**🎓 Slide 1: Trang bìa**

* **Tiêu đề**: Nghiên cứu bài toán phát hiện động đất sử dụng dữ liệu cảm biến gia tốc
* Sinh viên: Nguyễn Ngọc Thái
* Ngành: Công nghệ Kỹ thuật Cơ điện tử
* Giảng viên hướng dẫn: TS. Hoàng Văn Mạnh
* Trường Đại học Công nghệ – ĐHQGHN

**📌 Slide 2: Đặt vấn đề**

* Động đất là thiên tai khó dự đoán, gây hậu quả nặng nề.
* Việc phát hiện sớm giúp giảm thiểu thiệt hại.
* Hệ thống cảnh báo truyền thống còn nhiều hạn chế (độ trễ, cảnh báo sai, tốn chi phí).
* Mục tiêu: ứng dụng AI và cảm biến gia tốc để tự động phát hiện động đất.

**🌏 Slide 3: Thực trạng động đất**

* Trên thế giới: trung bình 20.000 trận mỗi năm, nhiều trận gây thiệt hại lớn (Tohoku 2011, Haiti 2010,...).
* Ở Việt Nam: không nằm trong vành đai lửa nhưng vẫn bị ảnh hưởng (Điện Biên, Tuần Giáo, Kon Plông,...).
* Động đất do cả tự nhiên và nhân sinh (thủy điện, công trình ngầm).

**🚨 Slide 4: Tầm quan trọng của phát hiện sớm**

* Cảnh báo kịp thời cứu sống nhiều người.
* Kích hoạt tự động dừng tàu, ngắt gas, cảnh báo cộng đồng.
* Hạn chế cảnh báo giả, tăng độ tin cậy hệ thống.

**⚠️ Slide 5: Hạn chế của hệ thống truyền thống**

* Độ trễ cao, dễ báo sai khi có nhiễu.
* Không phát hiện được ở vùng thiếu cảm biến.
* Chi phí lắp đặt và duy trì lớn.
* Khó mở rộng trong điều kiện hạ tầng hạn chế.

**🎯 Slide 6: Mục tiêu nghiên cứu**

* Xây dựng hệ thống học máy phân biệt động đất và nhiễu từ dữ liệu cảm biến gia tốc.
* Tiền xử lý dữ liệu, trích xuất đặc trưng.
* Huấn luyện và đánh giá mô hình học máy.
* Triển khai giao diện thử nghiệm cho người dùng.

**📚 Slide 7: Giới thiệu về động đất và cảm biến**

* Động đất do hoạt động kiến tạo, núi lửa hoặc nhân sinh.
* Cảm biến gia tốc (accelerometer) đo dao động mặt đất 3 trục.
* MEMS – phổ biến, rẻ, nhạy, phù hợp mạng IoT cảnh báo.

**🧠 Slide 8: Các mô hình học máy áp dụng**

* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machines
* Artificial Neural Network (ANN)
* Convolutional Recurrent Neural Network (CRNN)

**📦 Slide 9: Nguồn dữ liệu**

* **Bộ dữ liệu K-NET** (Nhật Bản)
* **Bộ dữ liệu Italy**
* Bao gồm tín hiệu EQ (động đất) và Noise (nhiễu).
* Dữ liệu gia tốc 3 trục: X, Y, Z, sampling rate 100Hz.

**🔍 Slide 10: Tiền xử lý dữ liệu**

* Chia nhỏ tín hiệu bằng sliding window 10 giây.
* Chuyển đổi đơn vị, loại bỏ outlier bằng IQR.
* Chuẩn hóa bằng Min-Max, Z-score.
* Gán nhãn EQ/Noise thủ công hoặc theo sự kiện.

**📊 Slide 11: Trích xuất đặc trưng**

* IQR, ZCR, Energy, Peak-to-Peak, Skew, Kurtosis, Dominant Frequency, PSD...
* Biến tín hiệu thành vector đặc trưng → input cho mô hình học máy.

**🤖 Slide 12: Xây dựng và huấn luyện mô hình**

* Dữ liệu huấn luyện: train/val/test
* K-Fold cross-validation
* Đánh giá: Accuracy, Precision, Recall, F1, ROC-AUC
* **Random Forest cho kết quả tốt nhất**

**🖥️ Slide 13: Giao diện người dùng**

* Sử dụng PyQt5 để tạo giao diện nhập dữ liệu.
* Cho phép người dùng chọn file, chạy phân loại, hiển thị kết quả EQ/Noise.
* Demo chức năng phân loại dữ liệu theo mô hình huấn luyện.

**🧩 Slide 14: Hạn chế & Hướng phát triển**

* Chưa triển khai real-time.
* Dữ liệu vẫn còn hạn chế, chưa phủ toàn quốc.
* Chưa xác định tâm chấn hay cường độ.
* Hướng tới mở rộng mạng cảm biến, tích hợp vào hệ thống cảnh báo quốc gia.

**✅ Slide 15: Kết luận**

* Đề tài cho thấy tiềm năng lớn trong phát hiện sớm động đất bằng học máy.
* Hệ thống có thể phân biệt tốt EQ và Noise.
* Gợi mở hướng ứng dụng rộng rãi, tiết kiệm chi phí, dễ triển khai tại Việt Nam.

BỘ CÂU HỎI CÓ THỂ HỎI TRONG QUÁ TRÌNH PHẢN BIỆN

**1. Câu hỏi về lý do chọn đề tài**

* Vì sao bạn chọn đề tài này trong khi Việt Nam không phải là quốc gia có nguy cơ động đất cao?
* Đề tài này có ý nghĩa gì về mặt ứng dụng thực tiễn tại Việt Nam?
* Bạn có thể so sánh hệ thống bạn nghiên cứu với một hệ thống cảnh báo động đất đang triển khai thực tế (ví dụ: ShakeAlert của Mỹ, EEW của Nhật)?

**📌 2. Câu hỏi về cơ sở lý thuyết**

* Sự khác biệt giữa cấp độ (Magnitude) và cường độ (Intensity) động đất là gì?
* Tại sao động đất gây ảnh hưởng nhiều đến sóng P, S, còn nhiễu lại không?
* Cảm biến gia tốc hoạt động dựa trên nguyên lý vật lý nào?

**📌 3. Câu hỏi về dữ liệu**

* Vì sao bạn chọn bộ dữ liệu K-NET và Italy? Dữ liệu này có những ưu nhược điểm gì?
* Trong bộ dữ liệu, bạn làm thế nào để đảm bảo dữ liệu động đất và nhiễu được gán nhãn chính xác?
* Với dữ liệu từ điện thoại (cảm biến MEMS), nếu nhiễu tăng, mô hình của bạn có còn hiệu quả không?

**📌 4. Câu hỏi về tiền xử lý dữ liệu**

* Vì sao bạn dùng kỹ thuật **sliding window 10 giây**? Có tối ưu không?
* Bạn sử dụng phương pháp nào để xử lý **outliers**, tại sao lại chọn IQR?
* Vì sao phải chuẩn hóa dữ liệu (z-score, min-max)? Nếu không chuẩn hóa thì chuyện gì xảy ra?

**📌 5. Câu hỏi về trích xuất đặc trưng (feature extraction)**

* Tại sao bạn chọn các đặc trưng như ZCR, Energy, Kurtosis? Đặc trưng nào là mạnh nhất?
* Nếu loại bỏ đặc trưng PSD thì độ chính xác có giảm không?
* Bạn có sử dụng thuật toán giảm chiều không (PCA, SelectKBest,...) không?

**📌 6. Câu hỏi về mô hình học máy**

* Vì sao mô hình Random Forest lại hoạt động tốt hơn các mô hình còn lại?
* Mô hình bạn chọn có overfitting không? Bạn kiểm tra như thế nào?
* Nếu triển khai real-time, bạn ưu tiên mô hình nào: Random Forest hay CRNN? Vì sao?

**📌 7. Câu hỏi về đánh giá mô hình**

* Ý nghĩa thực tế của các chỉ số: Precision, Recall, F1-score trong bài toán này là gì?
* Bạn có gặp class imbalance (mất cân bằng nhãn) không? Cách khắc phục?
* Nếu cảnh báo nhầm (false positive) thì hậu quả là gì? Trường hợp nào thì cần ưu tiên Recall hơn Precision?

**📌 8. Câu hỏi mở rộng và thực tiễn**

* Mô hình của bạn có thể tích hợp vào smartphone không?
* Nếu triển khai toàn quốc, cần điều kiện gì về hạ tầng cảm biến và phần mềm?
* Trong tương lai, bạn có định tích hợp thêm dữ liệu GPS, cảm biến rung, hoặc AI generative để mô phỏng sóng địa chấn?

**📌 9. Câu hỏi về giới hạn đề tài**

* Đề tài của bạn chưa thực hiện real-time, vậy theo bạn, thách thức lớn nhất để làm real-time là gì?
* Mô hình của bạn chưa xác định được tọa độ tâm chấn hay cường độ – bạn dự kiến giải quyết thế nào?
* Nếu áp dụng trong hệ thống cảnh báo quốc gia, cần cải tiến gì về hiệu năng và độ tin cậy?

**📌 10. Câu hỏi phản biện chuyên sâu**

* PSD của tín hiệu động đất có đặc điểm gì khác biệt với nhiễu? Làm sao phân biệt khi có sự chồng lấn?
* Bạn chọn random forest với bao nhiêu cây (n\_estimators)? Bạn đã thử tuning hyperparameter chưa?
* Tại sao lại dùng ROC-AUC thay vì PR-AUC trong bài toán này? Khi nào nên dùng PR-AUC?
* **📌 1. Vì sao chọn đề tài này?**
* **Trả lời:**  
  Dù Việt Nam không nằm trong vành đai lửa, nhưng vẫn ghi nhận nhiều trận động đất như tại Điện Biên, Kon Tum,... Đặc biệt, các rung chấn từ động đất nước ngoài vẫn lan tới các đô thị lớn như TP.HCM, Hà Nội. Do đó, em nhận thấy cần thiết phải có hệ thống phát hiện sớm, ứng dụng công nghệ rẻ – như cảm biến gia tốc – để xây dựng mạng cảnh báo động đất phù hợp với điều kiện Việt Nam.
* **📌 2. Cơ sở lý thuyết: cảm biến và sóng địa chấn**
* **Trả lời:**  
  Cảm biến gia tốc đo dao động theo 3 trục, chuyển đổi tín hiệu cơ học thành tín hiệu điện dựa trên hiệu ứng áp điện. Động đất thường sinh ra các sóng P (nhanh, nhẹ) và sóng S (mạnh, gây hại), trong khi nhiễu nhân sinh thường có đặc điểm tần số và biên độ khác biệt. Mô hình học máy giúp phân biệt hiệu quả hai loại tín hiệu này.
* **📌 3. Về dữ liệu**
* **Trả lời:**  
  Em chọn bộ dữ liệu K-NET của Nhật Bản vì đây là hệ thống đo gia tốc chuẩn quốc tế, với hơn 2000 trận động đất từ năm 1996–2021. Dữ liệu được cung cấp bởi trạm quan trắc nên đã được gán nhãn chính xác. Ngoài ra, em bổ sung tập dữ liệu từ Italy để tăng tính tổng quát.
* **📌 4. Về tiền xử lý dữ liệu**
* **Trả lời:**  
  Sliding window 10 giây là lựa chọn hợp lý vì ngắn hơn sẽ thiếu dữ liệu, dài hơn sẽ giảm độ chính xác thời gian phản ứng.  
  Em sử dụng IQR để loại bỏ outlier vì nó không bị ảnh hưởng bởi phân phối dữ liệu, giúp loại bỏ nhiễu hiệu quả mà không làm mất tín hiệu quan trọng.  
  Chuẩn hóa dữ liệu bằng Z-score giúp tăng tốc độ hội tụ và đảm bảo các mô hình như SVM, Logistic hoạt động ổn định.
* **📌 5. Về đặc trưng**
* **Trả lời:**  
  Em trích xuất các đặc trưng phổ biến như IQR, Energy, Kurtosis, ZCR,… Trong đó, Energy và PSD là đặc trưng mạnh nhất giúp mô hình phân biệt động đất thật và nhiễu.  
  Em chưa áp dụng giảm chiều PCA vì số lượng đặc trưng không lớn (20–30 đặc trưng). Tuy nhiên, có áp dụng SelectKBest để chọn ra top 15 đặc trưng có ý nghĩa nhất.
* **📌 6. Về mô hình**
* **Trả lời:**  
  Random Forest cho kết quả tốt nhất vì có khả năng tổng quát hóa tốt, kháng nhiễu, không cần chuẩn hóa gắt gao.  
  Mô hình không overfit do em áp dụng kỹ thuật cross-validation (K-Fold) và kiểm tra bằng ROC-AUC, các chỉ số test ổn định.  
  Nếu triển khai real-time, em ưu tiên SVM hoặc Logistic do tốc độ dự đoán nhanh, nhẹ tài nguyên. Tuy nhiên, CRNN vẫn là mục tiêu dài hạn nếu có GPU.
* **📌 7. Về đánh giá mô hình**
* **Trả lời:**  
  Precision phản ánh độ chính xác cảnh báo động đất. Recall phản ánh khả năng không bỏ sót. Trong hệ thống cảnh báo, em ưu tiên Recall cao để không bỏ sót cảnh báo thật.  
  Em có dùng kỹ thuật lấy mẫu cân bằng để giảm mất cân bằng nhãn EQ/Noise.  
  ROC-AUC giúp đánh giá mô hình ở nhiều ngưỡng khác nhau, toàn diện hơn so với chỉ một điểm cắt.
* **📌 8. Thực tiễn và ứng dụng**
* **Trả lời:**  
  Mô hình có thể tích hợp vào smartphone nếu sử dụng cảm biến gia tốc có sẵn trên thiết bị. Tuy nhiên, cần đơn giản hóa mô hình và tối ưu năng lượng.  
  Về triển khai toàn quốc: cần mạng lưới cảm biến MEMS (như Node-RED, ESP32) và trạm trung tâm xử lý tín hiệu.  
  Trong tương lai, em dự định mở rộng mô hình đa đầu ra: vừa phân loại vừa dự đoán tọa độ và độ lớn trận động đất.
* **📌 9. Về giới hạn đề tài**
* **Trả lời:**  
  Hạn chế lớn nhất là chưa xử lý real-time và chưa có mạng lưới cảm biến triển khai thực tế. Ngoài ra, mô hình hiện tại chỉ dừng ở mức phân loại EQ/Noise, chưa ước lượng được tọa độ và magnitude.  
  Tuy nhiên, kết quả khả quan từ offline test là bước đầu để phát triển hệ thống real-time sau này.
* **📌 10. Phản biện nâng cao**
* **Trả lời:**  
  PSD của EQ thường tập trung trong dải 1–10Hz, biên độ cao, dạng "gò đồi". Nhiễu thường tần số cao, không có hình dạng đặc trưng → có thể nhận diện bằng mô hình.  
  Em chọn Random Forest với 100 cây (n\_estimators=100), đã thử tuning bằng GridSearchCV.  
  ROC-AUC được ưu tiên vì bài toán cần đánh giá tổng thể. Tuy nhiên, nếu class imbalance nghiêm trọng hơn, em sẽ chuyển sang dùng PR-AUC.