NGHIÊN CỨU ỨNG DỤNG HỌC MÁY VÀ XAI TRONG PHÂN TÍCH DỮ LIỆU CHẤT LƯỢNG ĐỂ CẢI THIỆN QUY TRÌNH SẢN XUẤT

RESEARCH ON THE APPLICATION OF MACHINE LEARNING AND EXPLAINABLE AI IN QUALITY DATA ANALYSIS FOR MANUFACTURING PROCESS IMPROVEMENT

Phạm Minh Ngọc[1,*], Nguyễn Thành Công[2]
[1]Trường Đại học Hàng hải Việt Nam
[2]Trường Cao đẳng Hàng hải và Đường thủy I
**Email: ngocpm.mtb@vimaru.edu.vn*

TÓM TẮT

Nghiên cứu này trình bày phương pháp phát hiện và phân tích lỗi bề mặt thép dựa trên học máy (ML) kết hợp với giải thích mô hình (XAI). Bộ dữ liệu Steel Plate Defects được sử dụng để huấn luyện đồng thời năm mô hình ML gồm Logistic Regression, Decision Tree, Random Forest, XGBoost và LightGBM. Kết quả cho thấy các mô hình boosting (XGBoost, LightGBM) đạt hiệu năng vượt trội với F1-score và AUC cao ở hầu hết các loại lỗi. Bên cạnh đó, SHAP và LIME được tích hợp nhằm giải thích cơ chế dự đoán, giúp xác định các đặc trưng quan trọng tác động đến từng loại lỗi. Hệ thống được đề xuất góp phần nâng cao tính minh bạch và hiệu quả trong kiểm soát chất lượng bề mặt thép theo định hướng Công nghiệp 4.0.

Từ khóa: *Học máy; XAI; Lỗi sản xuất; LightGBM; XGBoost; SHAP; LIME; Kiểm soát chất lượng.*

ABSTRACT

This study proposes a machine-learning-based approach for detecting and analyzing steel surface defects combined with explainable artificial intelligence (XAI). The Steel Plate Defects dataset is used to train five ML models: Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM. Experimental results show that boosting models (XGBoost and LightGBM) achieve superior performance with high F1-scores and AUC values across most defect types. SHAP and LIME are applied to interpret model decisions, revealing key features associated with each defect category. The proposed framework enhances transparency and supports intelligent quality control for steel surfaces in Industry 4.0 environments.

Keywords: *Machine learning; XAI; Industrial Faults; LightGBM; XGBoost; SHAP; LIME; Quality control.*

1. MỞ ĐẦU

Trong bối cảnh chuyển đổi số trong sản xuất theo định hướng Công nghiệp 4.0, yêu cầu về kiểm soát chất lượng ngày càng trở nên quan trọng nhằm giảm phế phẩm, hạn chế lỗi công đoạn và duy trì ổn định quy trình. Trong các dây chuyền cán thép tấm, các khiếm khuyết bề mặt như xước, vết lõm, vết bẩn hoặc biến dạng hình học thường xuất hiện do điều kiện thiết bị, nguyên liệu hoặc sự thay đổi của thông số công nghệ. Việc phát hiện lỗi chủ yếu dựa trên kiểm tra thủ công khiến khả năng bỏ sót cao, thiếu ổn định và khó mở rộng trong sản xuất công suất lớn.

Sự phát triển của học máy (Machine Learning - ML) đã mở ra hướng tiếp cận mới cho dự đoán và phân loại lỗi bề mặt dựa trên dữ liệu. Các thuật toán như Logistic Regression, Decision Tree, Random Forest, XGBoost và LightGBM có khả năng xử lý dữ liệu đa biến và mô hình hóa tốt các mối quan hệ phi tuyến trong quá trình sản xuất. Tuy nhiên, hầu hết các mô hình ML hoạt động như "hộp đen" và không đưa ra giải thích cho dự đoán, gây khó khăn cho kỹ sư khi xác định nguyên nhân gây lỗi và triển khai cải tiến.

Để khắc phục hạn chế này, các phương pháp giải thích mô hình (Explainable AI - XAI) như SHAP và LIME được sử dụng nhằm làm rõ mức độ ảnh hưởng của từng đặc trưng đối với quyết định của mô hình. SHAP cung cấp ánh xạ toàn cục về tầm quan trọng của đặc trưng trong toàn bộ mô hình, trong khi LIME giải thích tại cấp độ từng mẫu, giúp xác định rõ đặc trưng nào đã khiến mô hình dự đoán sản phẩm là lỗi. Bài báo này tập trung ứng dụng năm mô hình ML để dự đoán bảy loại khiếm khuyết bề mặt trong bộ dữ liệu Steel Plates Faults (Dataset A). Mô hình có F1-score cao nhất cho từng loại lỗi được lựa chọn để phân tích bằng SHAP và LIME. Việc kết hợp ML và XAI không chỉ nâng cao hiệu quả phát hiện lỗi mà còn giúp hiểu rõ nguyên nhân ảnh hưởng đến chất lượng, từ đó hỗ trợ ra quyết định cải tiến trong quy trình sản xuất.

2. CƠ SỞ LÝ THUYẾT VÀ TỔNG QUAN NGHIÊN CỨU

Việc ứng dụng các phương pháp học máy trong nhận diện lỗi bề mặt đã được nghiên cứu rộng rãi trong nhiều lĩnh vực công nghiệp. Về mặt lý thuyết, các thuật toán học máy dạng phân loại (classification) được sử dụng nhằm học mô hình ánh xạ giữa bộ đặc trưng đầu vào và nhãn lỗi. Các thuật toán như Logistic Regression, Decision Tree, Random Forest, XGBoost và LightGBM đều có khả năng mô hình hóa mối quan hệ giữa các thuộc tính hình học của thép tấm và nguy cơ xuất hiện khiếm khuyết. Logistic Regression đại diện cho mô hình tuyến tính dễ diễn giải, trong khi Random Forest, XGBoost và LightGBM sử dụng cấu trúc cây quyết định để mô tả quan hệ phi tuyến, thường cho hiệu suất cao hơn trong các bài toán lỗi bề mặt có nhiều tương tác đặc trưng.

Song song với đó, khung tiếp cận XAI (Explainable AI) được xem là cần thiết khi triển khai ML trong sản xuất. Các phương pháp như SHAP (SHapley Additive exPlanations) và LIME (Local Interpretable Model-Agnostic Explanations) cho phép hiểu rõ cơ chế hoạt động của mô hình. SHAP cung cấp tầm quan trọng của đặc trưng dựa trên lý thuyết giá trị Shapley trong trò chơi hợp tác, cho phép giải thích mức độ đóng góp của từng đặc trưng trong toàn bộ mô hình. Ngược lại, LIME tập trung vào giải thích cục bộ bằng cách xây dựng mô hình tuyến tính đơn giản xung quanh một mẫu cụ thể, giúp kỹ sư hiểu vì sao mô hình dự đoán sản phẩm đó là lỗi hay không lỗi.

☞

Các nghiên cứu quốc tế gần đây đã chứng minh hiệu quả rõ rệt của các mô hình học máy trong nhận diện và phân loại lỗi. Huang et al. (2022) [1] cho thấy các thuật toán ML có thể tự động phát hiện khiếm khuyết trên bề mặt kim loại hình trụ với độ chính xác cao, nhờ khả năng học đặc trưng trực tiếp từ hình ảnh cảm biến. Trong lĩnh vực sản xuất thép tấm, Dorbane et al. (2025) [2] đề xuất khung giải pháp kết hợp giữa thuật toán học máy giải thích được và kỹ thuật xử lý mất cân bằng dữ liệu SMOTE, qua đó cải thiện đáng kể độ nhạy (recall) trong nhận diện lỗi hiếm gặp, một thách thức vốn rất phổ biến trong nhà máy cán thép. Wang et al. (2025) [3] chứng minh rằng LightGBM, khi được tối ưu bằng Bayesian Optimization và kết hợp với chiến lược resampling, có thể dự đoán chính xác sự xuất hiện của các vết nứt dọc trên slab. Mô hình cho thấy khả năng học sâu các mối quan hệ phi tuyến giữa đặc trưng quá trình và sự hình thành khuyết tật, giúp nâng cao độ tin cậy trong cảnh báo lỗi sớm. Song song với sự phát triển của các mô hình dự đoán, xu hướng ứng dụng AI giải thích được (XAI) trong sản xuất công nghiệp cũng ngày càng quan trọng. Lundberg và Lee (2017) [4] đã giới thiệu SHAP như một phương pháp giải thích thống nhất, cho phép định lượng mức độ đóng góp của từng đặc trưng vào quyết định dự đoán. Đây là nền tảng giúp tăng minh bạch cho các mô hình phức tạp như LightGBM hay XGBoost. Tiếp nối hướng này, Wang et al. (2024) [5] phát triển một mô hình học máy có khả năng giải thích để dự đoán tính chất cơ học của thép cán nóng, nghiên cứu cho thấy việc kết hợp XAI giúp xác định được các thuộc tính quan trọng nhất, từ đó hỗ trợ kỹ sư ra quyết định và tối ưu hóa quy trình sản xuất.

Tổng hợp các nghiên cứu trên cho thấy ML không chỉ có khả năng dự đoán lỗi chính xác mà còn có thể kết hợp với XAI để cung cấp minh bạch, hỗ trợ kỹ sư chất lượng hiểu rõ nguyên nhân gốc rễ và cải thiện quy trình một cách hiệu quả. Tổng quan các nghiên cứu cho thấy ML có thể mang lại độ chính xác cao trong dự đoán lỗi, trong khi XAI giúp làm rõ vai trò của từng đặc trưng và tăng cường tính minh bạch của mô hình. Tuy nhiên, mỗi loại mô hình có ưu/ nhược điểm riêng, và việc lựa chọn thuật toán phù hợp cũng như cơ chế giải thích rõ ràng là điều cần thiết khi triển khai trong thực tế.

3. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Nghiên cứu sử dụng Steel Plates Faults Dataset do Secom Data công bố, chứa dữ liệu đo từ hệ thống kiểm tra bề mặt thép công nghiệp. Dataset gồm 1941 mẫu với 27 đặc trưng hình học và 7 nhãn lỗi (Pastry, Z-Scratch, K-Scratch, Stains, Dirtiness, Bumps, Other Faults). Các đặc trưng bao gồm thông số hình dạng (perimeter, area), tọa độ biên theo trục X–Y, giá trị cực trị (min, max) và các chỉ số hình dạng phức tạp hơn như edge indices và logarithm of area.

Dữ liệu phân bố mất cân bằng mạnh, trong đó các loại lỗi như Stains và Dirtiness chiếm dưới 5% tổng số mẫu, trong khi nhóm no-fault chiếm tỷ trọng lớn. Mất cân bằng này đặt ra yêu cầu đánh giá mô hình bằng F1-score thay vì Accuracy để đảm bảo phản ánh đúng hiệu quả dự đoán lỗi.

Bảng 1. Một số các biến của Steel Plates Faults Dataset

Tên biến	Ý nghĩa
X_Minimum	Vị trí đo thấp nhất theo trục X
X_Maximum	Giá trị lớn nhất theo trục X
Length	Chiều dài tấm thép
Height	Chiều cao tấm thép
Outside Global Index	Chỉ số sai lệch tổng thể

Trước khi huấn luyện, dữ liệu được xử lý theo các bước chuẩn hóa thông dụng trong học máy. Trước hết, cột ID được loại bỏ do không liên quan đến đặc trưng kỹ thuật. Các đặc trưng liên tục được chuẩn hóa bằng StandardScaler, đưa về phân phối có trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp mô hình hội tụ nhanh và ổn định hơn. StandardScaler giúp mô hình học máy hội tụ tốt hơn, tránh việc đặc trưng có độ lớn lớn áp đảo các đặc trưng khác, đồng thời đảm bảo hiệu quả đối với các mô hình tuyến tính và khoảng cách như Logistic Regression, SVM, kNN và cả các thuật toán boosting.

Dữ liệu được biến đổi về phân phối có trung bình bằng 0 và phương sai bằng 1:

$$x' = \frac{x - \mu}{\sigma}$$

Quá trình huấn luyện tuân theo chuẩn đánh giá mô hình với dữ liệu thực tế, trong đó dataset được chia thành 70% tập huấn luyện và 30% tập kiểm tra bằng phương pháp train-test split. Chỉ tập kiểm tra được sử dụng để đánh giá mô hình nhằm tránh hiện tượng overfitting.

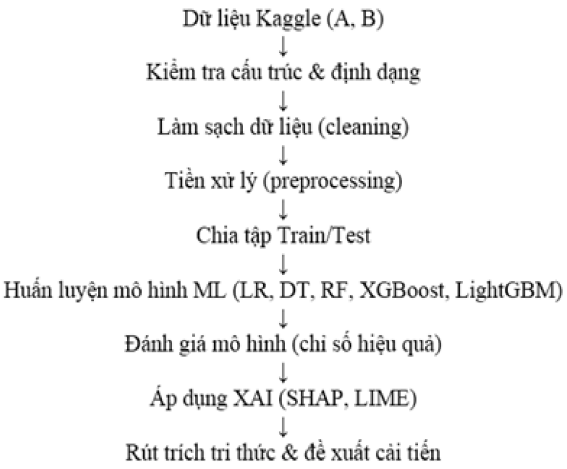
Để đảm bảo tính khách quan và tránh phụ thuộc vào một thuật toán duy nhất, nghiên cứu triển khai đồng thời năm mô hình học máy thường được sử dụng trong bài toán phân loại lỗi công nghiệp, bao gồm Logistic Regression, Decision Tree, Random Forest, XGBoost và LightGBM. Năm mô hình này đại diện cho ba nhóm thuật toán khác nhau: (i) mô hình tuyến tính, (ii) mô hình dựa trên cây quyết định và (iii) mô hình boosting hiện đại. Việc lựa chọn đa dạng thuật toán nhằm khai thác đầy đủ ưu thế của từng nhóm, đặc biệt trong bối cảnh dữ liệu lỗi bề mặt thép có quan hệ phi tuyến, tương tác thuộc tính phức tạp và phân bố mất cân bằng.

Tất cả các mô hình được huấn luyện và đánh giá trên cùng một tập kiểm tra để bảo đảm khả năng so sánh công bằng. Các thước đo hiệu năng được sử dụng bao gồm Accuracy, Precision, Recall, F1-score và AUC. Trong bối cảnh dữ liệu có mức độ mất cân bằng cao, F1-score được lựa chọn làm tiêu chí chính cho việc xác định mô hình tối ưu, vì chỉ số này cân bằng đồng thời giữa Precision và Recall, giúp phản ánh tốt hơn khả năng phát hiện lỗi thực sự của mô hình. Đối với mỗi loại lỗi, mô hình đạt F1-score cao nhất sẽ được sử dụng ở giai đoạn giải thích bằng XAI, bảo đảm rằng các phân tích SHAP và LIME được thực hiện dựa trên mô hình có chất lượng dự đoán tốt nhất.

Bảng 2. So sánh giữa các thuật toán trong mô hình

Mô hình	Đặc trưng chính	Tốc độ	Diễn giải	Độ chính xác	Khả năng XAI (SHAP/LIME)
Logistic Regression	Tuyến tính, đơn giản	Rất nhanh	Rất tốt	Trung bình	Kernel SHAP chậm, LIME tốt
Decision Tree	Cấu trúc cây, dễ hiểu	Rất nhanh	Tốt	Trung bình	Tree SHAP, LIME tốt
Random Forest	Nhiều cây, giảm overfitting	Trung bình	Trung bình	Cao	Tree SHAP mạnh, LIME tốt
XGBoost	Gradient boosting tối ưu, mạnh	Trung bình	Trung bình	Rất cao	Tree SHAP rất tốt
LightGBM	Boosting tối ưu tốc độ, chia theo lá	Nhanh nhất	Trung bình	Rất cao	Tree SHAP rất tốt

Bên cạnh việc dự đoán lỗi, nghiên cứu chú trọng khả năng giải thích nhằm hỗ trợ kỹ sư hiểu rõ nguyên nhân gây lỗi. Hai phương pháp XAI được sử dụng là SHAP và LIME, mỗi phương pháp mang lại góc nhìn bổ sung cho bài toán phân tích chất lượng. SHAP định lượng mức đóng góp của từng đặc trưng vào quyết định của mô hình. Trong nghiên cứu, SHAP được áp dụng ở dạng summary plot cho từng mô hình tốt nhất theo F1-score, cho phép xác định đặc trưng nào tác động mạnh nhất tới việc phát hiện từng loại lỗi. Đối với mô hình cây (Random Forest, XGBoost, LightGBM), TreeSHAP được dùng vì hiệu năng cao và tính chính xác. Với mô hình tuyến tính (Logistic Regression), KernelSHAP được sử dụng trong phạm vi 100 mẫu để đảm bảo tốc độ tính toán. LIME cung cấp giải thích cục bộ cho từng mẫu cụ thể, cho biết mỗi đặc trưng làm tăng hay giảm xác suất lỗi. Phương pháp này đặc biệt hữu ích cho kỹ sư chất lượng khi phân tích nguyên nhân của từng sản phẩm lỗi cụ thể trong dây chuyền.



Hình 1. Lưu đồ Data Pipeline

Hình trên trình bày kiến trúc tổng thể của Data Pipeline được sử dụng trong nghiên cứu nhằm đảm bảo dữ liệu được chuẩn hóa, làm sạch và chuyển đổi đúng cách trước khi đưa vào mô hình học máy. Quy trình gồm bốn nhóm bước chính, liên kết theo dòng chảy dữ liệu từ trên xuống dưới.

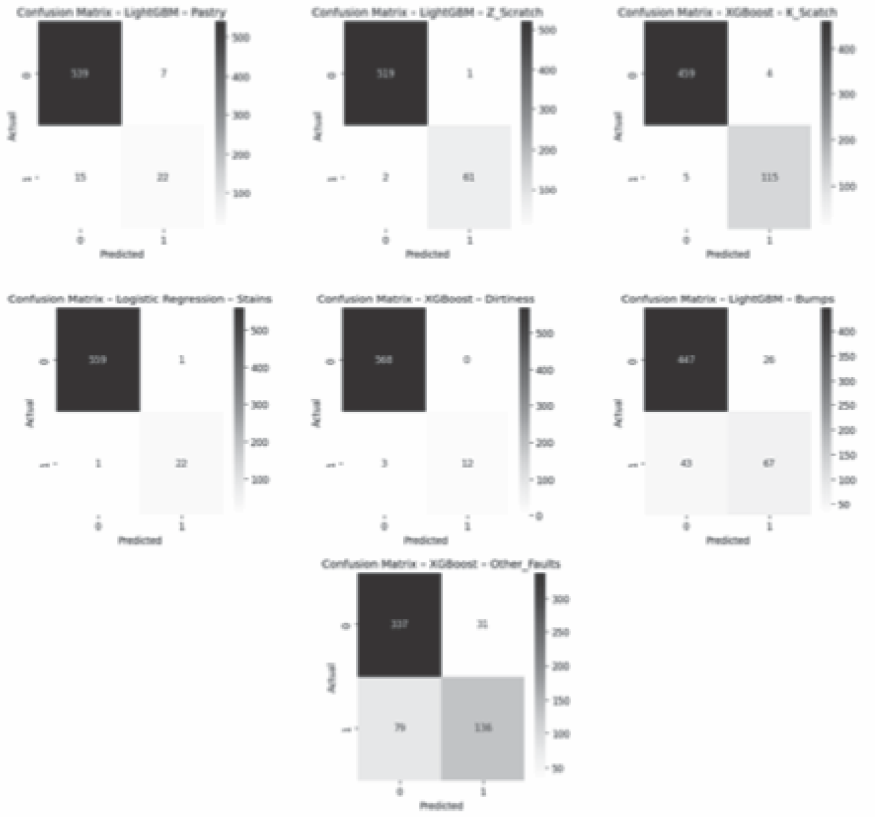
Đầu tiên, dữ liệu thô được tiếp nhận từ nguồn đầu vào và đưa vào giai đoạn tiền xử lý (Preprocessing), bao gồm làm sạch dữ liệu, loại bỏ các thuộc tính dư thừa và xử lý các giá trị bất thường hoặc thiếu hụt. Các biến đầu vào sau đó được chuẩn hóa bằng thuật toán StandardScaler, nhằm đưa tất cả đặc trưng về cùng thang đo để đảm bảo sự ổn định của mô hình.

Tiếp theo, dữ liệu sau khi chuẩn hóa được chia thành hai phần: tập huấn luyện và tập kiểm tra theo một tỷ lệ cố định. Tập huấn luyện được đưa vào module Machine Learning Training, nơi năm mô hình khác nhau (Logistic Regression, Decision Tree, Random Forest, XGBoost và LightGBM) được huấn luyện song song. Kết quả đánh giá trên tập kiểm tra được tổng hợp theo các chỉ số Accuracy, Precision, Recall, F1-score và AUC nhằm lựa chọn mô hình tối ưu cho từng nhóm lỗi.

Cuối cùng, mô hình tốt nhất được chuyển sang module XAI (Explainable AI), trong đó hai kỹ thuật SHAP và LIME được sử dụng để phân tích mức độ ảnh hưởng của các đặc trưng lên quyết định dự đoán của mô hình. Các biểu đồ giải thích giúp xác định các thuộc tính quan trọng, hỗ trợ phân tích nguyên nhân gốc rễ và đưa ra gợi ý cải tiến quy trình sản xuất.

Như vậy, Data Pipeline đảm bảo toàn bộ quá trình từ dữ liệu thô đến mô hình và giải thích dự đoán được tổ chức một cách khoa học, có thể kiểm tra và tái lập, đáp ứng yêu cầu của một nghiên cứu thực nghiệm trong lĩnh vực chất lượng công nghiệp.

4. KẾT QUẢ VÀ THẢO LUẬN

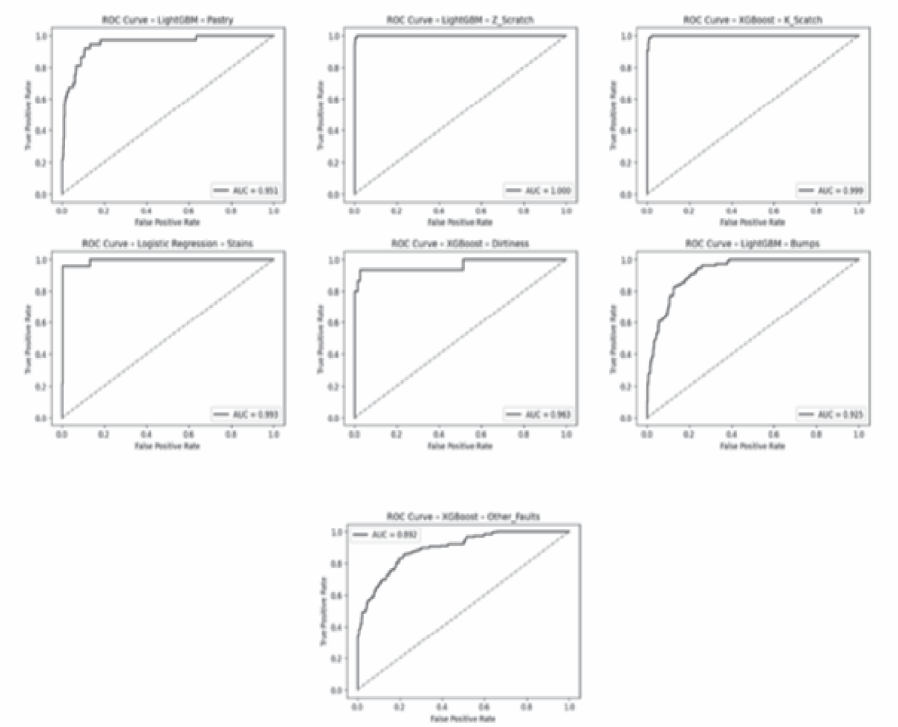


Hình 2. Kết quả ma trận nhầm lẫn đối với 7 lỗi

Kết quả ma trận nhầm lẫn cho bảy loại lỗi trong bộ dữ liệu Steel Plates Faults cho thấy mức độ khác biệt đáng kể giữa các mô hình ML khi xử lý các đặc trưng hình học bề mặt. Nhìn chung, cả LightGBM và XGBoost đều cho hiệu suất cao ở các lỗi có mẫu dữ liệu tương đối cân bằng, trong khi Logistic Regression chỉ thể hiện tốt ở các lỗi có ranh giới phân lớp tuyến tính rõ ràng. Đối với lỗi Pastry, mô hình LightGBM dự đoán chính xác 539/546 trường hợp thuộc lớp "không lỗi", song vẫn bỏ sót 15 mẫu lỗi thật (FN = 15). Điều này phản ánh việc mô hình ưu tiên giảm FP hơn là tăng khả năng phát hiện lỗi hiếm – một thách thức phổ biến trong dữ liệu mất cân bằng. Với lỗi Z_Scratch, LightGBM đạt độ chính xác rất cao với 519/520 mẫu "không lỗi" được phân loại đúng và chỉ 2 trường hợp lỗi bị bỏ sót. Điều này cho thấy các đặc trưng hình học liên quan đến vết xước trục Z đã được mô hình nhận diện tốt và ranh giới phân loại tương đối rõ ràng. Đối với lỗi K_Scatch, XGBoost thể hiện hiệu suất vượt trội khi phát hiện đúng 115/120 mẫu lỗi và chỉ tạo ra số lượng FP tối thiểu. Điều này phù hợp với đặc điểm mạnh của XGBoost trong việc học quan hệ phi tuyến và tương tác mạnh giữa các nhóm đặc trưng biên dạng. Lỗi Stains lại có bản chất tuyến tính hơn so với các loại lỗi khác, do đó Logistic Regression đạt hiệu suất dự đoán rất cao (TP = 22/23; FP = 1). Điều này cho thấy các thông số hình học liên quan đến vết bẩn có sự phân tách tuyến tính khá rõ ☞

trong không gian đặc trưng. Trong trường hợp lỗi Dirtiness, XGBoost tiếp tục cho kết quả ổn định với 12/15 mẫu lỗi được phân loại đúng và không tạo ra trường hợp dự đoán sai loại (FP = 0). Điều này cho thấy mô hình học được biên giới phân lớp sắc nét và ít bị nhiễu bởi các đặc trưng kém quan trọng. Ngược lại, lỗi Bumps là một trong những lỗi khó phân loại nhất. Kết quả từ LightGBM cho thấy mô hình vẫn gặp nhiều nhầm lẫn (FN = 43, FP = 26). Điều này có thể xuất phát từ sự chồng lấn đặc trưng hình học giữa lỗi "gồ ghề" và các loại lỗi hình dạng khác, làm giảm khả năng tách biệt của mô hình.

Cuối cùng, lỗi Other_Faults – vốn bao gồm nhiều dạng lỗi không đồng nhất – được XGBoost dự đoán tương đối tốt, nhưng vẫn còn tỷ lệ nhầm lẫn đáng kể (FN = 79; FP = 31). Điều này phản ánh tính chất phức tạp, đa dạng và khó mô hình hóa của nhóm lỗi này.

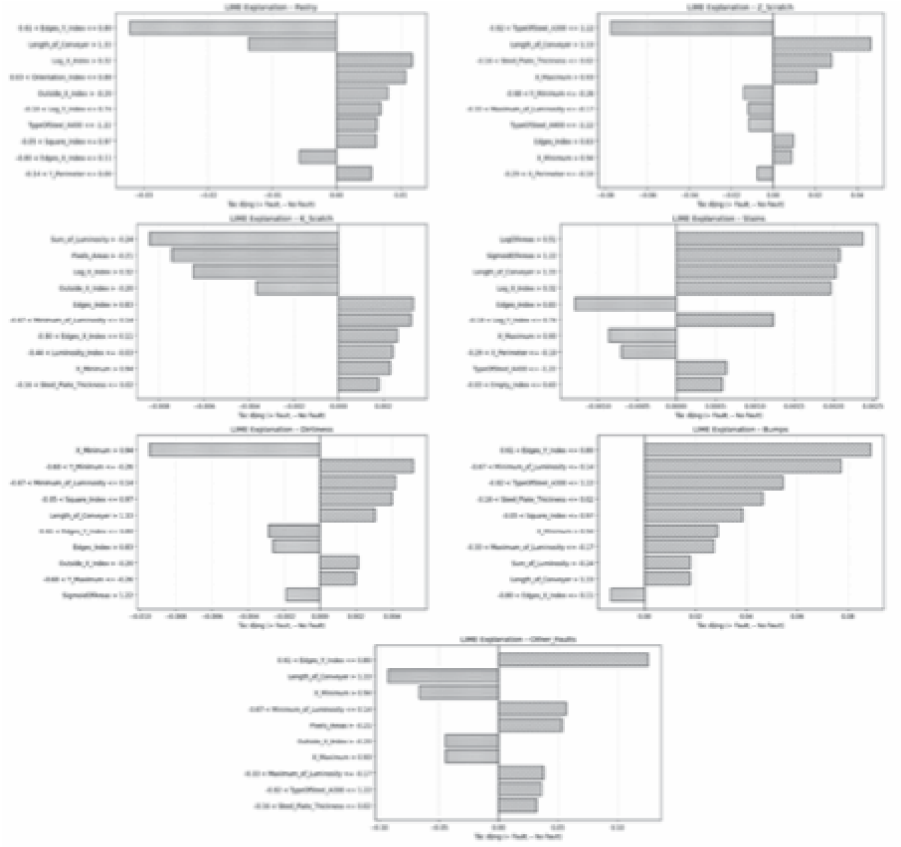


Hình 3. Đường cong ROC của mô hình học máy 7 lỗi

Các đường cong ROC được xây dựng cho từng loại lỗi nhằm đánh giá khả năng phân biệt (discriminative ability) của mô hình ML tối ưu được lựa chọn theo tiêu chí F1-score. Kết quả cho thấy hầu hết các mô hình đều đạt diện tích dưới đường cong (AUC) rất cao, phản ánh khả năng phân tách rõ ràng giữa sản phẩm lỗi và không lỗi. Đối với lỗi Bumps, mô hình LightGBM cho kết quả tốt với AUC = 0.925, thể hiện khả năng nhận diện các bất thường dạng gồ nổi khá ổn định. Mặc dù dữ liệu lỗi Bumps mất cân bằng mạnh, LightGBM vẫn duy trì độ phân biệt cao, cho thấy thuật toán boosting xử lý tốt các quan hệ phi tuyến trong dữ liệu hình học. Với lỗi Other_Faults, mô hình XGBoost đạt AUC = 0.892, thấp hơn so với các loại lỗi khác nhưng vẫn nằm trong ngưỡng "tốt". Điều này hợp lý vì nhóm lỗi "Other" có bản chất đa

dạng, bao gồm nhiều dạng khiếm khuyết khó phân loại, làm giảm độ đồng nhất của dữ liệu huấn luyện. Đối với lỗi Pastry, LightGBM thể hiện hiệu năng vượt trội (AUC = 0.951), chứng minh rằng các đặc trưng geometric-based trong dataset có khả năng mô tả rất rõ các trường hợp lỗi dạng bánh nhân. Kết quả ấn tượng nhất ghi nhận ở lỗi Z_Scratch, nơi LightGBM đạt AUC = 1.000, tức khả năng phân biệt hoàn hảo giữa dữ liệu lỗi và không lỗi. Tương tự, mô hình XGBoost cho lỗi K_Scatch đạt AUC = 0.999, phản ánh rằng hai loại lỗi này có đặc tính hình học rất đặc trưng và dễ nhận diện bởi mô hình boosting. Đối với lỗi Stains, Logistic Regression đạt AUC = 0.993, cho thấy ngay cả mô hình tuyến tính cũng có thể phân loại xuất sắc khi biên dạng dữ liệu không quá phức tạp và phân tách tuyến tính tốt. Điều này đồng thời khẳng định rằng không phải mọi loại lỗi đều cần đến mô hình phức tạp. Cuối cùng, lỗi Dirtiness đạt AUC = 0.963 với XGBoost. Đây là mức rất cao và phản ánh khả năng nhận diện ổn định đối với các lỗi dạng "bẩn bề mặt", vốn thường có biến thiên không đều.

Tổng hợp toàn bộ kết quả ROC cho thấy các mô hình Boosting (LightGBM, XGBoost) có xu hướng vượt trội hơn so với các mô hình tuyến tính và cây đơn, đặc biệt với những lỗi có biên dạng phi tuyến mạnh. Đồng thời, giá trị AUC cao ở tất cả các lỗi khẳng định rằng quy trình chuẩn hóa dữ liệu, phân tách tập huấn luyện – kiểm tra và lựa chọn mô hình theo F1-score là phù hợp với bối cảnh dữ liệu mất cân bằng của steel plate fault prediction.



Hình 4. Kết quả của mô hình LIME cho 7 lỗi

Sau khi chạy phương pháp SHAP, phương pháp LIME được sử dụng nhằm phân tích tác động cục bộ của từng đặc trưng lên quyết định dự đoán của mô hình tại từng mẫu lỗi cụ thể. Kết quả cho thấy mỗi loại lỗi bề mặt đều chịu ảnh hưởng bởi một tổ hợp đặc trưng riêng, phản ánh đúng bản chất hình học và phân bố điểm ảnh của từng dạng khiếm khuyết.

Đối với lỗi Pastry, các đặc trưng hình học như Edges_Y_Index, Length_of_Conveyer và các biến logarit (Log_X_Index, Log_Y_Index) góp phần mạnh mẽ vào việc đẩy dự đoán sang lớp lỗi. Những đặc trưng này phản ánh sự thay đổi về hình dạng và biên của vùng ảnh, phù hợp với bản chất bề mặt bị biến dạng kiểu cục bộ của lỗi Pastry. Với lỗi Z_Scratch, đặc trưng vật liệu (TypeOfSteel_A300) và các yếu tố hình học như Steel_Plate_Thickness, X_Maximum và Maximum_of_Luminosity đóng vai trò nổi bật. LIME chỉ ra rằng các mẫu có giá trị cao ở những đặc trưng này thường nghiêng về phía có lỗi, cho thấy mối liên quan giữa độ dày thép, độ sáng vùng biên và khả năng xuất hiện vết xước dạng Z. Tương tự, lỗi K_Scratch chịu ảnh hưởng chủ yếu từ Sum_of_Luminosity, Pixels_Areas và Edges_Index. Điều này cho thấy lỗi dạng vết xước K tạo ra thay đổi đáng kể trong mức sáng tổng vùng ảnh và phân bố điểm ảnh, phản ánh đúng cấu trúc tuyến tính đặc trưng của dạng lỗi này. Đối với lỗi Stains, các biến liên quan đến diện tích như LogOfAreas, SigmoidOfAreas cùng với các chỉ số biên (Edges_Index, X_Maximum) có tác động mạnh. Điều này hợp lý vì vết bẩn thường tạo ra sự thay đổi đột ngột về diện tích vùng tối/sáng trong ảnh, khiến mô hình ghi nhận đây là yếu tố phân biệt quan trọng. Lỗi Dirtiness cho thấy sự chi phối bởi các đặc trưng hình học (X_Minimum, Y_Minimum, Square_Index) cũng như các đặc trưng đo độ sáng. Các giá trị giới hạn này mô tả sự bất thường về hình dạng của vùng lỗi, thường xuất hiện ở mép hoặc rìa bề mặt. Cuối cùng, với lỗi Bumps, các đặc trưng Edges_Y_Index, Minimum_of_Luminosity, TypeOfSteel_A300 và Steel_Plate_Thickness là quan trọng nhất. Sự xuất hiện của các biến liên quan đến vật liệu và hình dạng cho thấy các vết gồ lên (bumps) có mối liên hệ mạnh với tính chất thép và đặc trưng vùng biên.

Nhìn chung, các biểu đồ LIME cho phép nhận diện rõ các đặc trưng giải thích tại từng mẫu cụ thể. Điều này không chỉ giúp xác nhận lại những phát hiện từ SHAP mà còn cung cấp thêm góc nhìn cục bộ về nguyên nhân gây lỗi hỗ trợ kỹ sư chất lượng xác định yếu tố rủi ro theo từng trường hợp và lập kế hoạch cải tiến quy trình phù hợp hơn.

5. KẾT LUẬN

Nghiên cứu đã xây dựng và đánh giá một hệ thống phân tích lỗi bề mặt thép dựa trên học máy kết hợp giải thích XAI, sử dụng dữ liệu từ bộ Steel Plate Defects (Dataset A). Kết quả thực nghiệm cho thấy các mô hình học máy hiện đại có khả năng nhận diện lỗi với độ chính xác cao, đặc biệt trong các nhóm lỗi có đặc trưng hình học rõ ràng. Trong số năm mô hình được thử nghiệm, LightGBM và XGBoost thể hiện ưu thế vượt trội ở hầu hết các loại lỗi, với các chỉ số F1-score và AUC đạt từ 0.88 đến gần 1.00. Điều này khẳng định hiệu quả của các thuật toán boosting trong xử lý quan hệ phi tuyến và tương tác phức tạp giữa các thuộc tính hình ảnh sau tiền xử lý.

Bên cạnh hiệu năng dự đoán, việc tích hợp SHAP và LIME giúp làm rõ cơ chế ra quyết định của mô hình, cho phép truy vết các yếu tố ảnh hưởng trực tiếp đến từng loại lỗi. SHAP cung cấp cái nhìn tổng quát về mức độ quan trọng của các đặc trưng, trong khi LIME cho phép phân tích cụ thể ở từng mẫu lỗi. Các kết quả giải thích cho thấy các thuộc tính

như Edges_Index, Edges_Y_Index, Length_of_Conveyer, Pixels_Areas, Log_X_Index, hoặc đặc tính vật liệu (TypeOfSteel_A300, TypeOfSteel_A400) có tác động nổi bật đến xác suất xảy ra lỗi. Điều này mở ra khả năng tối ưu quy trình sản xuất dựa trên những đặc trưng quan trọng nhất mà mô hình nhận diện.

Kết quả nghiên cứu không chỉ chứng minh tiềm năng của ML trong kiểm soát chất lượng thông minh, mà còn nhấn mạnh vai trò của XAI trong nâng cao tính minh bạch và mức độ tin cậy của hệ thống. Nhờ khả năng giải thích này, doanh nghiệp có thể sử dụng mô hình như một công cụ hỗ trợ quyết định – không chỉ để phát hiện lỗi, mà còn để hiểu nguyên nhân và đề xuất hướng cải tiến quy trình.❖

Ngày nhận bài: **12/11/2025**
Ngày phản biện: **24/11/2025**

Tài liệu tham khảo:

[1]. Huang, Y. C., Hung, K. C., & Lin, J. C. (2022), "*Automated machine learning system for defect detection on cylindrical metal surfaces*". Sensors, 22(24), 9783.

[2]. Dorbane, A., Harrou, F., & Sun, Y. (2025), "*Enhancing Defect Detection in Steel Plate Manufacturing with Explainable Machine Learning and SMOTE for Imbalanced Data*". Journal of Materials Engineering and Performance, 34(10), 9212-9233.

[3]. Wang, X., He, F., Yu, Y., Liu, X., Cong, J., Wu, Q., & Song, Y. (2025), "*Prediction of slab surface longitudinal crack based on resampling and the Bayesian optimisation of LightGBM*". Ironmaking & Steelmaking, 03019233251314962.

[4]. Lundberg, S. M., & Lee, S. I. (2017), "*A unified approach to interpreting model predictions*". Advances in Neural Information Processing Systems, 30.

[5]. Wang, X., Li, X., Yuan, H., Zhou, N., Wang, H., Zhang, W., & Ji, Y. (2024), "*Prediction and analysis of mechanical properties of hot-rolled strip steel based on an interpretable machine learning*". Materials Today Communications, 40, 109997.

[6]. Bode, G., Thul, S., Baranski, M., & Müller, D. (2020), "*Real-world application of machine-learning-based fault detection trained with experimental data*". Energy, 198, 117323.

[7]. Shakiba, F. M., Azizi, S. M., Zhou, M., & Abusorrah, A. (2023), "*Application of machine learning methods in fault detection and classification of power transmission lines: a survey*". Artificial Intelligence Review, 56(7), 5799-5836.