ÁP DỤNG THUẬT TOÁN PHÂN LOẠI RANDOM FOREST ĐỂ DỰ BÁO SỰ CỐ CỦA ĐỘNG CƠ DIESEL DO BỊ MÒN XÉC MĂNG

APPLYING RANDOM FOREST CLASSIFICATION ALGORITHM TO PREDICT FAUL OF THE MARINE DIESEL ENGINES DUE TO WEAR OF PISTON RINGS

TRẦN HỒNG HÀ*, ĐỖ THỊ HIỀN

Khoa Máy tàu biển, Trường Đại học Hàng hải Việt Nam

**Email liên hệ: tranhongha@vimaru.edu.vn*

Tóm tắt

Độ mòn của xéc măng trong động cơ diesel là một yếu tố quan trọng ảnh hưởng đến hiệu suất và tuổi thọ của động cơ. Nghiên cứu này đề xuất một phương pháp sử dụng thuật toán Random Forest (RF) để dự báo độ mòn của xéc măng dựa trên các thông số vận hành của động cơ, như áp suất buồng đốt, nhiệt độ khí xả, lượng nhiên liệu phun, và độ rung động. Tập dữ liệu được thu thập từ các động cơ hoạt động trong nhiều điều kiện khác nhau và được phân tích để xác định mối liên hệ giữa thông số vận hành và độ mòn. Kết quả cho thấy mô hình RF đạt độ chính xác cao, hỗ trợ hiệu quả cho việc bảo trì dự đoán và tối ưu hóa hiệu suất động cơ diesel.

Từ khóa: *Động cơ diesel, thuật toán phân loại, xéc măng, mòn, thông số vận hành.*

Abstract

The wear of piston rings in diesel engines is a critical factor affecting engine performance and longevity. This study proposes a method utilizing the Random Forest (RF) algorithm to predict piston ring wear based on engine operating parameters, such as combustion chamber pressure, exhaust gas temperature, fuel injection rate, and torsional vibration. The dataset was collected from engines operating under various conditions and analyzed to identify the relationship between operating parameters and wear. The results demonstrate that the RF model achieves high accuracy, effectively supporting predictive maintenance and optimizing diesel engine performance.

Keywords: *Diesel engine, Random Forest, piston ring, wear, operating parameters.*

1. Mở đầu

Độ mòn của xéc măng ảnh hưởng trực tiếp đến quá trình đốt cháy, tiêu hao nhiên liệu, và lượng khí thải của động cơ. Dự báo chính xác độ mòn của xéc măng giúp tối ưu hóa kế hoạch bảo trì và giảm chi phí vận hành. RF kết hợp dự đoán từ nhiều cây quyết định khác nhau thông qua phương pháp lấy mẫu tổng hợp (bagging) và bỏ phiếu (voting), làm giảm nguy cơ dự đoán sai do bị quá khớp (overfitting) từ một cây đơn lẻ. Điều này giúp RF ổn định hơn các thuật toán như cây quyết định, vốn dễ bị trùng lặp khi dữ liệu huấn luyện có nhiễu hoặc các biến không đồng nhất. RF cung cấp thông tin về mức độ quan trọng của từng đặc trưng, cho phép hiểu rõ hơn những yếu tố nào ảnh hưởng lớn nhất đến sự cố.

Một số nghiên cứu đã sử dụng thuật toán RF để dự báo nồng độ khí xả của động cơ diesel như mô hình của Karunamurthy [1] được đề xuất sử dụng thuật toán Random Forest Regressor và được huấn luyện với 324 dữ liệu thực nghiệm thu thập từ các thử nghiệm thực tế. Mô hình được đánh giá bằng chỉ số R^2, đạt giá trị 0,997 với tập dữ liệu được chia theo tỷ lệ 85:15 để huấn luyện và kiểm tra. Các đầu ra của mô hình được sử dụng để tính toán dữ liệu đầu ra cho bất kỳ giá trị mới nào của các thông số đầu vào. Các giá trị tối ưu của các thông số đầu vào để đạt hiệu suất nhiệt cao nhất và khí thải thấp nhất được tìm thấy bằng phương pháp tối ưu hóa Lagrangian. Các giá trị tối ưu là mô-men xoắn 12,48Nm, lưu lượng khí sinh học 8,29 lít/phút, hàm lượng methane 72,8% và nhiệt độ khí nạp 68,3°C. Trong nghiên cứu của Viana [2] dự báo động cơ diesel bằng cách đánh giá giá trị tuyệt đối của mức độ nghiêm trọng của sự cố, sử dụng các mô hình thuật toán phân loại (RF) và mạng nơ-ron đa lớp. Một cơ sở dữ liệu đã được xây dựng với 3500 kịch bản hỏng hóc nhằm khắc phục vấn đề gây ra các hỏng hóc phá hủy trong động cơ diesel. Các tín hiệu hỏng hóc của động cơ diesel được phát triển bằng mô hình nhiệt động lực học không gian hai chiều bên trong xi lanh, kết hợp với mô hình rung xoắn trục khuỷu. Các mạng nơ ron nhân tạo và mô hình hồi quy phân loại ngẫu nhiên đã được sử dụng để phân loại và định lượng các lỗi. Phương pháp này được áp dụng cùng với một mô hình mô phỏng động cơ để đánh giá hiệu quả và độ chính xác. Kết quả hiệu suất phù hợp nhất đạt được với mô hình hồi quy rừng ngẫu nhiên với giá trị RMSE là 0,10 ± 0,03%.

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

Nghiên cứu này nhằm phát triển một mô hình dự báo độ mòn của xéc măng dựa trên các thông số vận hành của động cơ diesel, sử dụng RF để đạt độ chính xác cao và khả năng dự báo tốt.

2. Thuật toán được sử dụng để huấn luyện dữ liệu

2.1. Thu thập dữ liệu và tiền xử lý dữ liệu

Dữ liệu được thu thập từ các động cơ diesel 4 kỳ, loại Yanmar TF 120M có 4 xy lanh hoạt động trong các điều kiện thực tế. Các thông số đo lường được thể hiện trên Bảng 1.

Bảng 1. Các thông số làm việc của động cơ diesel

Cylinder Pressure (kgf/cm^2)	Cylinder Temperature (^{0}C)	Fuel Injection Rate (9/s)	Torsional Vibration (mm/s)	Wear Level
104.9671415	220.6118918	57.80920356	0.86001783	0
98.61735699	176.8929047	49.52885844	1.054782276	0
106.4768854	211.5087438	43.35232201	0.987923768	0
115.2302986	187.6152308	43.05680934	0.971673624	0
97.65846625	193.4519439	48.28674609	1.076756922	1
97.65863043	200.9520853	42.09740114	0.956533127	0
115.7921282	197.610179	52.93574034	1.045773319	1
107.6743473	166.4778872	50.99086249	0.950772481	0
95.30525614	227.6362404	50.24759601	1.025027888	0
105.4256004	222.6881844	46.90318649	0.920599359	0
95.36582307	197.0736701	45.2755119	0.965567407	0
95.34270246	218.6879725	45.3365263	1.091710751	1
102.4196227	192.0460968	50.88418139	1.0414	0
80.86719755	180.3560511	51.01859193	1.046773862	1
82.75082167	201.0882603	48.24906934	1.152537215	1
94.37712471	176.6219428	43.18921534	0.990094717	0
89.8716888	212.1810125	53.7373227	0.864643616	0

Bộ dữ liệu được thu thập qua 10,000 lần đo thực hiện trên động cơ với các thông số như: Áp suất cháy, nhiệt độ khí xả, lượng nhiên liệu cấp vào động cơ và độ rung động của động cơ chạy ở 85% tải. Sự cố về độ mòn được xác định khi thay các xéc măng bị mòn quá giới hạn cho phép và tiến hành thu thập dữ liệu trong trường hợp động cơ làm việc với các xéc măng bị mòn. Dữ liệu sau khi thu thập sẽ được xử lý qua các bước sau:

Bước 1: Xử lý dữ liệu thiếu hoặc không đầy đủ:

Dữ liệu thực tế thường có giá trị bị thiếu hoặc không hợp lệ. Nếu không xử lý, các thuật toán sẽ gặp lỗi hoặc tạo ra dự đoán không chính xác.

Bước 2: Loại bỏ dữ liệu nhiễu:

Các giá trị ngoại lai (outliers) có thể ảnh hưởng đến hiệu suất của mô hình. Mô hình sử dụng Z-score để phát hiện và xử lý giá trị ngoại lai.

Bước 3: Chuẩn hóa và cân bằng dữ liệu:

Random Forest không yêu cầu dữ liệu được chuẩn hóa, nhưng đối với các đặc trưng có khoảng giá trị quá khác biệt, việc chuẩn hóa có thể cải thiện hiệu suất.

Cân bằng dữ liệu giữa các lớp (ví dụ, số lượng mẫu giữa các trạng thái mòn "Normal", và "Wear") giúp mô hình không bị lệch về một lớp cụ thể.

Bước 4: Chia dữ liệu hợp lý:

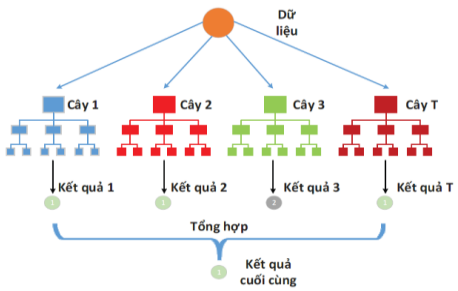
Chia dữ liệu thành tập huấn luyện và kiểm tra (70:30) giúp đánh giá mô hình chính xác.

2.2. Thuật toán của Random Forest

Random Forest (RF) được đánh giá cao trong việc dự báo sự cố, đặc biệt là trong các hệ thống phức tạp như động cơ diesel, vì các lý do sau: RF là một thuật toán dựa trên việc kết hợp nhiều cây quyết định. Mỗi cây học từ một phần dữ liệu khác nhau (kỹ thuật lấy mẫu ngẫu nhiên có hoàn lại bootstrap sampling), giúp giảm thiểu độ lệch và độ phân tán. So với các cây quyết định đơn lẻ, RF có khả năng giảm trùng lặp nhờ vào việc tổng hợp dự đoán từ nhiều cây.

RF có khả năng xử lý tốt các mối quan hệ phi tuyến giữa các biến đầu vào, điều này rất quan trọng trong dự báo sự cố, nơi mà các thông số hoạt động của thiết bị thường không có mối quan hệ tuyến tính. RF có thể cân nhắc ảnh hưởng của các nhãn hiếm (ví dụ: Sự cố hiếm gặp), nhờ vào khả năng sử dụng trọng số hoặc kỹ thuật oversampling. RF xử lý dữ liệu có số lượng lớn biến đầu vào một cách hiệu quả mà không yêu cầu giảm số chiều dữ liệu. Điều này phù hợp với bài toán chẩn đoán sự cố, nơi có nhiều thông số hoạt động được ghi nhận. Ngoài ra, RF có khả năng loại bỏ nhiễu trong dữ liệu vì mỗi cây chỉ học từ một phần dữ liệu. Điều này đảm bảo rằng mô hình không bị ảnh hưởng quá mức bởi các dữ liệu.

Khi tạo một cây trong RF như trong Hình 1, một tập con của dữ liệu huấn luyện ban đầu được chọn ra. Trong quá trình lấy mẫu này, một mẫu dữ liệu có thể được chọn nhiều lần, vì mỗi lần lấy đều được thực hiện có hoàn lại. Mỗi cây trong rừng được huấn luyện trên một tập dữ liệu khác nhau. Điều này đảm bảo rằng các cây sẽ không giống hệt nhau, giúp tăng khả năng khái quát hóa của RF. Nhờ sự đa dạng, mô hình có thể giảm overfitting so với việc sử dụng một cây quyết định duy nhất. Phương pháp lấy mẫu ngẫu nhiên có hoàn lại bootstrap cho phép RF khai thác toàn bộ thông tin của tập dữ liệu huấn luyện, trong khi vẫn tạo ra nhiều cây độc lập. Việc tổng hợp dự đoán từ các cây này giúp giảm sai số tổng thể.



Hình 1. Thuật toán random forest [3]

Một số mẫu có thể xuất hiện nhiều lần trong một tập mẫu ngẫu nhiên có hoàn lại bootstrap, giúp mô hình tập trung hơn vào các đặc trưng quan trọng hoặc xử lý tốt dữ liệu hiếm gặp. Xây dựng từng cây quyết định: Với mỗi tập mẫu ngẫu nhiên có hoàn lại bootstrap khởi tạo một cây quyết định. Nguyên tắc lựa chọn biến đầu vào tại mỗi nút phân chia:

- Thay vì sử dụng toàn bộ các đặc trưng, chỉ chọn ngẫu nhiên m đặc trưng từ tập dữ liệu (với $m <$ tổng số đặc trưng).

- Điều này làm cho các cây trở nên độc lập hơn, tránh overfitting và tăng tính đa dạng.

Xác định cách chia (split) tốt nhất dựa trên các đặc trưng được chọn ngẫu nhiên. Tiếp tục phân chia đến khi đạt điều kiện dừng (chẳng hạn số mẫu nhỏ hơn ngưỡng hoặc đạt độ sâu tối đa). Trong RF mức độ quan trọng của các đặc trưng phụ thuộc vào độ không thuần khiết của đặc trưng đó giảm hay tăng khi nó được chọn làm đặc trưng phân tách. Mức độ quan trọng của một đặc trưng X_j tính theo công thức sau [3, 4]:

$$X_j = \frac{1}{B}\sum_{b=1}^{B}\sum_{t\epsilon\ nodes}\Delta I(t) \qquad (1)$$

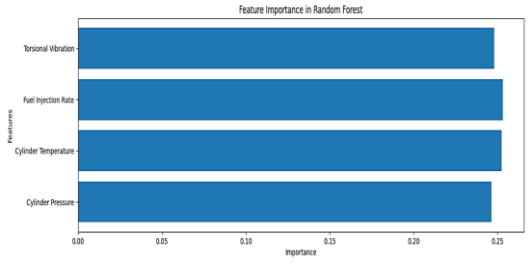
Trong đó:

- *ΔI(t):* mức giảm độ không thuần khiết tại nút *t* khi sử dụng đặc trưng *Xj*;

- *B:* tổng số cây trong rừng;

- *Nodes:* tập hợp các nút trong cây *b* nơi X_j được sử dụng.

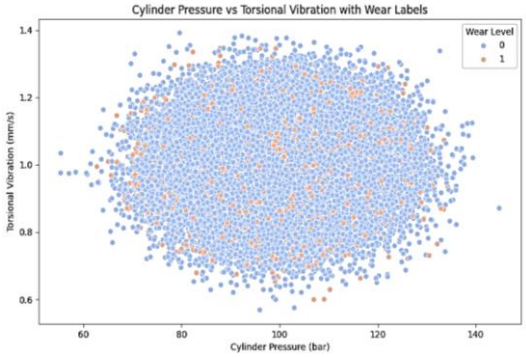
Trong mô hình đặc tính quan trọng của các dữ liệu Feature Importance là thước đo định lượng mức độ đóng góp của từng đặc tính vào việc dự đoán mục tiêu trong một mô hình Random Forest.



Hình 2. Đặc tính quan trọng của các dữ liệu huấn luyện

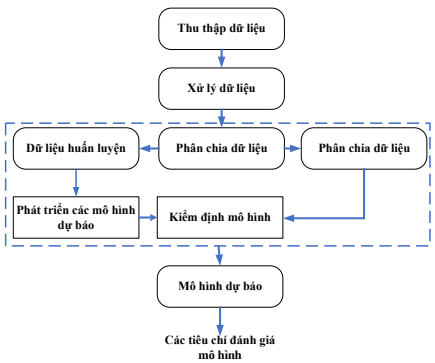
Biểu đồ trong Hình 2 hiển thị giá trị Feature Importance của từng đặc tính. Những đặc tính quan trọng (Feature Importance cao) có ảnh hưởng lớn đến độ chính xác của dự báo. Trong đó độ quan trọng của các thông số: Tortional vibration là 0,245; fuel injection rate là 0,252; Cylinder temperature là 0,253; Cylinder pressure là 0,242. Chỉ sử dụng các đặc tính quan trọng nhất để giảm thời gian huấn luyện và tăng hiệu suất. Phân tích sâu hơn về mối quan hệ giữa các đặc tính quan trọng để hiểu rõ nguyên nhân gây mòn. Việc xác định và sử dụng đặc tính quan trọng đúng cách giúp Random Forest đạt độ chính xác cao hơn trong dự đoán độ mòn xéc măng và sơ mi động cơ diesel.

Biểu đồ phân tán như trong Hình 3 hiển thị và phân tích mối quan hệ giữa hai biến số. Trục X đại diện cho một biến độc lập là áp suất cháy của động cơ. Trục Y đại diện cho một biến độc lập là độ rung. Mỗi điểm trên biểu đồ biểu diễn trạng thái của xéc măng là bình thường hoặc bị mòn. Scatter plot giúp xác định xem có mối quan hệ giữa hai biến không, mối quan hệ đó là phi tuyến tính (non-linear).



Hình 3. Biểu đồ phân tán về mối quan hệ giữa áp suất cháy, độ rung với độ mòn của xéc măng

Lưu đồ thuật toán của Random Forest như trong Hình 4 là một thuật toán học máy có giám sát, cực kỳ phổ biến và được sử dụng cho các bài toán phân loại và hồi quy trong học máy. Một khu rừng bao gồm rất nhiều cây, và số lượng cây càng nhiều thì khu rừng càng bền vững. Tương tự, số lượng cây trong thuật toán Random Forest càng lớn thì độ chính xác và khả năng giải quyết vấn đề của nó càng cao.



Hình 4. Lưu đồ thuật toán Random Forest [5]

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

Lưu đồ thuật toán Random Forest, được mô tả một cách tổng quát như sau:

Bước 1: Nhập dữ liệu đầu vào (dữ liệu huấn luyện và nhãn).

Bước 2: Tiền xử lý dữ liệu:

- Xử lý các giá trị bị thiếu;
- Mã hóa dữ liệu (nếu cần);
- Chia dữ liệu thành tập huấn luyện và tập kiểm tra.

Bước 3: Tạo các cây quyết định.

Chọn số lượng cây (n_estimators). Với mỗi cây lấy mẫu dữ liệu sử dụng phương pháp Bootstrap Sampling để chọn ngẫu nhiên dữ liệu (có hoàn lại). sau đó chọn tập con đặc trưng: Chọn ngẫu nhiên một số đặc trưng (feature) để xây dựng cây.

Bước 4: Xây dựng cây quyết định: Phân tách tại mỗi nút dựa trên tiêu chí (như Gini Impurity, Entropy).

Bước 5: Đánh giá mô hình bằng cách sử dụng tập kiểm tra để đánh giá độ chính xác (Accuracy, MSE). Tinh chỉnh siêu tham số (Hyperparameters) nếu cần (như số lượng cây, độ sâu cây).

Bước 6: Xuất kết quả dự đoán và mô hình.

Để đánh giá độ chính xác, mô hình sử dụng sai số bình phương trung bình Root Mean Square Error (RMSE) là một chỉ số phổ biến dùng để đo lường mức độ chính xác của các mô hình dự đoán trong các bài toán hồi quy. RMSE biểu thị độ lệch trung bình chuẩn giữa các giá trị thực tế và giá trị dự đoán của mô hình [6].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Trong đó:

n: Số lượng mẫu trong tập dữ liệu;

y_i: Giá trị thực tế của mẫu thứ i;

$\hat{y}_i$: Giá trị dự đoán của mẫu thứ i.

RMSE càng nhỏ thì mô hình dự đoán càng chính xác, tức là sai số giữa dự đoán và thực tế càng thấp. RMSE là một lựa chọn tốt khi mô hình hóa dữ liệu sao cho các giá trị dự đoán sát với thực tế. RMSE đo lường sai số của mô hình hồi quy, nhấn mạnh vào các lỗi lớn. Một chỉ số RMSE thấp hơn thường cho thấy mô hình tốt hơn.

3. Huấn luyện mô hình

Đầu vào: Tập dữ liệu D có kích thước 100,000 dữ liệu, chứa các giá trị của các biến dự đoán [Áp suất cháy trong xy lanh (Cylinder pressure), Nhiệt độ khí xả (Cylinder temperature), Lượng nhiên liệu (Fuel injection rate), Mức độ dao động (Torsion vibration)].

Đầu ra: Các giá trị được dự đoán cho các biến phản hồi dựa trên các đầu vào từ các biến dự đoán.

Phương pháp:

Tạo một random forest theo các bước sau:

Tạo một mẫu bootstrap *D'* từ tập dữ liệu D.

Xây dựng một cây quyết định (decision tree) cho mẫu bootstrap. Các bước (a) và (b) được lặp lại 'n' lần, trong đó 'n' là số lượng cây con yêu cầu hoặc có thể tạo ra. Tính toán giá trị dự đoán từ mỗi cây. Giá trị trung bình của các giá trị dự đoán đại diện cho đầu ra.

Khi thay đổi giá trị max_depth một tham số quan trọng trong mô hình huấn luyện từ 10 đến 40 như trong Bảng 2, các giá trị khác giữ nguyên không đổi. Nó quy định độ sâu tối đa của mỗi cây quyết định trong mô hình. Độ sâu của cây quyết định là số mức phân nhánh từ gốc đến lá cây có độ sâu nhất. Max_depth giới hạn số mức này để kiểm soát độ phức tạp của cây.

Ảnh hưởng của max_depth: Nếu max_depth nhỏ (giới hạn thấp): Cây sẽ nông, không đủ khả năng

Bảng 2. Ảnh hưởng của giá trị Max-depth tới độ chính xác của mô hình

n-estimator	Max-depth	Test_size	Precision		RMSE
100	10	0.2	0	0.9	0.49
			1	0.1	
100	15	0.2	0	0.9	0.39
			1	0.09	
100	20	0.2	0	0.9	0.34
			1	0.11	
100	25	0.2	0	0.9	0.32
			1	0.15	
100	30	0.2	0	0.9	0.32
			1	0.19	

Bảng 3. Ảnh hưởng của giá trị Test-size và n-estimator tới độ chính xác của mô hình

n-estimator	Max-depth	Test_size	Precision		RMSE
100	10	0.2	0	0.9	0.49
			1	0.1	
100	10	0.3	0	0.9	0.49
			1	0.11	
100	10	0.4	0	0.9	0.49
			1	0.1	
150	10	0.2	0	0.9	0.49
			1	0.1	
200	10	0.2	0	0.9	0.48
			1	0.1	
250	10	0.2	0	0.9	0.48
			1	0.1	

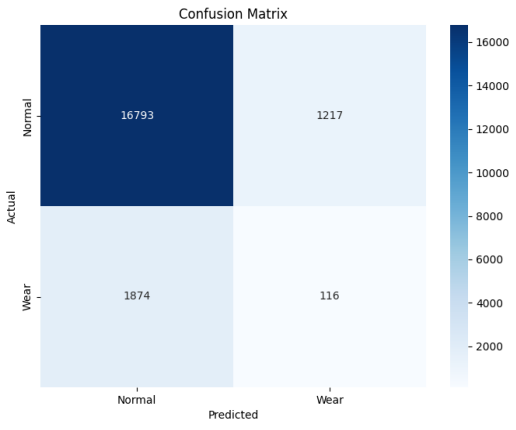
biểu diễn dữ liệu phức tạp. Mô hình có thể bị underfitting (không học đủ thông tin từ dữ liệu). Nếu max_depth lớn (giới hạn cao hoặc không giới hạn): Cây sẽ rất sâu, dẫn đến việc mô hình học quá nhiều chi tiết và nhiễu trong dữ liệu. Mô hình có thể bị overfitting (quá khớp với dữ liệu huấn luyện, kém hiệu quả trên dữ liệu kiểm tra). Chọn giá trị phù hợp cho max_depth: Không giới hạn (None hoặc max_depth rất lớn): Mô hình tự do phát triển cây đến khi tất cả các lá đều "thuần" (pure), hoặc không còn dữ liệu để phân chia.

Thay đổi n_estimators từ 100 đến 250 như trong Bảng 3 cho thấy số cây càng lớn, mô hình càng mạnh trong việc giảm thiểu phương sai và cho ra kết quả ổn định hơn, nhưng thời gian huấn luyện cũng tăng lên. Nếu chọn giá trị n_estimators thấp thì mô hình có thể không đủ phức tạp để học dữ liệu. Nếu chọn giá trị n_estimators cao thì mô hình có thể overfitting hoặc mất quá nhiều thời gian để huấn luyện.cho thấy là một tham số quan trọng trong các thuật toán học máy liên quan đến mô hình tập hợp. Nó quy định số lượng mô hình con (hoặc cây quyết định) được sử dụng trong mô hình tổng hợp và số lượng cây quyết định trong rừng.

Trong mô hình huấn luyện, test_size là một tham số trong các hàm phân chia dữ liệu, thường được sử dụng trong train-test split, như trong hàm train_test_split từ thư viện scikit-learn. Nó xác định tỷ lệ hoặc số lượng mẫu của tập dữ liệu sẽ được sử dụng làm tập kiểm tra (test set), phần còn lại sẽ được dùng làm tập huấn luyện (train set). Nó biểu thị tỷ lệ phần trăm của dữ liệu dành cho tập kiểm tra. Đánh giá mô hình: Giúp tách một phần dữ liệu để kiểm tra xem mô hình hoạt động tốt thế nào trên dữ liệu chưa từng thấy. Cân bằng dữ liệu: Một tỷ lệ hợp lý (ví dụ: 70-30, 80-20) giữa tập huấn luyện và tập kiểm tra giúp đảm bảo rằng mô hình có đủ dữ liệu để học và cũng có đủ dữ liệu để kiểm tra hiệu năng.

Các thông số được chọn Test-size=0,2 và n-estimator=100 và Max-depth=15 cho độ chính xác đối với trường hợp bình thường là 0,9 và bị mòn là 0,09, sai số RMSE=0,39 được chấp nhận với mô hình huấn luyện.

Để cung cấp cái nhìn trực quan hơn về việc xác định lỗi, Hình 5 cho thấy ma trận nhầm lẫn (confusion matrix) của mô hình phân loại với hai lớp: Normal (Bình thường) và Wear (Mòn). Ma trận cho biết mối quan hệ giữa các giá trị thực tế và dự đoán của mô hình.



Hình 5. Ma trận nhầm lẫn

Số liệu trong ma trận:

- True Negative (TN): 16,793 (Thực tế là "Normal" và được dự đoán đúng là "Normal").

TẠP CHÍ ISSN: 1859-316X
KHOA HỌC CÔNG NGHỆ HÀNG HẢI
JOURNAL OF MARINE SCIENCE AND TECHNOLOGY

- False Positive (FP): 1,217 (Thực tế là "Normal" nhưng bị dự đoán nhầm thành "Wear").

- False Negative (FN): 1,874 (Thực tế là "Wear" nhưng bị dự đoán nhầm thành "Normal").

- True Positive (TP): 116 (Thực tế là "Wear" và được dự đoán đúng là "Wear").

- Tổng số mẫu = TN + FP + FN + TP = 16,793 + 1,217 + 1,874 + 116 = 20,000.

Với độ chính xác của mô hình là 84,5%. Độ chính xác cao của mô hình giúp giảm thiểu các trường hợp bỏ sót và phân loại nhầm.

4. Kết quả dự đoán và thảo luận

Sau khi huấn luyện, mô hình được kiểm tra lại độ chính xác của mô hình khi chuẩn đoán các trường hợp sự cố của động cơ diesel ở các trường hợp khai thác khác nhau. Bộ dữ liệu được đo ở chế độ động cơ chạy ở 85% tải để đánh giá độ chính xác của mô hình như trong Hình 6.

```
# Dự đoán trên dữ liệu mới
num_samples = 100000
new_data = pd.DataFrame({
    'Cylinder Pressure': np.random.uniform(80, 160, num_samples),
    'Cylinder Temperature': np.random.uniform(120, 217, num_sample
    'Fuel Injection Rate': np.random.uniform(30, 54, num_samples),
    'Torsional Vibration': np.random.uniform(0.7, 1.7, num_samples
    })
df = pd.DataFrame(new_data)
print(df.head())
predictions = model.predict(new_data)
print("Predicted Wear Levels:", predictions)
```

Hình 6. Code tạo bộ dữ liệu mới của động cơ diesel

```
   Cylinder Pressure  ...  Torsional Vibration
0         159.048393  ...             1.667229
1         133.146027  ...             1.211981
2          97.939849  ...             0.741356
3          95.555650  ...             1.400093
4         132.952690  ...             1.115416

[5 rows x 4 columns]
Predicted Wear Levels: [0 0 0 ... 0 1 0]
```

```
              precision    recall  f1-score   support

           0       0.90      0.93      0.91     18009
           1       0.10      0.07      0.08      1991

    accuracy                           0.84     20000
   macro avg       0.50      0.50      0.50     20000
weighted avg       0.82      0.84      0.83     20000

Root Mean Square Error (RMSE): 0.40
```

Hình 7. Kết quả dự đoán đối với tập dữ liệu mới

Kết quả kiểm tra cho thấy mô hình chuẩn đoán được độ chính xác là 84% như trong Hình 7. Tỷ lệ dự đoán đúng tổng thể khá cao (16,800/20,000 mẫu đúng). Tuy nhiên, điều này bị ảnh hưởng bởi sự mất cân bằng lớp (lớp 0 chiếm ưu thế lớn hơn lớp 1). Precision, Recall, và F1 đều là 0,50, chỉ ra rằng mô hình việc xử lý lớp ít xuất hiện (lớp 1) nhỏ hơn do thu thập về dữ liệu các trường hợp động cơ bị sự cố khi xéc măng bị ăn mòn quá giới hạn cho phép chưa đủ nhiều. RMSE phản ánh sai số trung bình của mô hình. Với giá trị 0,40, mức sai số này có thể chấp nhận được đối với mô hình dự đoán.

5. Kết luận

Kết quả được trình bày trong bài báo này chứng minh khả năng áp dụng cho một số tập dữ liệu và mô hình có thể ánh xạ tương tự trong không gian đặc trưng, từ đó mang lại hiệu suất tốt cho cả phân loại/chẩn đoán và dự đoán. Các kỹ thuật được áp dụng ở đây cho động cơ diesel cũng có thể được áp dụng cho các hệ thống kỹ thuật phức tạp khác, nơi dữ liệu vận hành và cảm biến được thu thập bởi các hệ thống điều khiển và giám sát.

Mô hình RF đạt độ chính xác cao độ chính xác tới 84%. trong việc phân loại các mức độ mòn của xéc măng và sơ mi xy lanh. Thuật toán RF không chỉ ổn định mà còn có khả năng xác định mức độ quan trọng của từng đặc trưng, giúp các nhà nghiên cứu hiểu rõ hơn về tác nhân chính gây mòn. Mô hình RF có tiềm năng áp dụng trong thực tế để giám sát và bảo dưỡng động cơ diesel, giúp dự đoán sự cố trước khi chúng xảy ra, giảm thời gian chết máy và chi phí bảo trì.

Nghiên cứu trong tương lai sẽ tập trung vào việc điều tra khả năng thu thập dữ liệu đầy đủ hơn, thông qua các quy trình thu thập dữ liệu được cải thiện hoặc làm sạch dữ liệu tốt hơn, để tạo ra các bộ phân loại học máy có tính giải thích được, thực tế và hữu ích cho việc dự đoán lỗi trên các động cơ diesel lớn.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Hàng hải Việt Nam Đề tài mã số: **DT24-25.20**.

TÀI LIỆU THAM KHẢO

[1] Krishnasamy Karunamurthy, Mohammed Musthafa Feroskhan1, Ganesan Suganya, and Ismail Saleel (2022), *Prediction and optimization of performance and emission characteristics of a dual fuel engine using machine learning*, International Journal for Simulation and Multidisciplinary Design Optimization (IJSMDO), Vol.13.
https://doi.org/10.1051/smdo/2022002.

[2] Denys P. Viana 1ORCID, Dionísio H. C. Martins (2023), *Diesel Engine Fault Prediction Using Artificial Intelligence Regression Methods*, Machines, Vol.11(5).
https://doi.org/10.3390/machines11050530.

[3] Đỗ Quang Hưng (2024), *Dự báo hoạt động ngân hàng bằng thuật toán rừng ngẫu nhiên*. Tạp chí Kinh Tế và Phát triển, Số 320, tr.64-78.

[4] Parlak, A., et al. (2006), *Application of artificial neural network to predict specific fuel consumption and exhaust temperature for a Diesel engine*. Applied Thermal Engineering, Vol.26(8-9): pp.824-828.
https://doi.org/10.1016/j.applthermaleng.2005.10.006.

[5] Fang, X., et al. (2021), *On the application of artificial neural networks for the prediction of NOx emissions from a high-speed direct injection diesel engine*. International Journal of Engine Research. Vol.22(6).
https://doi.org/10.1177/1468087420929768.

[6] Roy, S., R. Banerjee, and P.K. Bose (2014), *Performance and exhaust emissions prediction of a CRDI assisted single cylinder diesel engine coupled with EGR using artificial neural network*. Applied Energy, Vol.119, pp.330-340.

Ngày nhận bài:	05/02/2025
Ngày nhận bản sửa:	03/03/2025
Ngày duyệt đăng:	05/03/2025