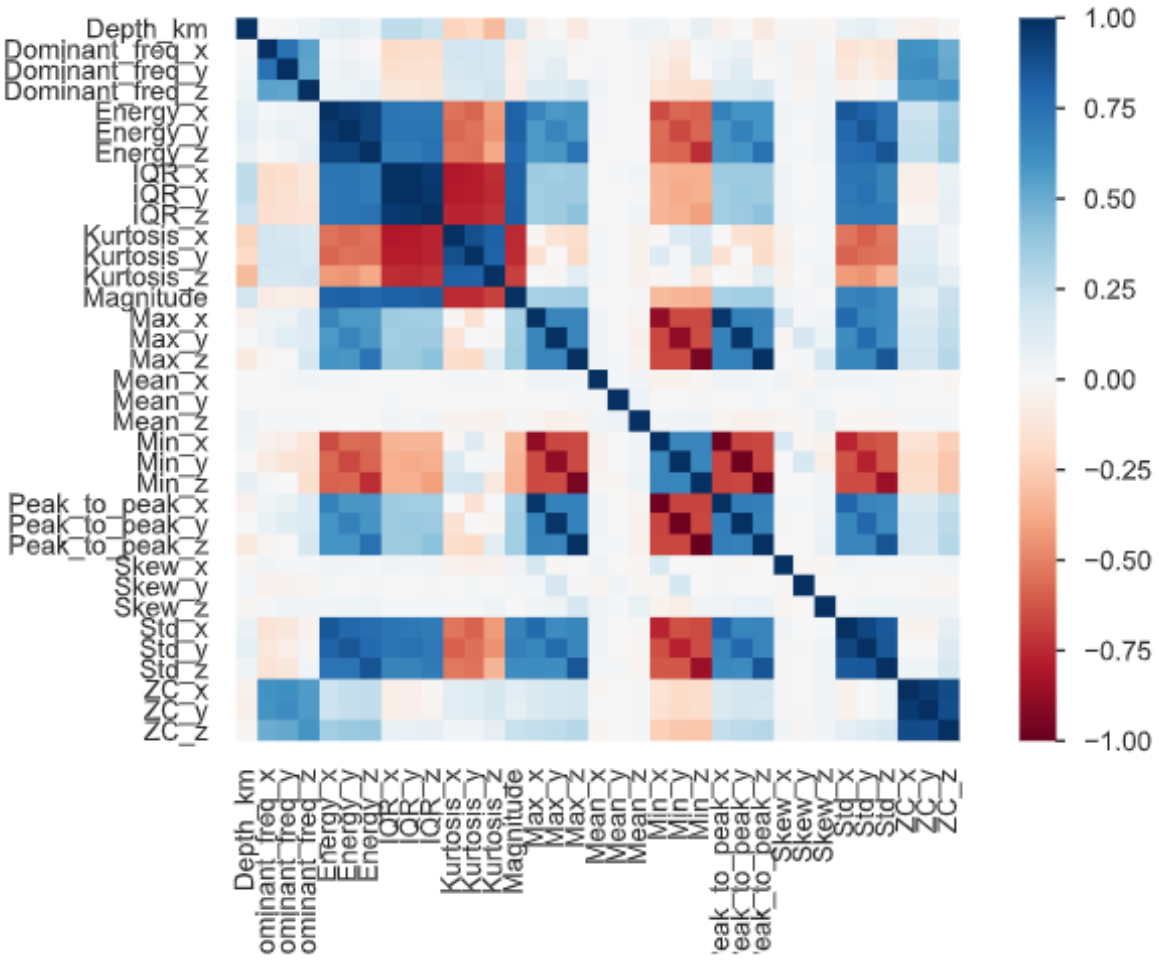
Đo tính bất đối xứng của phân phối tín hiệu. Dữ liệu động đất có thể tạo ra phân phối lệch rõ hơn so với nhiễu đều hoặc ngẫu nhiên.

2.4.9. 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.

Mối tương quan giữa các đặc trưng:



Hình 2.14. 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.

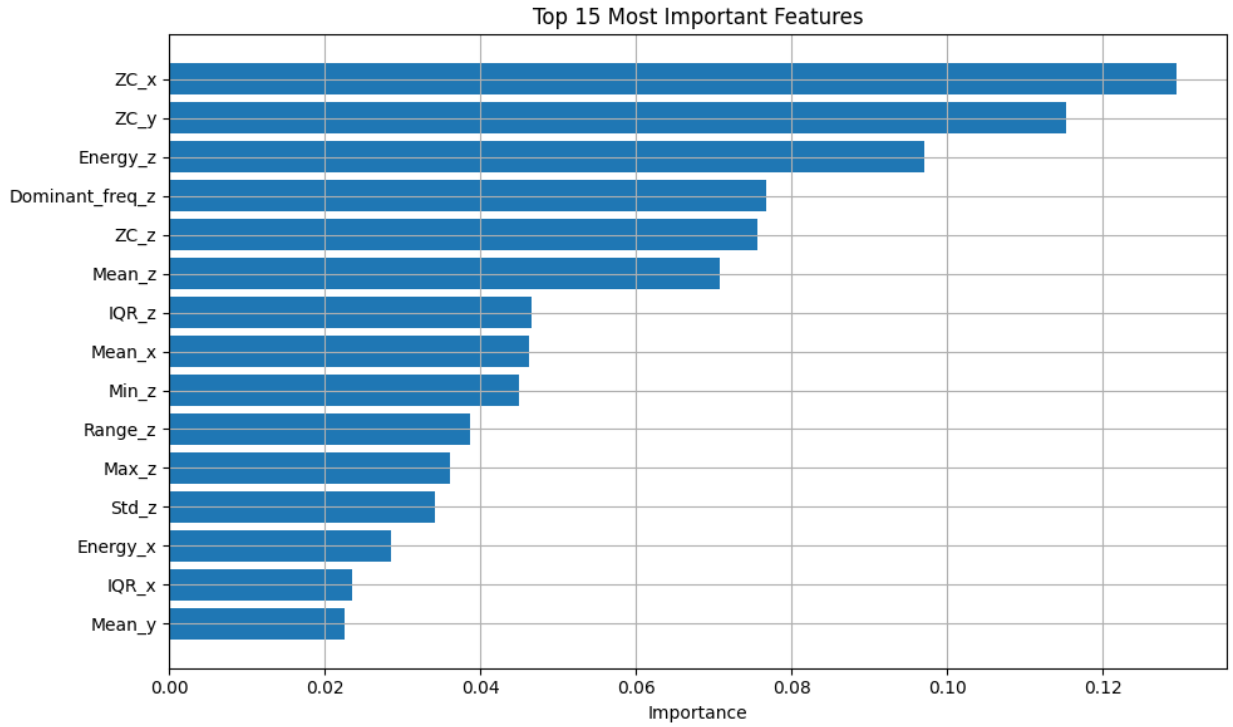
Còn nếu too many features: dễ gây overfitting làm tăng thời gian huấn luyện, khó diễn giải, có thể chứa nhiều nhiễu (noise).

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 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 2.15. Top 15 đặc trưng quan trọng lấy từ mô hình Random Forest.

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

2.5. Phân chia dữ liệu

Việc phân chia dữ liệu nhằm tách tập dữ liệu ban đầu thành các tập con chuyên biệt cho mục đích huấn luyện và đánh giá mô hình. Điều này giúp đảm bảo rằng các mô hình học máy được đánh giá trên dữ liệu chưa từng thấy trước đó, giảm thiểu nguy cơ overfitting và cung cấp cái nhìn thực tế về khả năng tổng quát hóa (generalization) của mô hình.

2.5.1. Train/Test split

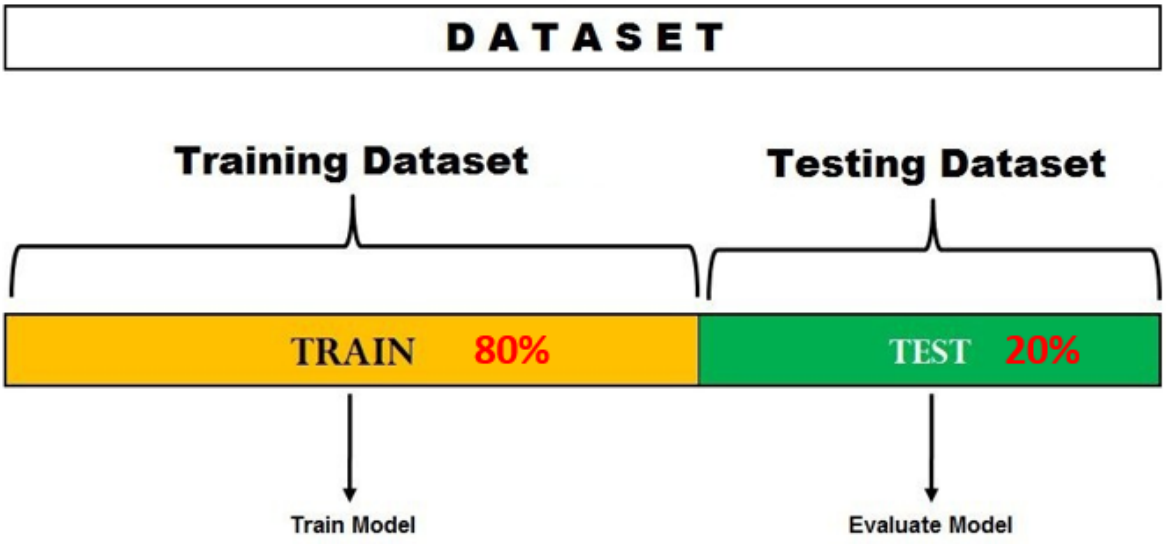
Ban đầu, tập dữ liệu được chia thành hai phần:

Tập huấn luyện (Training set): chiếm 80% tổng số mẫu.

Tập kiểm tra (Test set): chiếm 20% tổng số mẫu.

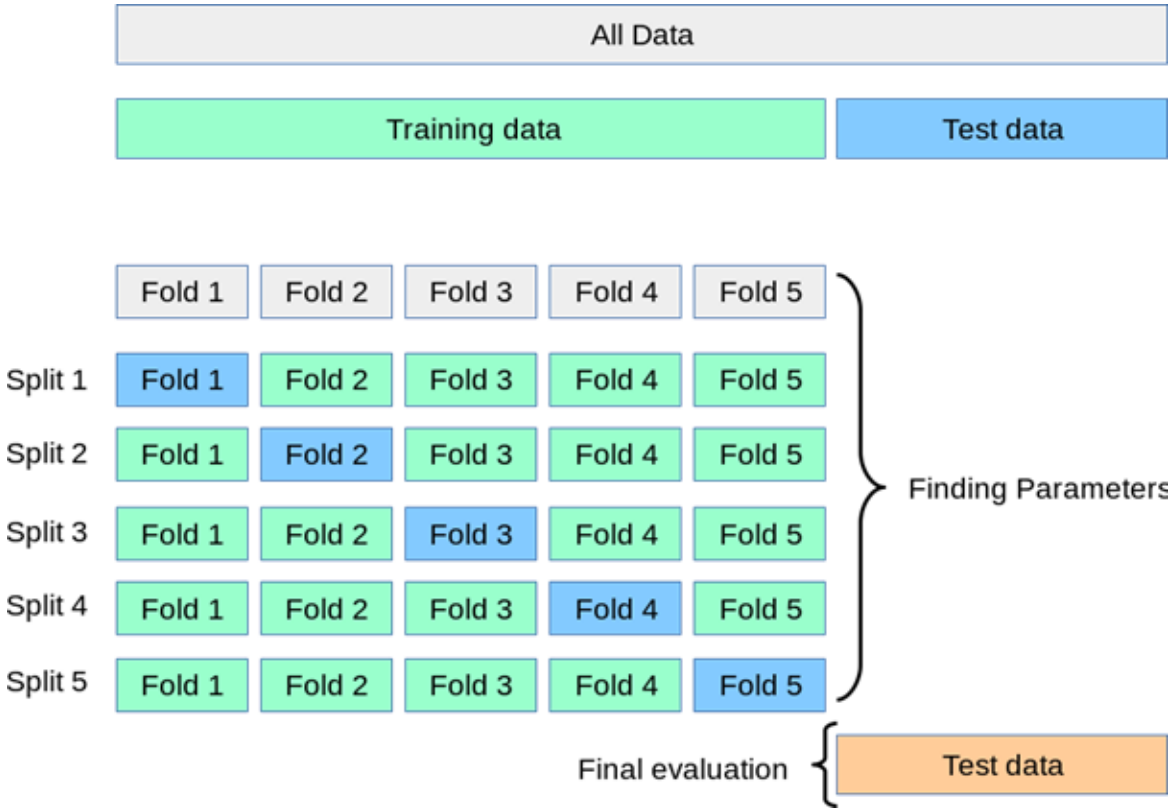
Việc phân chia được thực hiện một cách ngẫu nhiên, đồng thời đảm bảo rằng tỷ lệ giữa hai lớp (Earthquake và Noise) trong hai tập là tương đối cân bằng thông qua phương

pháp stratified sampling. Tập huấn luyện được sử dụng để tối ưu hóa trọng số mô hình, trong khi tập kiểm tra chỉ được sử dụng cho mục đích đánh giá cuối cùng.



Hình 2.16. Training – Testing Dataset

2.5.2. K-Folder Cross Validation



Hình 2.17. Phương pháp K-folder

Để tăng độ tin cậy cho việc đánh giá mô hình, kỹ thuật K-Fold Cross-Validation được áp dụng trên tập huấn luyện. Tập huấn luyện được chia thành K phần (folds) bằng nhau (trong nghiên cứu này, em có sử dụng K = 5).

Mỗi lần, một fold được dùng làm tập validation, bốn folds còn lại được dùng làm tập train.

Quá trình này lặp lại K lần, đảm bảo mỗi fold được sử dụng đúng một lần để đánh giá.

Kết quả cuối cùng được tính bằng trung bình các chỉ số đánh giá qua K lần.

Ưu điểm của K-Fold:

- Giảm sự phụ thuộc vào cách chia dữ liệu ban đầu.
- Sử dụng hiệu quả toàn bộ tập huấn luyện.
- Đánh giá độ ổn định và khả năng tổng quát hóa của mô hình tốt hơn.

2.5.3. Xử lý dữ liệu mất cân bằng

Trong dữ liệu thu thập được, số lượng mẫu giữa hai lớp Earthquake và Noise không hoàn toàn cân đối. Cụ thể, số lượng mẫu Noise lớn hơn số lượng mẫu Earthquake đáng kể, dẫn đến hiện tượng mất cân bằng lớp (class imbalance). Mức độ mất cân bằng được định lượng bằng công thức:

$$\text{Tỉ lệ} = \frac{\text{Số mẫu EQ}}{\text{Số mẫu Noise}}$$

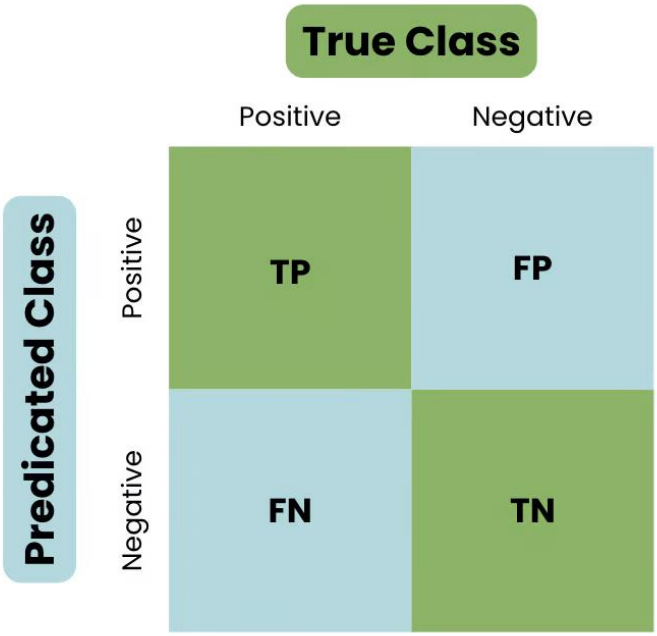
2.6. Đánh giá mô hình

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 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:

2.6.1. 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.



Hình 2.18. Ma trận nhầm lẫn (Confusion matrix)

2.6.2. Accuracy

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. Trong bài toán phân loại nhị phân, công thức accuracy được biểu diễn như sau:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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. 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.

2.6.3. Precision

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:

$$Precision = \frac{TP}{TP + FP}$$

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.

2.6.4. Recall

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:

$$Recall = \frac{TP}{TP + FN}$$

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

2.6.5. 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 và recall. F1-score được tính theo công thức:

$$F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

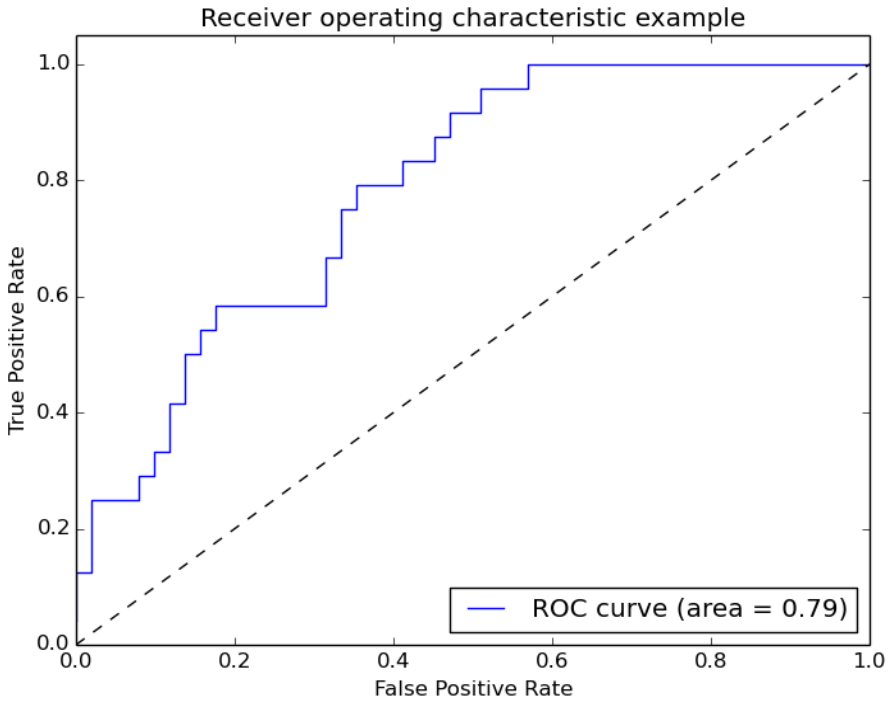
Trong ngữ cảnh phân biệt động đất, F1-score càng cao chứng tỏ mô hình vừa nhạy bén trong việc phát hiện động đất, vừa đáng tin cậy trong dự đoán.

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

2.6.6. 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 khái niệm cơ sở trước, đó là ROC Curve

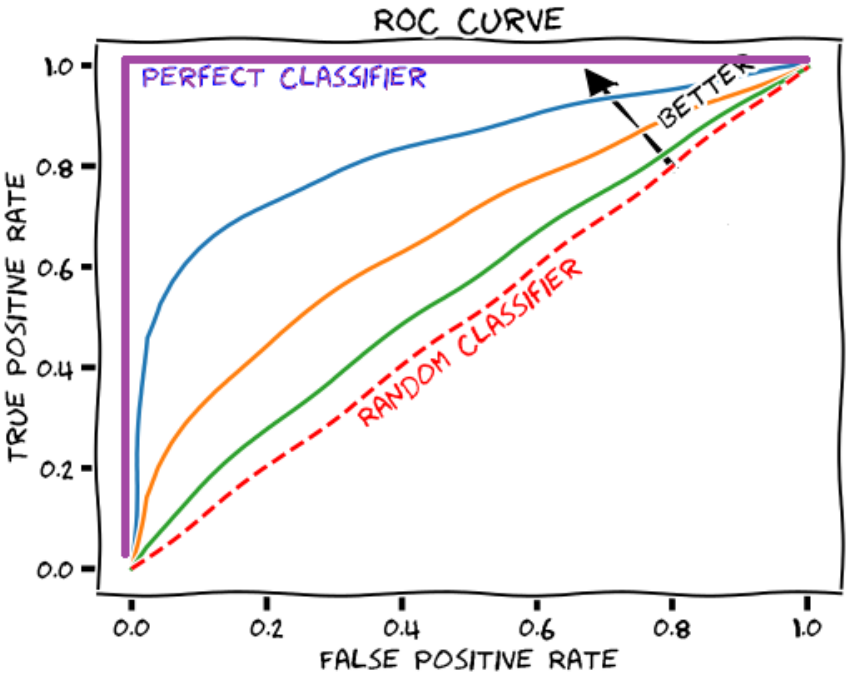
ROC Curve (The receiver operating characteristic curve) là một đường cong biểu diễn hiệu suất phân loại của một mô hình phân loại tại các ngưỡng threshold. Về cơ bản, nó hiển thị True Positive Rate (TPR) so với False Positive Rate (FPR) đối với các giá trị ngưỡng khác nhau.



Hình 2.19. 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].

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.



Hình 2.20. Biểu đồ ROC thể hiện hiệu quả của các bộ phân loại.

2.7. Công cụ và môi trường thực nghiệm

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:

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
Obspy	Đọc và xử lý tín hiệu địa chấn (file MSEED)

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 T4**	**TPU v2-8**
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 3: THỰC NGHIỆM VÀ KẾT QUẢ

3.1. Thiết lập thực nghiệm

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.

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)

3.2. Kết quả huấn luyện và đánh giá mô hình

3.2.1. Phân biệt giữa Động đất và nhiễu

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, 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.

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.

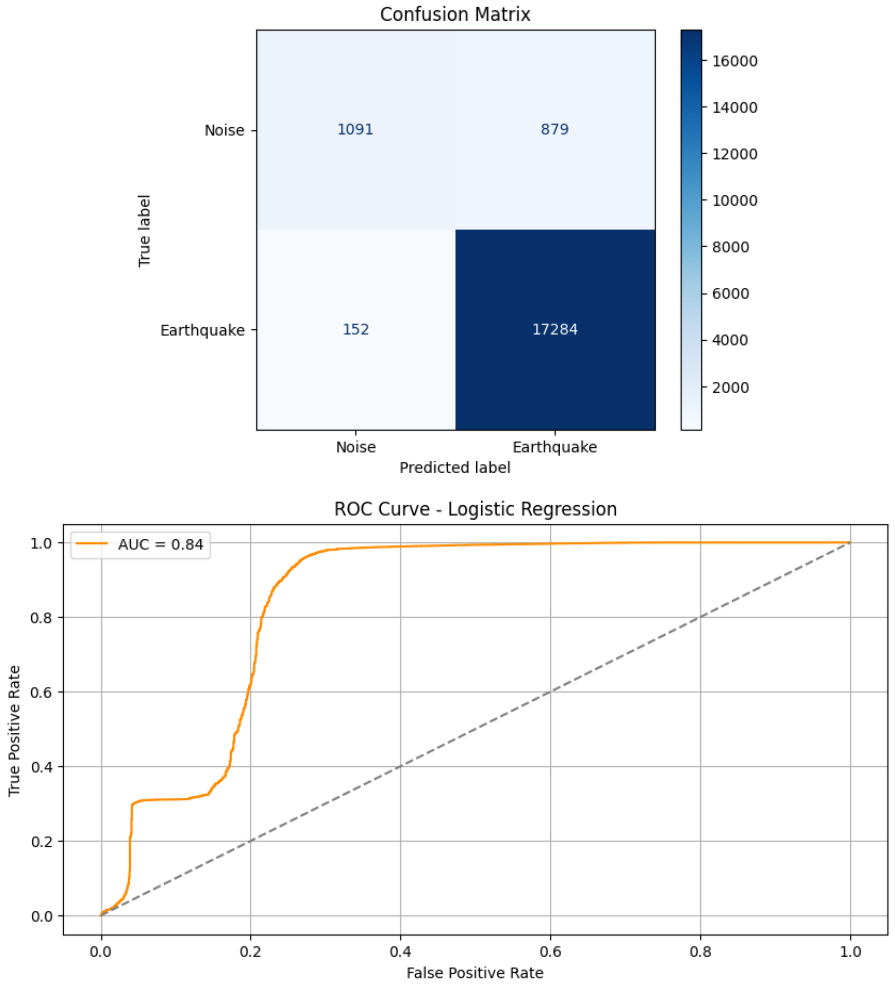
Artificial Neural Network (ANN) thể hiện khả năng học đặc trưng phi tuyến từ dữ liệu thô, tuy nhiên hiệu suất còn phụ thuộc vào thiết kế mạng và dữ liệu đầu vào đã được xử lý đặc trưng hay chưa.

Bảng 3.1. Thống kê 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
ANN [9]	0.976	0.8935	0.8646	0.975
CRNN [9]	0.9989	1.0	0.99	0.99

Kết quả của hai mô hình ANN và CRNN thu được là quá trình thực hiện chương trình có sẵn trong bài báo “CrowdQuake: A Networked System of Low-Cost Sensors for Earthquake Detection via Deep Learning” [9].

Mô hình Logistic Regression:

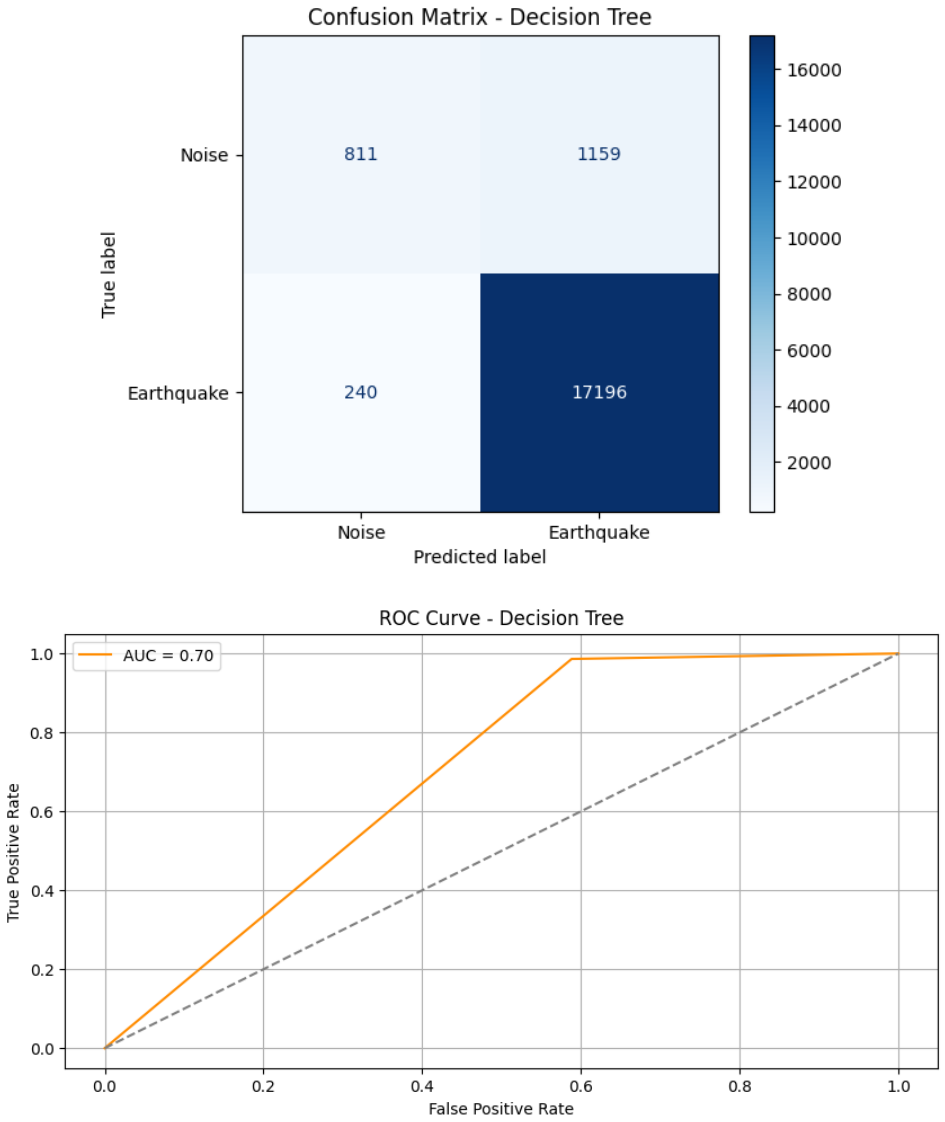


Hình 3.1: Kết quả huấn luyện mô hình Logistic Regression.

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:

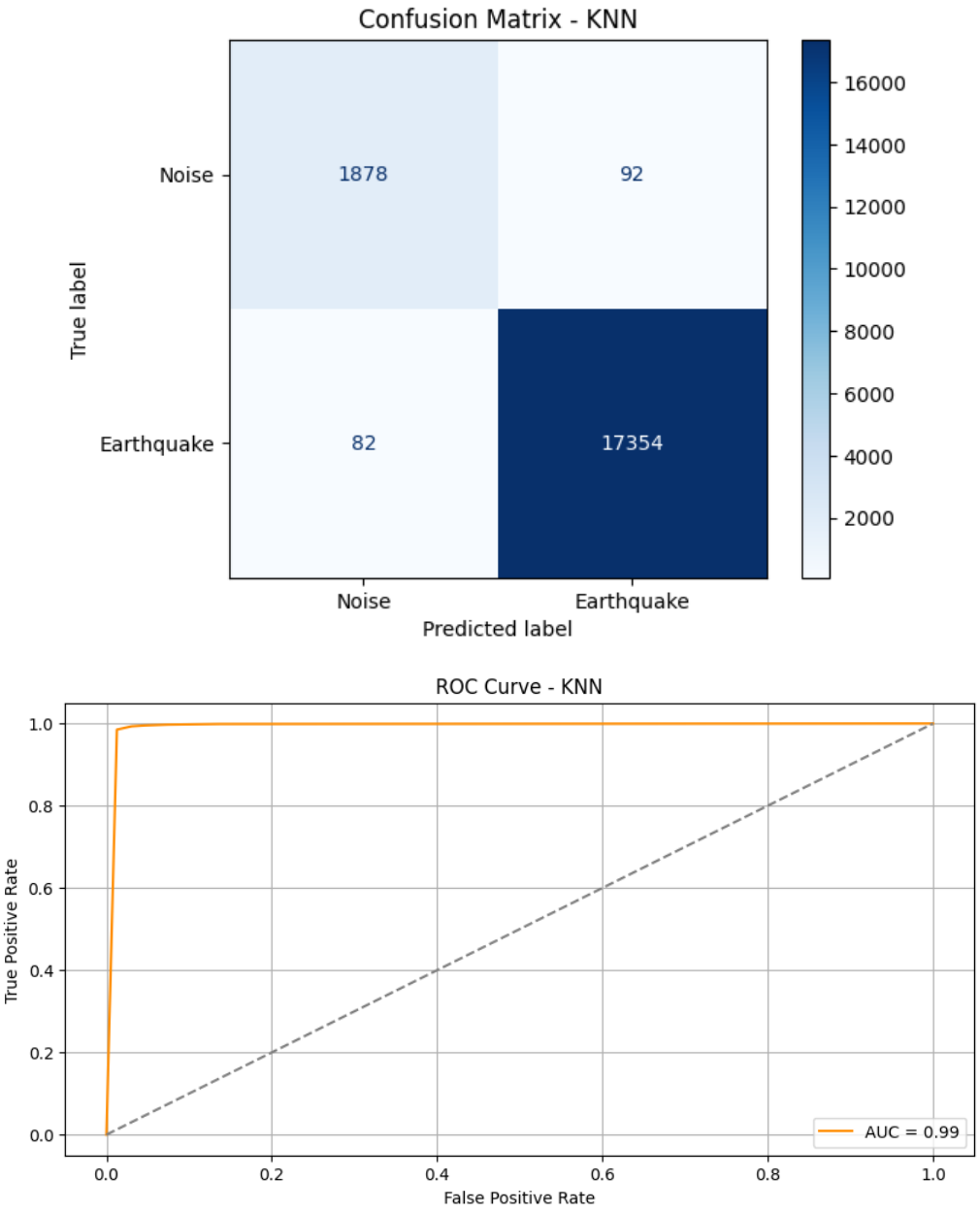


Hình 3.2: Kết quả huấn luyện mô hình Decision Tree.

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:

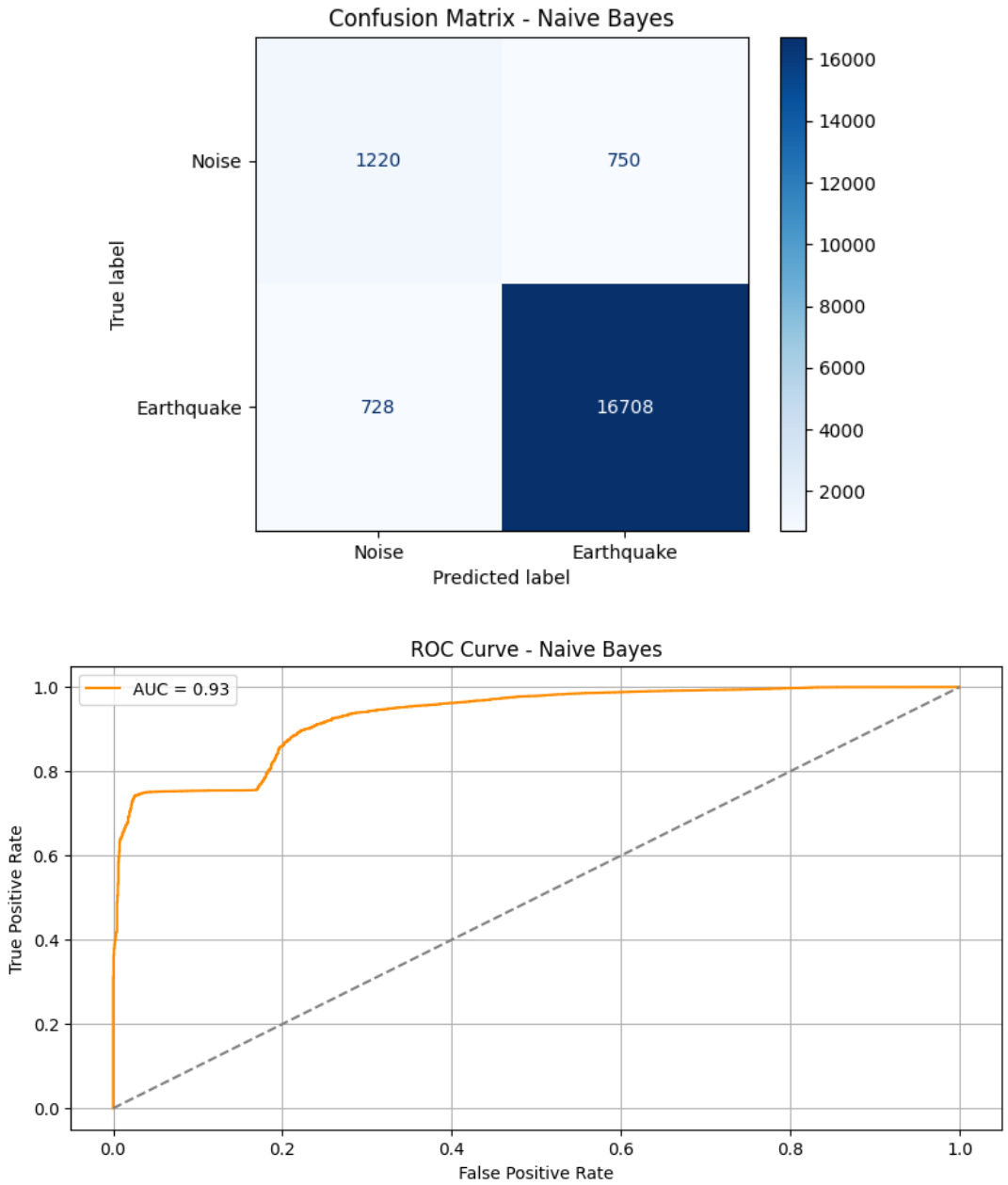


Hình 3.3: Kết quả huấn luyện mô hình Logistic Regression.

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:

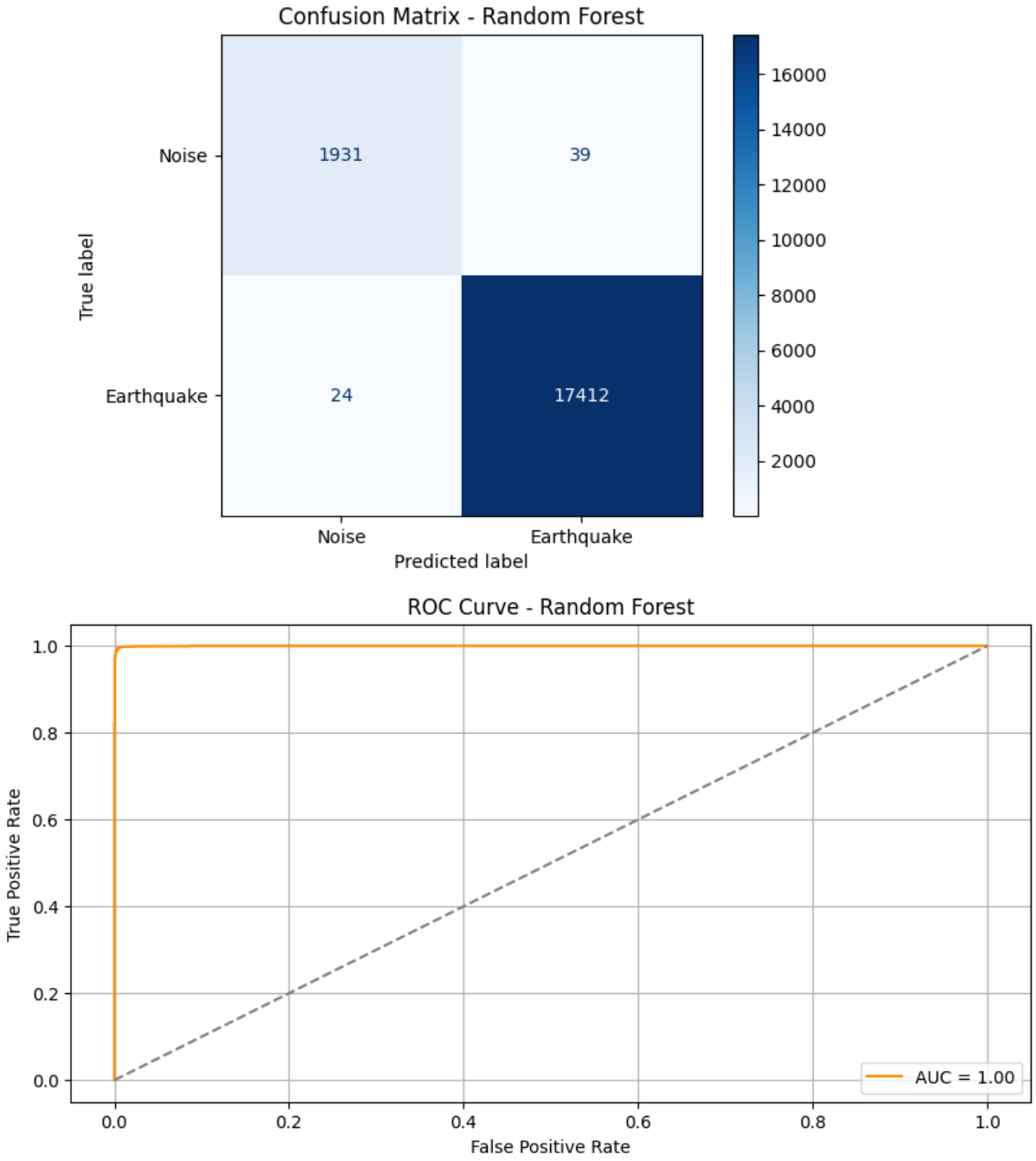


Hình 3.4: Kết quả huấn luyện mô hình Naïve Bayes.

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:

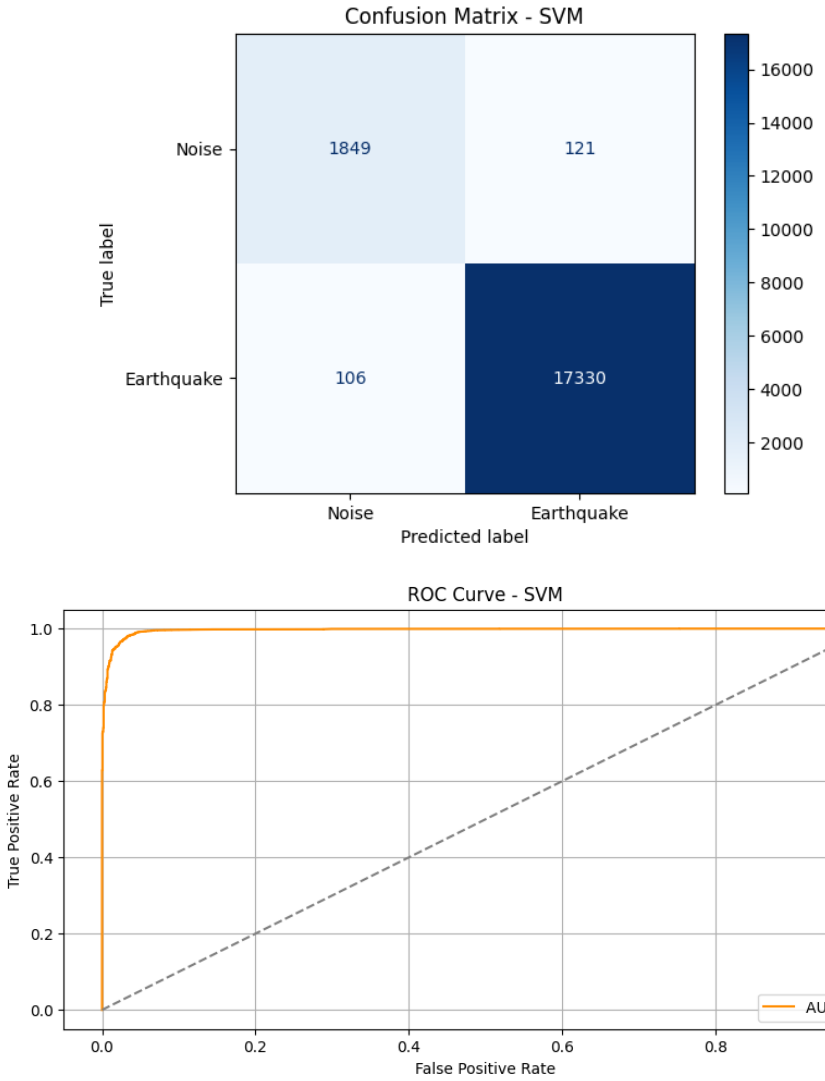


Hình 3.5: Kết quả huấn luyện mô hình Random Forest.

Nhận xét:

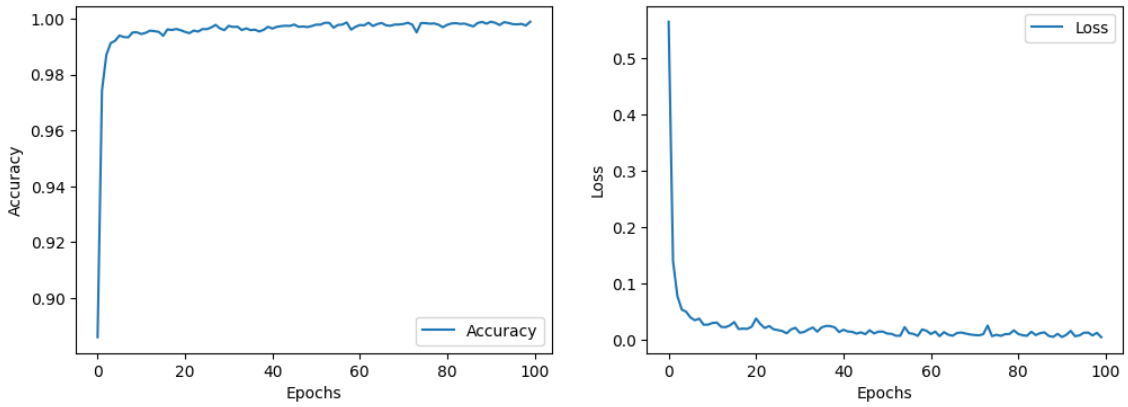
Mô hình Random Forest đạt hiệu suất rất cao với Accuracy 99.67%, Precision và Recall đều đạt 99%, cho thấy khả năng phân loại gần như tuyệt đối và rất ít bỏ sót động đất. F1-score cũng đạt 0.99, phản ánh sự ổn định và chính xác toàn diện của mô hình. Với khả năng xử lý tốt dữ liệu phi tuyến và chống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

Mô hình Support Vector Machine:



Hình 3.6: Kết quả huấn luyện mô hình SVM.

Mô hình CRNN:



Hình 3.7: Kết quả mô hình CRNN.

3.2.2. Đánh giá mô hình với bộ Italia:

Để làm tăng thêm tính xác thực và khả năng phán đoán của mô hình, em có thử nghiệm thêm trên bộ dữ liệu động đất của **Italia** và thu về khả năng khá khả quan, bảng dưới đây là kết quả thử nghiệm trên 10 dữ liệu bao gồm 5 EQ và 5 tập noise.

Bảng 3.2: Đánh giá thực nghiệm trên bộ dữ liệu Italia.

Dữ liệu	Thực tế	Dự đoán	Tỉ lệ dự đoán là EQ
Data 1	EQ	EQ	0.67
Data 2	EQ	EQ	0.73
Data 3	EQ	EQ	0.80
Data 4	EQ	EQ	0.67
Data 5	EQ	EQ	0.69
Data 6	Nhiễu	Nhiễu	0.37
Data 7	Nhiễu	EQ	0.71
Data 8	Nhiễu	Nhiễu	0.45
Data 9	Nhiễu	Nhiễu	0.33
Data 10	Nhiễu	Nhiễu	0.35

Tỉ lệ dự đoán EQ là do bộ dữ liệu em sẽ chia nhỏ thành nhiều cửa sổ, sau đó mô hình sẽ đưa ra phán đoán trên từng cửa sổ, cuối cùng sẽ tính trung bình cộng của các giá trị. Nếu như tập dữ liệu nào có tỉ lệ dự đoán lớn hơn 0.5 thì sẽ em sẽ kết luận là EQ và ngược lại.

Kết quả thử nghiệm với 10 tập dữ liệu khác nhau cho thấy mô hình có độ chính xác trung bình đạt khoảng 90%. Trong hầu hết các trường hợp, mô hình phân biệt tốt giữa tín hiệu địa chấn thực và các tín hiệu nhiễu nền. Tuy nhiên, một số tập có tỷ lệ sai cao hơn do tín hiệu bị nhiễu mạnh từ môi trường. Điều này cho thấy mô hình hoạt động ở mức ổn định và có tiềm năng ứng dụng trong thực tế, nhưng vẫn cần cải thiện thêm.

3.2.3. 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). Một trong những phương pháp phổ biến là sử dụng công thức:

$$M = \log_{10}(\frac{A}{A0})$$

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 R^2.

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 được tính như sau:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|yi - y'i|$$

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 được tính như sau:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(yi - y'i)^2$$

R^2(R-squared) là một metric dùng trong bài toán hồi quy. R^2 đo lường mức độ mà mô hình giải thích được phương sai của dữ liệu thực tế. Công thức:

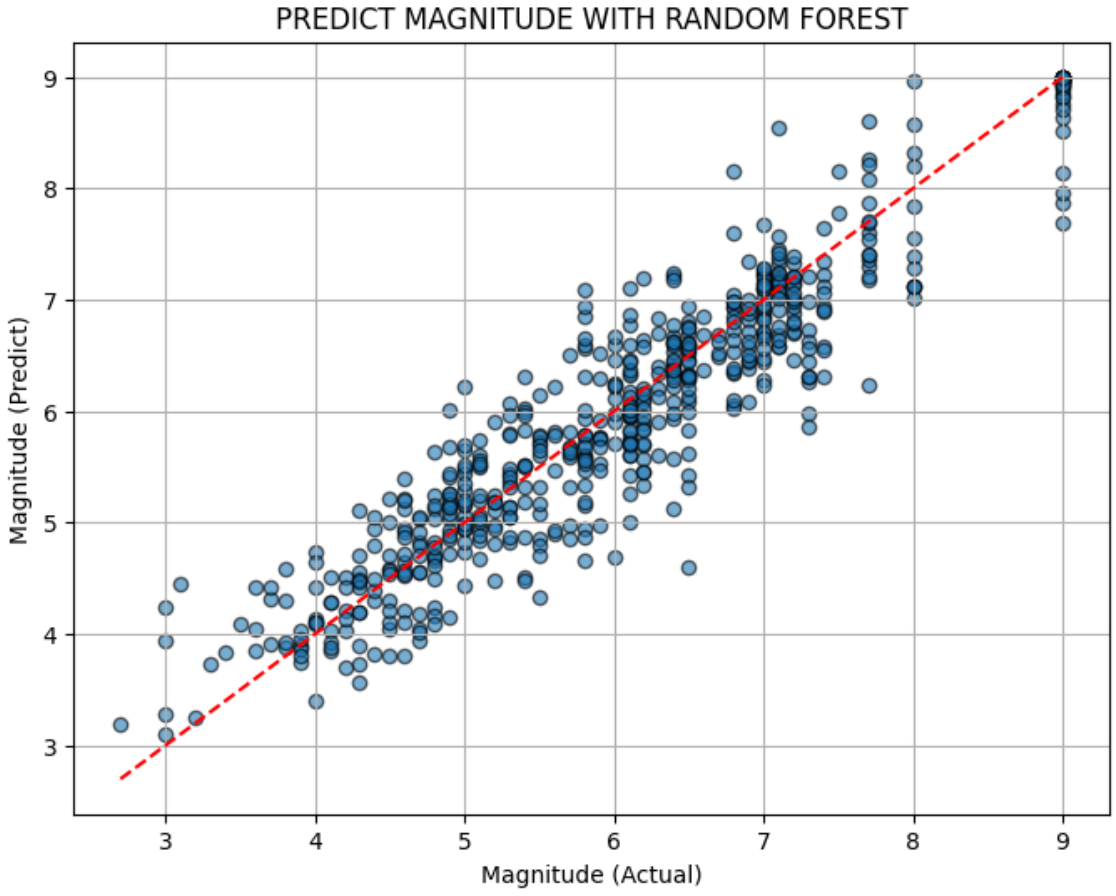
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Bảng 3.3. Thống kê kết quả các mô hình phán đoán độ lớn trận động đất.

Mô hình	MAE	MSE	R^2
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 R^2 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.

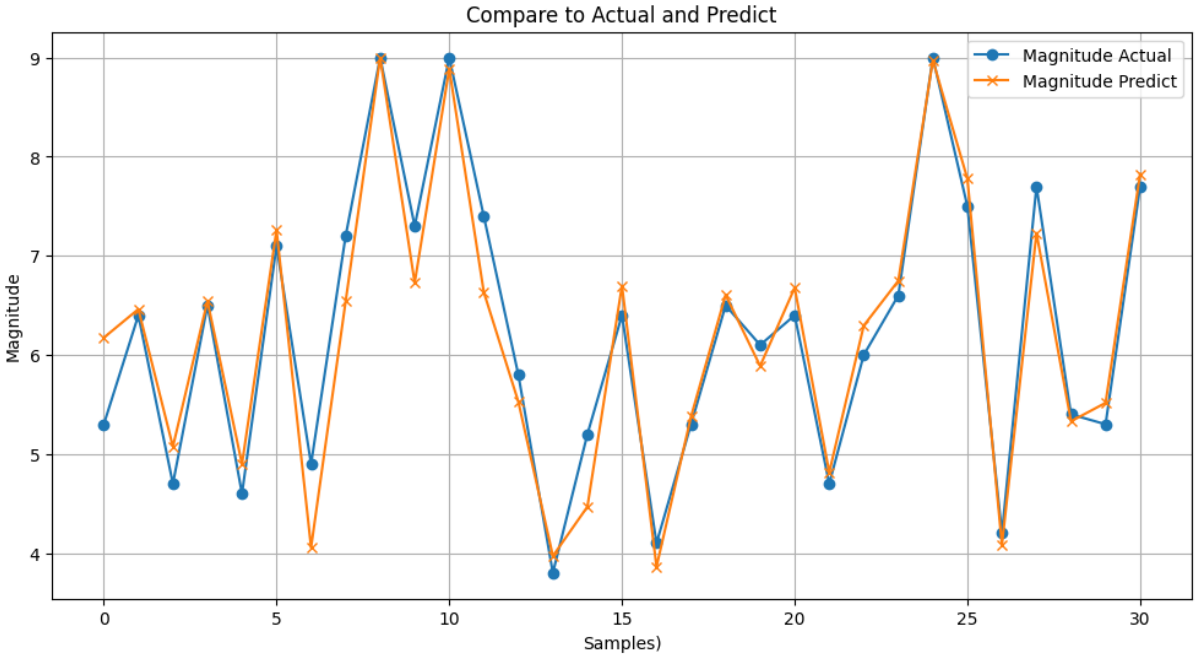


Hình 3.8. Biểu đồ phân tán giữa giá trị dự đoán và giá trị thực tế

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.

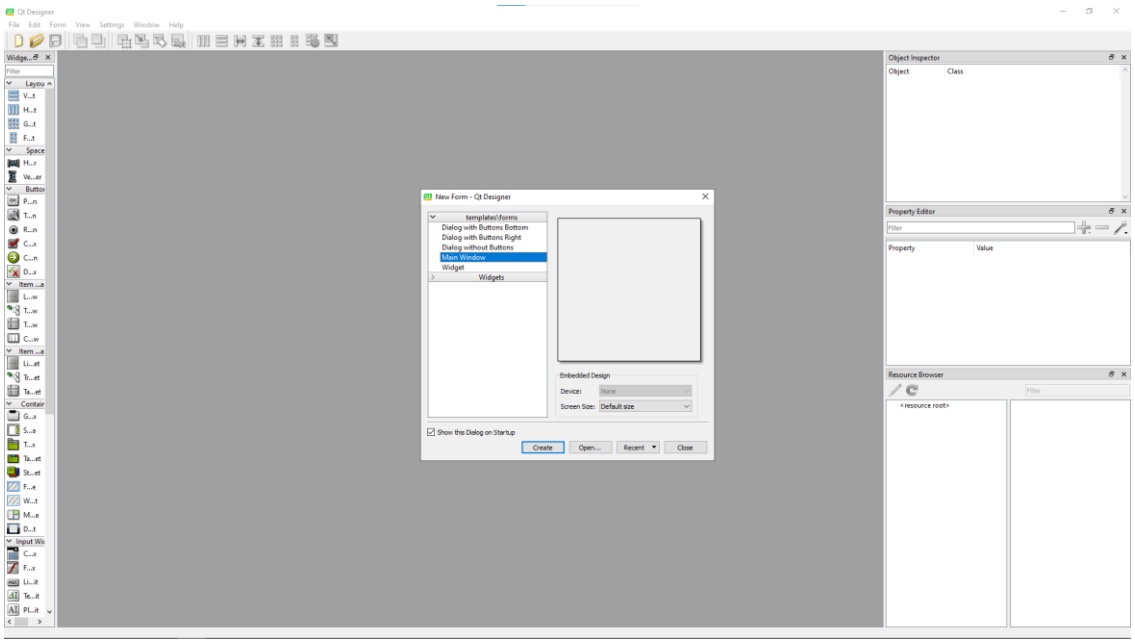
- Tuy nhiên vẫn có một số điểm lệch khá xa, nhất là ở vùng độ lớn lớn hơn 7.
- Mô hình chưa học tốt các trận động đất lớn, có thể vì số mẫu ít hoặc bị underfitting.



Hình 3.9. Biểu đồ miền thể hiện giá trị thực tế và giá trị dự đoán.

3.2.3. Thiết kế 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.

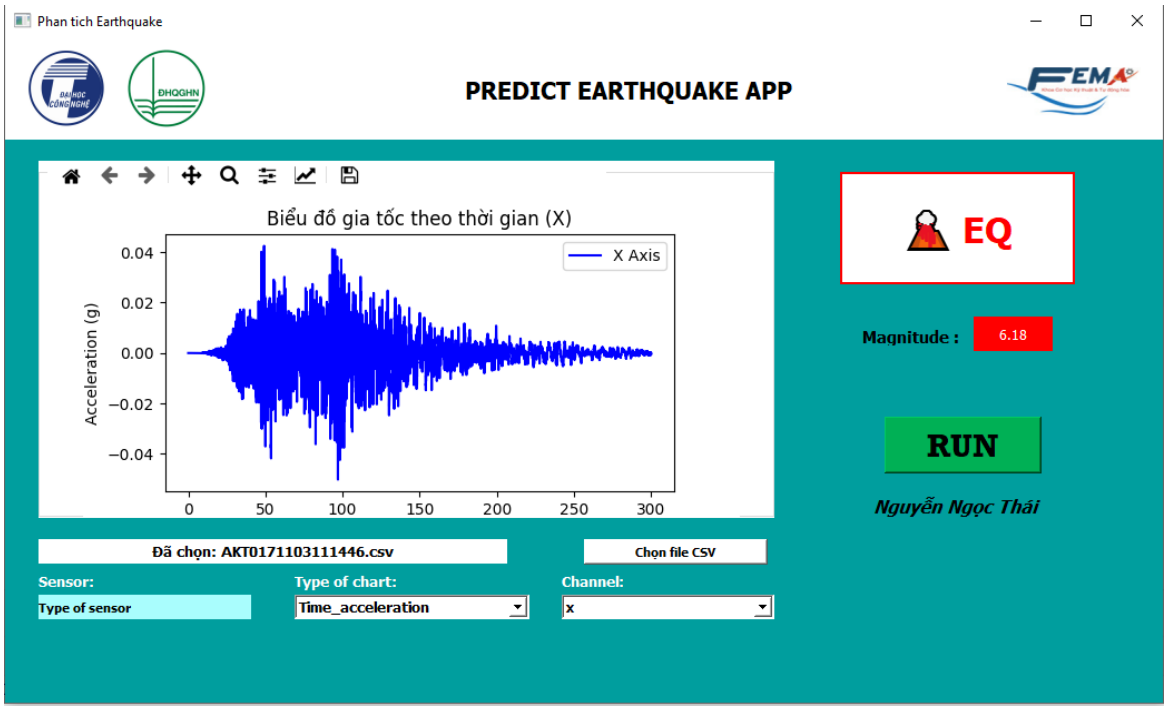


Hình 3.10. Giao diện phầm mềm QT Designer

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 khoá luận 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 3.11. Giao diện người dùng hệ thống

THẢO LUẬN VÀ KẾT LUẬN

Trong khuôn khổ khoá luận 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, 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 đạt độ chính xác cao nhất, SVM 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 và 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. Tuy nhiên, do hạn chế về cấu hình máy huấn luyện và kết quả đưa ra cũng chính xác nên em đã quyết định sử dụng mô hình máy học Random Forest. 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, khoá luận 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 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 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.

[3] Hoài N. (2025, March 29). Những trận động đất lớn nhất ở Việt Nam thế kỷ qua. Báo Điện Tử Tiền Phong. https://tienphong.vn/nhung-tran-dong-dat-lon-nhat-o-viet-nam-the-ky-qua-post1729297.tpo

[4] Tất cả về hệ thống giám sát địa chấn - IMV CORPORATION. (n.d.). IMV CORPORATION. https://we-are-imv.com/vi/support/library/seismograph/

Tiếng Anh:

[5] Abbadia, Jessica. "Understanding Earthquake Magnitudes and Their Impact." *Mind the Graph Blog*, 27 Feb. 2023, mindthegraph.com/blog/earthquake-magnitude/.

[6] Panchuk, Karla. "12.2 Seismic Waves and Measuring Earthquakes." *Opentextbc.ca*, 20 Aug. 2021, opentextbc.ca/physicalgeologyh5p/chapter/seismic-waves-and-measuring-earthquakes/.

[7] "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/.

[8] *IEEE Xplore Full-Text PDF:* (n.d.). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10623344

[9] *Data Download after Search for Data*. (n.d.). https://www.kyoshin.bosai.go.jp/kyoshin/data/index_en.htm

[10] *CrowdQuake+: Data-driven earthquake early warning via IoT and deep learning*. (2021, December 15). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9671971