ISSN: 2476-6909; Modares. Mech. Eng. 2027;27(04):327-338.

Modares Mechanical Engineering

DOI: 10.48311/mme.2026.123003.83090
journal homepage: mme.modares.ac.ir

Laboratory Assessment of Corrosion in 316L Stainless Steel Based on Lamb Waves and Machine Learning

Mohammad Khodadosti*, Abdolrahim Taheri

Department of Mechanical Engineering, Petroleum University of Technology, Abadan, Iran

ARTICLE INFO

Article Type
Original Research

Article History
Received: August 22, 2026
Revised: September 26, 2026
Accepted: October 10, 2026
ePublished: November 14, 2026

ABSTRACT

Reliable corrosion monitoring becomes more difficult when a diagnostic model is expected to remain effective beyond the experiment used for its development. This study combines Lamb-wave measurements with machine learning to assess corrosion in 316L stainless steel, with particular emphasis on transfer between independent experiments. A public laboratory dataset containing two corrosion experiments was analyzed, comprising 429 corrosion states and 12 directed piezoelectric actuator-sensor paths per state. Damage indices, baseline-relative descriptors, and frequency-domain features were evaluated using conventional machine-learning models and a physics-guided path-attention multi-task learning framework (PG-PAMTL). Performance was assessed under within-experiment validation, strict zero-shot transfer, and unsupervised domain adaptation. In the strict cross-experiment setting, PG-PAMTL achieved a macro ROC-AUC of 0.830 and a PR-AUC of 0.871, indicating substantially stronger transferability than the conventional benchmarks. Domain adaptation increased the classification ROC-AUC only marginally to 0.832, but its effect was more evident for severity estimation, where the macro MAE decreased from 39.3 to 32.5 µm. Permutation analysis showed that predictive information was concentrated in a limited number of sensing paths and in both waveform-change and frequency-domain features. A separate sensitivity analysis also confirmed that a known bubble-related disturbance affected transfer performance. Experimental progression, however, was strongly associated with corrosion severity, limiting the interpretation of the regression results. The findings therefore support Lamb-wave machine learning for cross-experiment corrosion detection, while showing that reliable absolute severity estimation still requires validation under more independent experimental conditions.

Keywords: Lamb Waves; Corrosion Monitoring; 316L Stainless Steel; Machine Learning; Domain Adaptation

How to cite this article
Khodadosti M, Taheri A, Laboratory Assessment of Corrosion in 316L Stainless Steel Based on Lamb Waves and Machine Learning. Modares Mechanical Engineering; 2027;27(04):327-338.

*Corresponding author's email: mohammadkhodadosti@gmail.com
*Corresponding ORCID ID: 0009-0005-3460-259X

Copyright© 2027, TMU Press. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

شاپا: ۶۹۰۹-۲۴۷۶؛ مهندسی مکانیک مدرس. ۳۳۸-۳۲۷:(۰۴)۲۷؛۱۴۰۶

مهندسی مکانیک مدرس

DOI: 10.48311/mme.2026.123003.83090

صفحه اصلی مجله: mme.modares.ac.ir

ارزیابی آزمایشگاهی خوردگی فولاد زنگ‌نزن ۳۱۶L مبتنی بر امواج لمب و یادگیری ماشین

محمد خدادوستی*، عبدالرحیم طاهری

گروه مهندسی مکانیک، دانشگاه صنعت نفت، آبادان، ایران

اطلاعات مقاله

نوع مقاله

مقاله پژوهشی

تاریخچه مقاله

دریافت: ۱۴۰۵/۰۵/۳۱

بازنگری: ۱۴۰۵/۰۷/۰۴

پذیرش: ۱۴۰۵/۰۷/۱۸

ارائه آنلاین: ۱۴۰۵/۰۸/۲۳

چکیده

پایش قابل‌اعتماد خوردگی زمانی دشوارتر می‌شود که یک مدل تشخیصی باید خارج از آزمایشِ مبنای توسعه نیز عملکرد خود را حفظ کند. در این پژوهش، اندازه‌گیری‌های امواج لمب با یادگیری ماشین برای ارزیابی خوردگی فولاد زنگ‌نزن ۳۱۶L و با تمرکز بر انتقال بین دو آزمایش مستقل به کار گرفته شد. مجموعه‌داده عمومی شامل ۴۲۹ حالت خوردگی و ۱۲ مسیر جهت‌دار محرک–حسگر پیزوالکتریک برای هر حالت بود. شاخص‌های آسیب، توصیفگرهای نسبی نسبت به حالت مبنا و ویژگی‌های حوزه فرکانس با مدل‌های متداول و چارچوب PG-PAMTL ارزیابی شدند. عملکرد مدل‌ها در اعتبارسنجی درون‌آزمایشی، انتقال سخت‌گیرانه صفرنمونه‌ای و سازگارسازی دامنه بدون نظارت بررسی شد. در انتقال صفرنمونه‌ای، PG-PAMTL به ROC-AUC کلان ۰٫۸۳۰ و PR-AUC برابر ۰٫۸۷۱ دست یافت و نسبت به مدل‌های مرجع انتقال‌پذیری بهتری نشان داد. سازگارسازی دامنه ROC-AUC را تنها تا ۰٫۸۳۲ افزایش داد، اما MAE شدت خوردگی را از ۳۹٫۳ به ۳۲٫۵ میکرومتر کاهش داد. تحلیل اهمیت نشان داد که تعداد محدودی از مسیرهای حسگری و ترکیبی از ویژگی‌های تغییر شکل موج و حوزه فرکانس بیشترین سهم را در پیش‌بینی‌ها دارند. اختلال شناخته‌شده مرتبط با حباب نیز بر عملکرد انتقال اثرگذار بود. با این حال، ارتباط زیاد پیشرفت مرحله‌ای آزمایش با شدت خوردگی، تفسیر نتایج رگرسیون را محدود می‌کند. بنابراین، روش پیشنهادی برای تشخیص خوردگی بین آزمایش‌های مستقل امیدوارکننده است، اما برآورد مطلق شدت خوردگی همچنان به اعتبارسنجی مستقل‌تر نیاز دارد.

کلیدواژه‌ها :امواج لمب؛ پایش خوردگی؛ فولاد زنگ‌نزن 316L؛ یادگیری ماشین؛ سازگارسازی دامنه

نحوه ارجاع به این مقاله

خدادوستی محمد، طاهری عبدالرحیم، ارزیابی آزمایشگاهی خوردگی فولاد زنگ‌نزن ۳۱۶L مبتنی بر امواج لمب و یادگیری ماشین، مهندسی مکانیک مدرس. ۳۳۸-۳۲۷:(۰۴)۲۷؛۱۴۰۶.

*پست الکترونیکی نویسنده عهده‌دار مکاتبات: mohammadkhodadosti@gmail.com

*شناسه ارکید نویسنده عهده‌دار مکاتبات: ۰۰۰۹-۰۰۰۵-۳۴۶۰-۲۵۹x

Copyright© 2027, TMU Press. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

1- Introduction

Revision note: The scope of this study has been clarified. The present work is a secondary analysis of the publicly available COQTEL laboratory dataset rather than a new experimental campaign. Accordingly, the contribution of this study is defined as evaluating the cross-experiment robustness of Lamb-wave-based machine-learning approaches and identifying the conditions under which corrosion-related information can be transferred between independent laboratory domains. Corrosion remains a major threat to the service life and safety of metallic structures. Localized material loss and pit formation create stress concentrations that can gradually weaken a component and increase the risk of failure [1, 2]. Although 316L stainless steel is widely selected for its mechanical performance and corrosion resistance, it is still susceptible to localized and pitting attack in aggressive environments, particularly when chloride ions are present [3, 4]. Early detection and reliable tracking of this damage are therefore important for maintaining structural integrity. Guided waves, and Lamb waves in particular, are well suited to structural health monitoring of plate-like components because they can interrogate relatively large areas from a limited number of transducers [5, 6]. Bonded piezoelectric transducers (PZTs) can both excite and receive these waves, making multi-path sensing networks practical for repeated monitoring [7]. Corrosion alters amplitude, energy, waveform similarity, and spectral content, providing several complementary indicators of damage [8]. Earlier studies have demonstrated sensitivity to early corrosion and wall-thickness loss [9, 10], and the use of multiple modes or actuator–sensor paths can further improve damage sensitivity [11, 12]. Experimental work on 316L stainless steel has also shown that PZT arrays can track small corrosion pits through Lamb-wave measurements [13]. The main difficulty is interpretation: Lamb waves are dispersive and multimodal, and their response is easily influenced by experimental conditions. This has motivated the use of machine learning to combine several signal descriptors and capture relationships that are difficult to express with a single damage index [14, 15]. Deep-learning approaches have extended this idea by learning damage-sensitive patterns directly from ultrasonic data [16], including information related to localization and severity [17].

Related studies have investigated Lamb-wave imaging for corrosion monitoring in automotive lap joints [18], multimodal corrosion detection and localization [19], and corrosion quantification using Lamb wavefields and monogenic signal processing [20]. Transfer learning for temperature-compensated damage identification [21] and machine learning with aggregated baselines [22] have also been explored to address variability in guided-wave monitoring. Recent work has moved beyond simple damage detection toward tracking corrosion development and improving transfer across changing conditions . Ding et al. [23] showed that nonlinear Lamb-wave quantities, including higher-order harmonics, can respond to early surface corrosion before substantial material loss is present. Zhang et al. [24] used machine-learning models with time-, frequency-, and time–frequency-domain features and demonstrated that these descriptors can support automated classification. Wang et al. [25] combined guided-wave measurements with convolutional neural networks to recover corrosion-related spatial information, while Huang et al. [26] showed that combining information from multiple Lamb-wave modes can improve both detection and localization. More recent studies have addressed severity estimation and model transfer. Azad et al. [27] compared deep-learning architectures for Lamb-wave-based localization and severity assessment, and Höll and Humer [28] investigated transfer from simulation to experiment. Of particular relevance here, Nicard et al. [29] monitored electrochemically generated pits in a 316L stainless-steel plate using four PZT transducers and reported a clear relationship between Lamb-wave damage indicators and equivalent pit radius. Lu et al. [30] subsequently showed that unsupervised domain adaptation can improve Lamb-wave damage identification when source and target distributions differ. Despite this progress, corrosion detection, severity estimation, multi-path sensing, and experiment-to-experiment generalization are still rarely evaluated together. The possible confounding effect of experimental progression has also received comparatively little attention. The present study addresses these issues within a single evaluation framework and asks not only whether the models predict well, but also whether their cross-experiment performance is genuinely supported by Lamb-wave information.

2- Materials and Methods

The present analysis uses real experimental data from controlled pitting-corrosion tests on 316L stainless-steel plates monitored with ultrasonic Lamb waves. The two experiments were conducted by the original COQTEL dataset authors on 1.5-mm-thick, 80 × 125 mm specimens instrumented with four bonded PZT transducers [31]. Corrosion was generated electrochemically while Lamb-wave responses were recorded repeatedly as the pits developed. In this study, the electrochemical measurements serve only as the reference used to define corrosion level; the predictive models are built from the ultrasonic measurements. This separation makes it possible to evaluate detection, severity estimation, and transfer from one independent experiment to the other without giving the models direct electrochemical information.

2-1- Dataset Source and Experimental Configuration

The data were taken from the publicly available COQTEL dataset reported by Nicard et al. [31]. It contains real laboratory measurements from controlled corrosion tests on 316L stainless-steel plates; no simulated or synthetically generated ultrasonic signals were introduced in the present analysis. The original release includes two independent HDF5 experiments, Essai_Corrosion1.h5 and Essai_Corrosion2.h5. Experiment 1 contains 166 measurement states and Experiment 2 contains 263, for a total of 429 states. Each 80 × 125 × 1.5 mm plate was instrumented with four bonded PZT transducers [31]. During a measurement sequence, one transducer acted as the actuator and the other three as receivers. Cycling the actuator role through all four PZTs produced 12 directed actuator–sensor paths for every state. Excitation was a five-cycle tone burst centered at 200 kHz; the responses were sampled at 2 MHz with 2000 samples per acquisition. Fig. 1 summarizes the experimental configuration and monitoring arrangement. The two experiments were kept as separate experimental domains rather than pooled and randomly divided. This choice is central to the study because it exposes the models to genuine experiment -to -experiment variability at test time. Electrochemical variables supplied with the dataset were retained only for constructing the corrosion reference and were excluded from the predictive inputs. The main characteristics of the data used in the analysis are listed in Table 1.

2-2- Corrosion Reference and Target Definition

Corrosion severity was represented by the equivalent pit radius estimated from cumulative anodic charge. For classification, 30 µm was used as the primary threshold: states below this value were assigned to the below-threshold class and states at or above 30 µm to the corrosion class. A 20 µm threshold was examined only as a sensitivity check and was not used for the main classification results. For regression, the equivalent pit radius remained a continuous target. Electrochemical measurements were therefore used to define the reference labels only; none were included among the machine-learning inputs. Table 2 summarizes the target definitions and class distributions in the two experiments.

2-3- Lamb-Wave Data Acquisition and Signal Processing

For every corrosion state, the four bonded PZT transducers were operated sequentially. One transducer excited the plate while the other three recorded the response, producing 12 directed actuator–sensor paths per state. The excitation was a five-cycle tone burst centered at 200 kHz, sampled at 2 MHz with 2000 points per acquisition

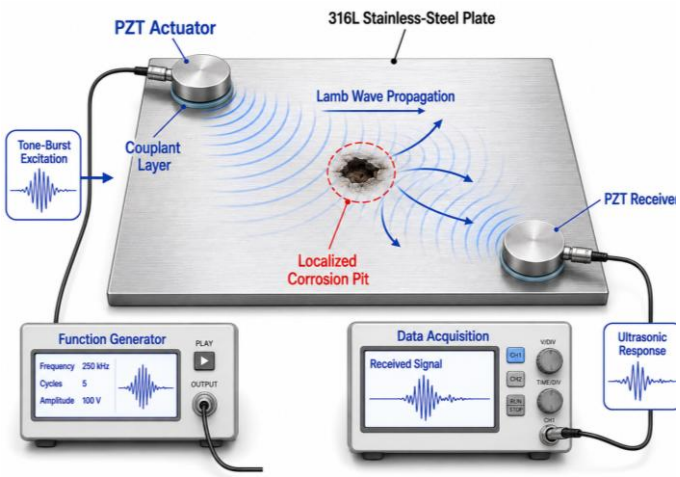


Fig. 1 Schematic representation of the experimental COQTEL configuration for Lamb-wave monitoring of localized corrosion in a 316L stainless-steel plate.

Table 1 Main characteristics of the experimental COQTEL dataset used in this study.

Parameter	Description
Dataset	COQTEL
Data source	Public experimental dataset reported by Nicard et al. [31]
Data type	Real laboratory measurements
Material	316L stainless steel
Specimen dimensions	80 × 125 × 1.5 mm
Number of experiments	2 independent experiments
Experimental states	166 (Exp1), 263 (Exp2), 429 total
PZT transducers	4 bonded PZT transducers
Actuator–sensor paths	12 directed paths per state
Excitation	Five-cycle tone burst
Center frequency	200 kHz
Sampling frequency	2 MHz
Samples per acquisition	2000
Damage type	Electrochemically generated localized corrosion

Table 2 Definition of corrosion targets and dataset distribution

Item	Experiment 1	Experiment 2	Definition
Total number of states	166	263	Experimental measurement states
Below corrosion threshold	27	97	Equivalent pit radius < 30 μm
Corrosion class	139	166	Equivalent pit radius ≥ 30 μm
Primary classification threshold	30 μm	30 μm	Main threshold used for corrosion detection
Sensitivity threshold	20 μm	20 μm	Used only for sensitivity analysis
Regression target	Equivalent pit radius (μm)	Equivalent pit radius (μm)	Continuous corrosion-severity target

Signals from different paths were kept separate so that path-specific changes were not lost before the learning stage.Each waveform was characterized both in its original form and relative to the first measurement of the same path. State 1 served as the baseline, and the residual signal was defined as

$$r(t) = x(t) - x_0(t) \quad (1)$$

where x(t) is the current response and x0(t) is the corresponding baseline signal. The overall waveform change was quantified using NRMSD,

$$\mathrm{NRMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_{0i})^2}{\sum_{i=1}^{N} x_{0i}^2}} \quad (2)$$

Waveform correlation was considered alongside NRMSD. The feature set also included RMS amplitude, peak response, signal energy, spectral centroid, bandwidth, spectral entropy, and relative spectral energy, allowing the model to use complementary information from both the time and frequency domains. Figure 2 shows a representative PZT3→PZT1 response from Experiment 1, comparing State 1 with State 126, corresponding to an equivalent pit radius of about 140.5 μm. The complete and zoomed waveforms remain broadly similar, but the residual signal reveals measurable changes in amplitude and waveform shape. In this example, the NRMSD was 0.321 and the correlation was 0.948, indicating that corrosion introduced noticeable but still relatively subtle changes in the response. The frequency-domain comparison also shows differences around the 200 kHz excitation region. Together, these observations support the use of combined temporal, spectral, and baseline-referenced features for the subsequent machine-learning analysis.

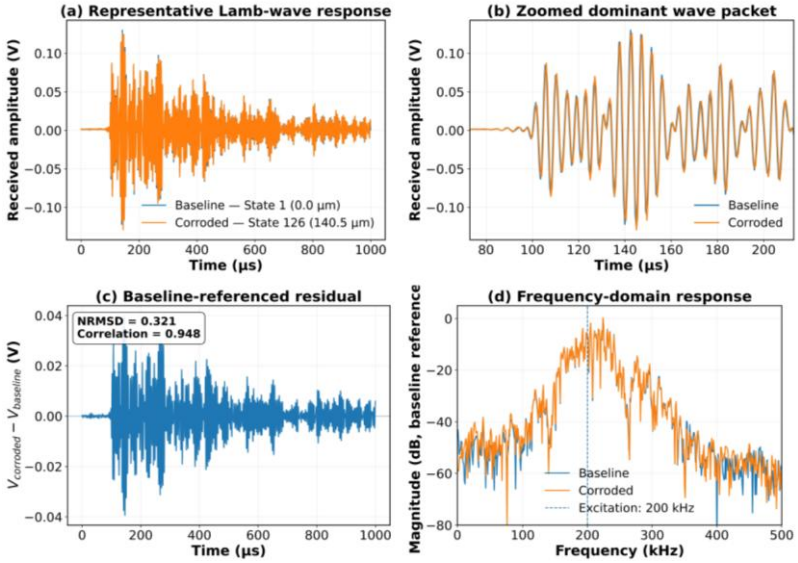


Fig. 2. Representative Lamb-wave responses for the PZT3→PZT1 path in Experiment 1, comparing the baseline state with a developed corrosion state. (a) Full received waveform, (b) zoomed dominant wave packet, (c) baseline-referenced residual, and (d) corresponding frequency-domain response.

2-4- Machine-Learning Framework

Additional methodological clarification: The learning framework is designed to preserve the physical structure of guided-wave measurements. Unlike conventional models that flatten all actuator–sensor responses into a single feature vector, PG-PAMTL represents each corrosion state as a collection of directional sensing paths. A shared encoder generates path-level representations, while an attention mechanism learns the relative contribution of individual paths. The multi-task formulation combines corrosion classification and severity estimation with auxiliary regularization objectives to improve representation stability. The model configuration was selected through validation-based optimization to balance predictive performance and overfitting risk. Because the available dataset consists of sequential measurements from a limited number of physical experiments, model selection was performed with emphasis on temporal-block validation rather than random sample splitting. The learning framework addressed two related tasks: binary corrosion detection and continuous severity estimation. Detection labels were assigned from the equivalent pit radius using the 30 μm threshold,

$$y_{\text{cls}} = 0 r_{\text{eq}} < 30\mu\text{m}; y_{\text{cls}} = 1 r_{\text{eq}} \geq 30\mu\text{m} \quad (3)$$

For severity estimation, the equivalent pit radius was treated as a continuous regression target,

$$y_{\text{reg}} = r_{\text{eq}} \quad (4)$$

2-5- Proposed Model and Cross-Experiment Learning Strategy

Additional clarification: The physics-guided constraints are introduced as regularization mechanisms rather than additional sources of corrosion information. Experimental progression is not provided as an input feature. Instead, the monotonicity constraint is used only to discourage physically inconsistent severity predictions. This distinction is important because experimental order itself can contain information correlated with corrosion development. PG-PAMTL (Physics-Guided Path-Attention Multi-Task Learning) was designed around the directed sensing paths rather than a single flattened feature vector. Each experimental state is represented by its 12 actuator–sensor paths. For each path, the Lamb-wave descriptors are combined with basic geometric information describing the actuator, receiver, propagation direction, and path length. A shared encoder maps these inputs into path-level representations, and an attention mechanism allows the network to emphasize paths that carry more useful damage information. The attended path representations are then pooled into a common latent representation used for corrosion detection and severity estimation. An auxiliary corrosion-stage task is included during training to encourage a more organized representation of damage progression. Because the assumed pit geometry links corrosion volume to the cube of equivalent radius, the regression target is expressed in normalized volume-related form,

$$v = \left(\frac{r_{\text{eq}}}{160}\right)^3 \quad (5)$$

Learning the related targets together allows the network to share information between detection and severity estimation, consistent with the general motivation for multi-task learning [34]. A physics-guided monotonicity constraint was also introduced. Corrosion develops progressively in the controlled experiments, so large decreases in predicted severity between successive states would be inconsistent with the experimental process. The corresponding penalty discourages such behavior during training. Experimental state and time are not used as predictive inputs; progression enters only through this training constraint. The total loss is

$$L = L_{\text{cls}} + 1.5L_{\text{reg}} + 0.20L_{\text{stage}} + 0.35L_{\text{mono,s}} + 0.30L_{\text{monot}} + 0.40L_{\text{DA}} \quad (6)$$

Only Lamb-wave-derived information was supplied to the models. Electrochemical measurements were reserved for defining the reference targets. Time-domain, frequency-domain, and baseline-relative descriptors were retained at the individual path level so that potentially useful differences among the 12 propagation paths remained available. Table 3 summarizes the models used for comparison. SVM/SVR and XGBoost follow established benchmark formulations [32, 33], Random Forest provides a nonlinear ensemble reference, and a 1D-CNN serves as a waveform-based deep-learning comparison. The proposed PG-PAMTL model learns corrosion detection and pit-radius estimation jointly while preserving the multi-path structure of the measurements.

Table 3 Machine-learning models considered in this study.

\	Type	Purpose
SVM / SVR	Kernel-based learning	Conventional classification and regression benchmark
Random Forest	Ensemble learning	Nonlinear tree-based benchmark
XGBoost	Gradient boosting	Benchmark for engineered Lamb-wave features
1D-CNN	Deep learning	Waveform-based deep-learning comparison
PG-PAMTL	Proposed multi-task learning	Joint corrosion detection and severity estimation using multi-path information

The terms correspond to classification, regression, auxiliary stage prediction, monotonicity, and domain-alignment losses. Transfer was evaluated in both directions, Exp1→Exp2 and Exp2→Exp1. Two protocols were kept distinct. In strict zero-shot evaluation, training used the source experiment only and no target-domain information was introduced. PG-PAMTL-DA, in contrast, used unlabeled target features for unsupervised domain adaptation. Correlation alignment was applied to reduce the discrepancy between source and target latent representations without using target corrosion labels, following the CORAL principle [35]. For deployment, the classification operating point was estimated from unlabeled target scores using two-cluster K-means. Separating these protocols avoids conflating true zero-shot generalization with improvements that depend on access to unlabeled target data.

2-6- Model Evaluation, Interpretability, and Robustness

Additional validation clarification: The number of recorded states is not considered equivalent to the number of independent physical experiments. Therefore, uncertainty analysis and bootstrap procedures are interpreted considering temporal dependence among measurements. The current conclusions are restricted to transfer between the two investigated laboratory experiments. Detection and severity estimation were evaluated separately. Detection metrics included ROC-AUC, PR-AUC, balanced accuracy, F1-score, precision, recall, specificity, and MCC; severity was assessed using MAE, RMSE, R^2, Spearman correlation, and Pearson correlation. Random splitting was avoided because neighboring corrosion states are temporally related. Within-experiment validation therefore used purged contiguous blocks, while transfer was tested in both Exp1→Exp2 and Exp2→Exp1 directions. Uncertainty in the main transfer results was estimated with a moving-block bootstrap using 2000 resamples and a block length of 10 states [36]. Permutation importance was used to examine model dependence on individual actuator–sensor paths and Lamb-wave features; importance was measured by the resulting decrease in ROC-AUC [37]. Robustness was examined through ablation, a state-only progression control, Gaussian feature perturbations at 40, 30, 20, and 10 dB, and matched exclusion analyses for the known disturbed interval in Experiment 2. Conformal intervals were also explored for severity predictions. Because the data are sequential and the experiments exhibit domain shift, these intervals are interpreted as diagnostics rather than guaranteed future coverage [38].

3- Results and Discussion

The results are organized from the easier within-experiment setting to the more demanding cross-experiment tests. The analysis then examines what changes when unlabeled target data are available, how well corrosion severity transfers, and which components of PG-PAMTL account for the observed gains. Progression confounding, sensing-path and feature importance, and sensitivity to the known experimental artifact are considered alongside predictive performance so that the results can be interpreted in the context of the experimental design.

3-1- Within-Experiment Benchmark Performance

Within-experiment performance was examined first using purged-block validation to reduce leakage between neighboring corrosion states. Table 4 reports the numerical results and Fig. 3 compares the detection and regression metrics. For corrosion detection, the feature-based models were strong. SVM/SVR achieved the highest balanced accuracy (0.772) and F1-score (0.898), while Random Forest produced the highest ROC-AUC (0.940), closely followed by XGBoost (0.937). PG-PAMTL remained competitive, with balanced accuracy of 0.763 and F1-score of 0.888, although its ROC-AUC was lower at 0.849. The 1D-CNN was markedly weaker, with a ROC-AUC of 0.559. The comparison in Fig. 3(a) also shows why a single classification metric is not sufficient. Random Forest and XGBoost ranked the two classes well, as reflected by their high ROC-AUC values, but their balanced accuracy was lower than that of SVM/SVR at the selected operating point. PG-PAMTL was more even across the three classification metrics, whereas the waveform-based 1D-CNN lost much of its discriminative ability. Under these within-experiment conditions, engineered Lamb-wave descriptors therefore provided a more stable representation than the raw-waveform CNN benchmark. The separation between models was even clearer for severity estimation. XGBoost achieved the best regression performance ($R^2 = 0.77$, MAE = 14.4 μm), followed by Random Forest ($R^2 = 0.71$, MAE = 17.1 μm). PG-PAMTL reached an MAE of 18.4 μm with $R^2 = 0.43$, while the 1D-CNN produced the largest error (36.5 μm) and a negative R^2 of −0.30. Thus, conventional feature-based models remained highly competitive when training and validation came from the same experiment. The purpose of PG-PAMTL is therefore not to dominate the easier within-domain benchmark, but to retain useful information when the experimental domain changes.

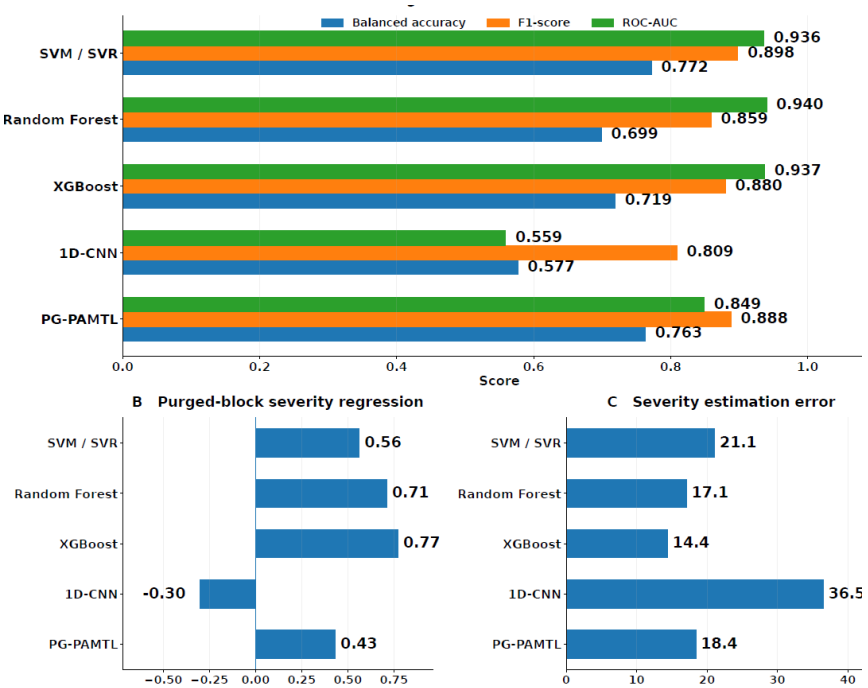


Fig. 3 Within-experiment performance under leakage-resistant purged-block validation: (a) corrosion-detection performance based on balanced accuracy, F1-score, and ROC-AUC; (b) R^2 for equivalent pit-radius estimation; and (c) MAE for corrosion-severity estimation.

Table 4 Within-experiment performance under purged-block validation.

Model	Balanced Accuracy	F1-score	ROC-AUC	R^2	MAE (μm)
SVM / SVR	0.772	0.898	0.936	0.56	21.1
Random Forest	0.699	0.859	0.940	0.71	17.1
XGBoost	0.719	0.880	0.937	0.77	14.4
1D-CNN	0.577	0.809	0.559	−0.30	36.5
PG-PAMTL	0.763	0.888	0.849	0.43	18.4

3-2- Strict Zero-Shot Cross-Experiment Generalization

A stricter test was then considered: the model was trained on one experiment and evaluated directly on the other, with no target-domain labels used during training. Table 5 and Fig. 4 summarize the results. The separation between PG-PAMTL and the conventional benchmarks became much clearer in this setting. PG-PAMTL achieved a ROC-AUC of 0.830 and a PR-AUC of 0.871, whereas SVM/SVR, Random Forest, XGBoost, and 1D-CNN reached ROC-AUC values of 0.587, 0.405, 0.391, and 0.543, respectively. The transfer results show that the strong within-experiment performance of Random Forest and XGBoost did not carry over to an independent experiment. PG-PAMTL retained substantially more class-separation information, suggesting that its learned representation was less tied to experiment-specific characteristics. The fixed-threshold result, however, tells a different story. Despite its ROC-AUC and PR-AUC, PG-PAMTL achieved a balanced accuracy of only 0.502 when the source-domain threshold was transferred unchanged; the other models were also close to 0.50. This combination of good ranking and near-chance balanced accuracy is more consistent with a shift in the score distribution and operating point than with a complete loss of corrosion sensitivity. In practical terms, the model could still order target samples meaningfully, but the source threshold was poorly calibrated for the new experiment. This distinction motivates the unsupervised calibration and domain-adaptation analyses that follow.

3-3- Domain-Adaptation and Transfer Performance

We then asked whether unlabeled data from the target experiment could improve transfer. As shown in Fig. 5 and Table 6, the effect was modest for detection but more noticeable for severity estimation. The PG-PAMTL core model achieved macro ROC-AUC and PR-AUC values of 0.830 and 0.871, compared with 0.832 and 0.878 for PG-PAMTL-DA. The small change indicates that most of the discriminative information was already present in the source-trained representation. After unsupervised target-score calibration, balanced accuracy was also similar: 0.758 for the core model and 0.756 for PG-PAMTL-DA. Viewed together with the strict zero-shot result, this supports the interpretation that the fixed-threshold failure was driven mainly by operating-point shift rather than loss of class separation. The larger benefit appeared in severity estimation, where macro MAE fell from 39.3 to 32.5 μm, a 17.3% reduction. Directional results remained useful in both transfers: Exp1→Exp2 reached ROC-AUC = 0.843 and AP = 0.795, while Exp2→Exp1 reached 0.822 and 0.960. The latter AP should be read in light of the higher positive-class prevalence (0.837 versus 0.631). Domain adaptation therefore contributed mainly to regression stability, with little additional gain in classification discrimination.

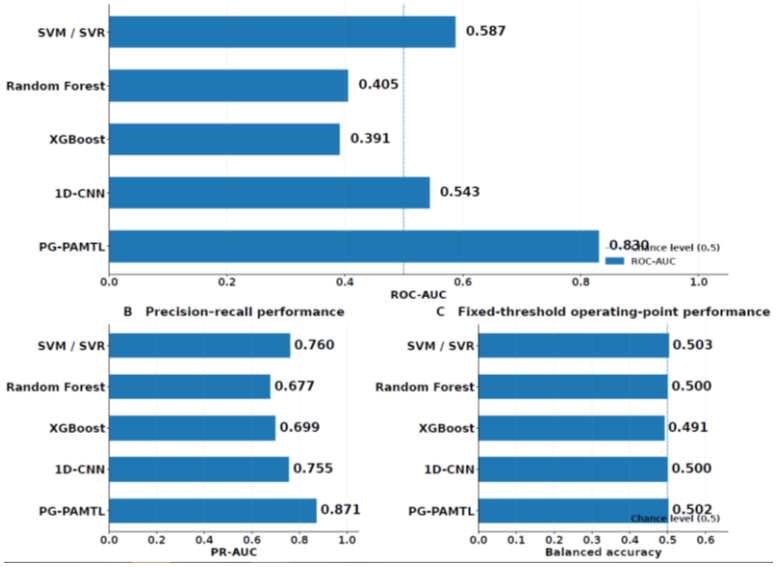


Fig. 4 Strict zero-shot cross-experiment corrosion-detection performance: (a) ROC-AUC comparison across models; (b) PR-AUC performance; and (c) balanced accuracy obtained using the fixed source-domain decision threshold. PG-PAMTL retained the strongest cross-experiment discrimination, whereas the near-chance balanced accuracy indicates a shift in the operating point between the two experimental domains.

Table 5 Strict zero-shot cross-experiment corrosion-detection performance

Model	Balanced Accuracy	PR-AUC	ROC-AUC
SVM / SVR	0.503	0.760	0.587
Random Forest	0.500	0.677	0.405
XGBoost	0.491	0.699	0.391
1D-CNN	0.500	0.755	0.543
PG-PAMTL	0.502	0.871	0.830

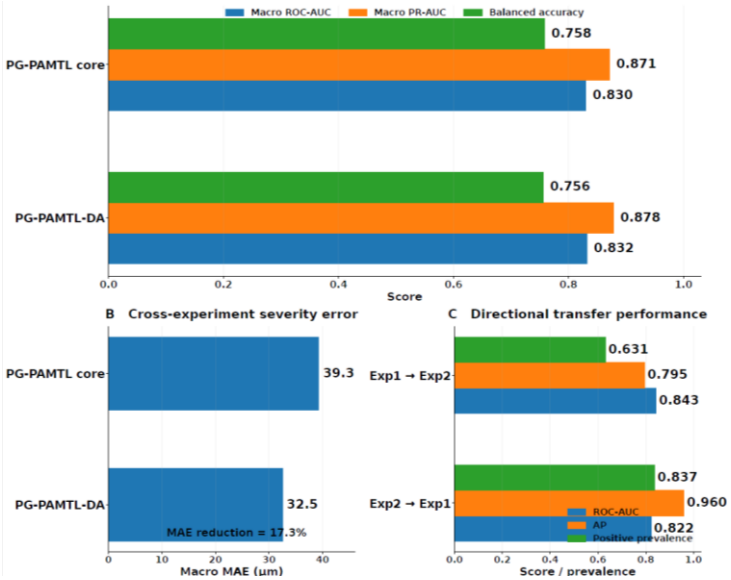


Fig. 5 Domain-adaptation and cross-experiment transfer performance: (a) comparison of macro ROC-AUC, macro PR-AUC, and deployment balanced accuracy for PG-PAMTL and PG-PAMTL-DA; (b) change in cross-experiment severity MAE following unsupervised domain adaptation; and (c) directional transfer performance for Exp1→Exp2 and Exp2→Exp1, together with the corresponding positive-class prevalence.

Table 6 Cross-experiment transfer performance of PG-PAMTL and PG-PAMTL-DA

Evaluation setting	ROC-AUC	PR-AUC / AP	Balanced accuracy	MAE (μm)	Positive prevalence
PG-PAMTL core	0.830	0.871	0.758	39.3	—
PG-PAMTL-DA	0.832	0.878	0.756	32.5	—
Exp1 → Exp2, PG-PAMTL-DA	0.843	0.795	—	26.7	0.631
Exp2 → Exp1, PG-PAMTL-DA	0.822	0.960	—	38.3	0.837

3-4- Corrosion-Severity Estimation

Interpretation revision: Although domain adaptation reduced the severity estimation MAE, this improvement should not be interpreted as proof of highly accurate absolute corrosion quantification. The negative or limited R^2 values indicate that the model captures only part of the severity variation across experiments. Therefore, regression results are interpreted as estimation of corrosion-related trends rather than a fully calibrated physical measurement system.

Cross-experiment severity estimation was more difficult than binary corrosion detection. Domain adaptation nevertheless improved the regression results: PG-PAMTL-DA reduced macro MAE from 39.3 to 32.5 μm and increased macro R^2 from −0.357 to 0.105, corresponding to a 17.3% reduction in error. Fig. 6(a,b) summarizes this overall change. The two transfer directions behaved differently. Exp1→Exp2 produced the lower MAE (26.7 μm), whereas Exp2→Exp1 showed the stronger rank correlation (ρ = 0.529).

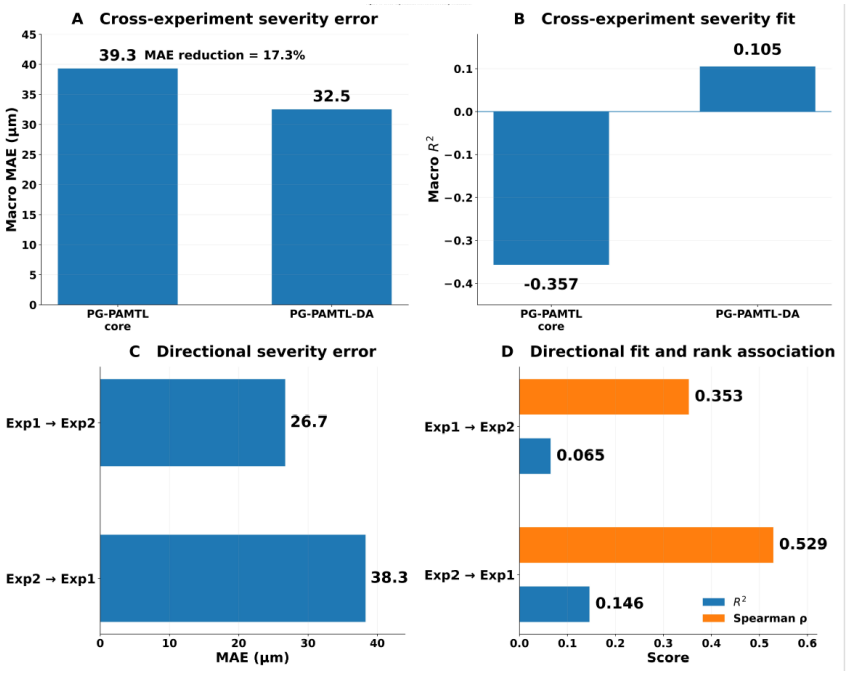


Fig. 6 Cross-experiment corrosion-severity estimation: (a) macro MAE before and after domain adaptation; (b) corresponding macro R^2; (c) directional MAE for Exp1→Exp2 and Exp2→Exp1; and (d) directional R^2 and Spearman correlation.

The regression outputs therefore captured part of the overall severity trend, but not the equivalent pit radius with high absolute accuracy across independent experiments. The low cross-experiment R^2 values are an important limitation: they support interpreting the severity output as a trend estimate rather than a precise physical measurement. Domain adaptation improved stability, but accurate absolute severity quantification remains a more demanding problem.

3-5- Ablation and Statistical Validation

The ablation study was used to identify which components drove the observed gains. Replacing the flat MLP with path attention increased mean ROC-AUC from 0.464 ± 0.093 to 0.757 ± 0.140 and reduced severity MAE from 69.2 ± 12.9 to 41.8 ± 9.3 µm. Adding the physics-guided component improved the results further, reaching ROC-AUC = 0.837 ± 0.029 and MAE = 34.3 ± 7.4 µm. The largest step therefore came from representing the sensing paths explicitly, while the physics-guided formulation provided an additional improvement in both accuracy and run-to-run stability. The bootstrap comparison puts these gains into context. PG-PAMTL-DA improved ROC-AUC clearly relative to the conventional benchmark models, whereas the increase over the core PG-PAMTL model was only 0.007. The effect of domain adaptation was more apparent for severity estimation, where it reduced cross-experiment error. This distinction is important: the architecture itself accounts for most of the cross-experiment detection gain, while domain adaptation contributes mainly to regression stability rather than a large increase in classification discrimination.

3-6- Progression-Confounding Analysis

Additional discussion: The State-only baseline demonstrates that experimental progression is strongly correlated with corrosion severity in the available dataset. Consequently, severity prediction results should be interpreted cautiously. The present analysis supports the presence of corrosion-sensitive information in Lamb-wave features, but additional experiments with randomized corrosion progression are required to fully separate damage sensitivity from temporal progression effects.

Corrosion increased monotonically during both experiments, so experimental progression could itself carry substantial information about the targets. A state-only baseline was therefore used as a negative control. As shown in Fig. 8(a), state alone produced ROC-AUC = 1.000 in both transfer directions, compared with 0.843 for Exp1→Exp2 and 0.822 for Exp2→Exp1 using the Lamb-wave model. The contrast was even larger for severity: the state-only model achieved MAEs of 6.5 and 11.1 µm, whereas the Lamb-wave model produced 26.7 and 38.3 µm. Adding Lamb-wave information to state did not consistently improve the state-only baseline, and the same pattern was reflected in R^2. These results identify experimental progression as a major confounder, particularly for quantitative severity estimation. State was never used as a predictor in PG-PAMTL, but the monotonic experimental design means that part of the apparent predictive structure can still track sequence progression. Cross-experiment detection is therefore the stronger result of this study; severity estimation requires validation in experiments where corrosion level and experimental order are varied more independently.

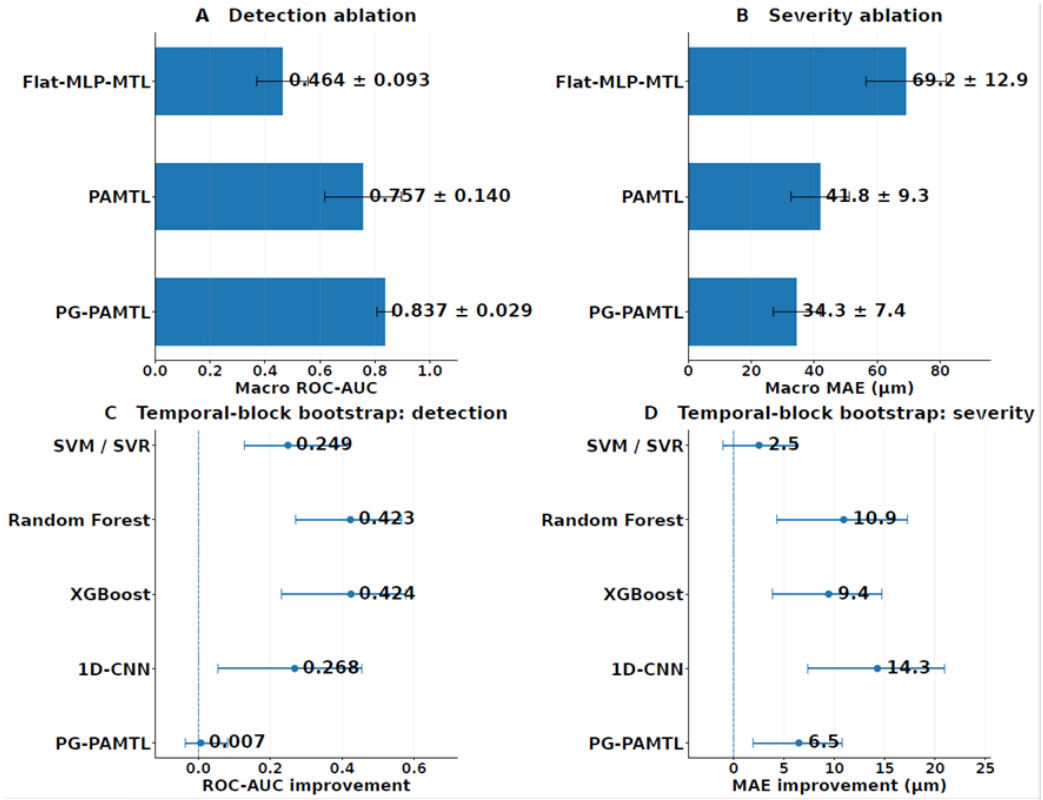


Fig. 7 Ablation and statistical validation of the proposed framework: (a) multi-seed detection ablation based on macro ROC-AUC; (b) severity ablation based on macro MAE; (c) temporal-block bootstrap comparison of ROC-AUC improvements; and (d) corresponding improvements in severity MAE.

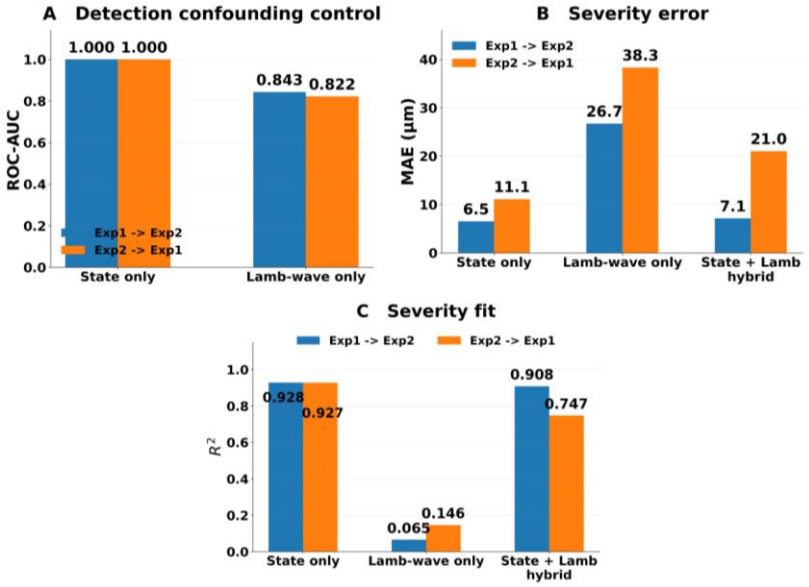


Fig. 8. Progression-confounding analysis: (a) comparison of state-only and Lamb-wave-based corrosion detection; (b) severity MAE for state-only, Lamb-wave-only, and hybrid models; and (c) corresponding R^2 values for severity estimation.

3-7- Model Interpretability: Path and Feature Importance

Permutation importance was used to determine which sensing paths and features the model relied on most. Fig. 9(a) shows that PZT3→PZT1 and PZT1→PZT3 were the most influential paths, with mean ROC-AUC drops of 0.088 and 0.068, respectively. The remaining paths had smaller effects, indicating that the network did not use all actuator–sensor combinations equally. Feature importance was likewise concentrated rather than uniform. DI diffrms produced the largest AUC drop (0.075), followed by band-power features in the 220–300 kHz and 300–500 kHz ranges and by envelope correlation. Baseline-relative residual, spectral centroid, and sensor energy also contributed. The model therefore drew on a mixture of waveform-change and spectral information rather than a single feature family.

3-8- Artifact and Bubble-Sensitivity Analysis

Additional discussion: The bubble-related disturbance is considered an experimental artifact affecting domain variability. However, the observed performance change after removing this interval is not interpreted as a causal explanation of transfer success. Instead, it demonstrates that uncontrolled experimental disturbances can influence cross-domain performance and should be considered in future validation studies. Experiment 2 contains a known bubble-related disturbance, so its effect on transfer performance was examined separately. For the Exp1→Exp2 transfer, ROC-AUC increased from 0.843 to 0.995 when the affected interval was removed. A class-matched reference helps interpret this change: across 5000 random exclusions with similar class composition, the mean ROC-AUC remained near 0.843 and the 97.5th percentile was 0.936. The much larger value obtained after excluding the bubble interval therefore cannot be explained by class composition alone.

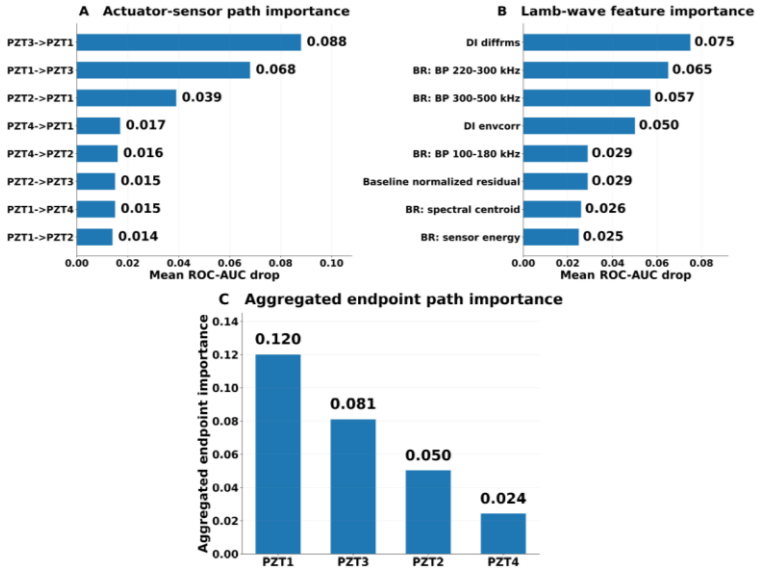


Fig. 9 Model interpretability based on permutation importance: (a) actuator–sensor path importance expressed as the mean ROC-AUC drop; (b) importance of the leading Lamb-wave features; and (c) aggregated endpoint path importance across the four PZT locations.

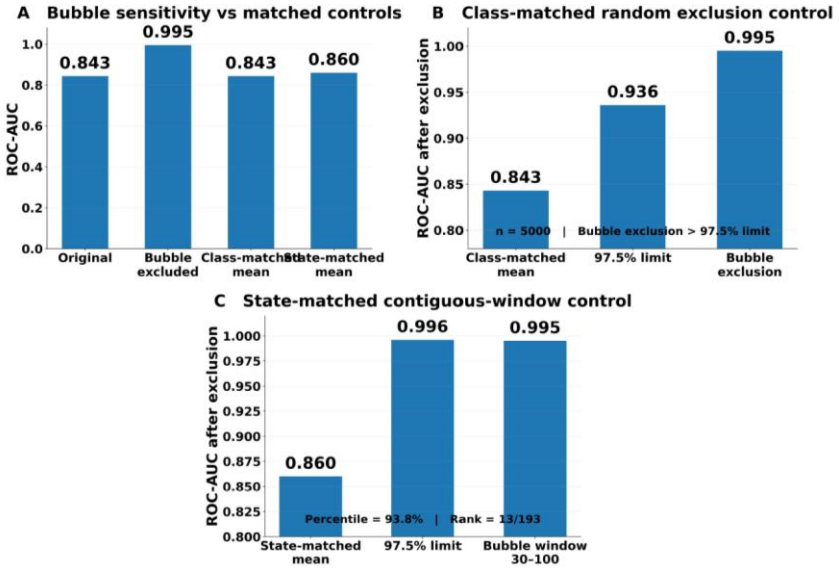


Fig. 10 Artifact and bubble-sensitivity analysis: (a) ROC-AUC before and after excluding the bubble-affected interval, together with matched-control means; (b) class-matched random exclusion control; and (c) state-matched contiguous-window control.

The contiguous-window control gives a more cautious picture. The bubble interval ranked 13th among 193 state-matched windows, corresponding to the 93.8th percentile, while the 97.5th-percentile reference for contiguous exclusions was 0.996. The disturbed segment was therefore influential, but not uniquely exceptional among all temporal windows. It is appropriate to treat the bubble event as an important experimental artifact without attributing all cross-experiment variability to that interval alone.

4- Conclusion

Revision of conclusion scope: The findings demonstrate promising cross-experiment corrosion detection capability within the investigated COQTEL dataset. However, broader claims regarding universal deployment or general transferability require validation using additional specimens, independent laboratories, and different environmental conditions. This study examined whether Lamb-wave measurements and machine learning can provide a corrosion-assessment strategy that remains informative beyond the experiment used for model development. Two independent laboratory experiments on 316L stainless steel made it possible to compare the conventional within-experiment setting with the more demanding problem of experiment-to-experiment transfer. The strongest result concerned corrosion detection. In strict zero-shot transfer, PG-PAMTL achieved a macro ROC-AUC of 0.830 and a PR-AUC of 0.871, retaining substantially more discriminative information than the conventional benchmarks. Unsupervised domain adaptation increased ROC-AUC only slightly, to 0.832, but had a clearer effect on severity estimation, reducing macro MAE from 39.3 to 32.5 µm. The interpretability analysis further showed that predictive information was concentrated in a limited number of sensing paths and in a combination of waveform-change and frequency-domain descriptors. The severity results require more caution. Experimental progression was strongly associated with corrosion severity, and the state-only control outperformed the Lamb-wave model for quantitative estimation. The bubble-related disturbance in Experiment 2 also affected one transfer direction. The present regression results should therefore not be treated as an independent absolute measurement of corrosion severity. Future validation should decouple corrosion level from experimental order, ideally through randomized or non-monotonic progression, repeated specimens, broader operating conditions, and independent tests. Such experiments would provide a stronger basis for assessing whether the proposed framework can support reliable cross-domain corrosion monitoring and, eventually, more accurate quantitative damage assessment.

Artificial Intelligence Statement: Generative AI tools were used for language editing and formatting assistance only. The author reviewed the manuscript and remains responsible for the scientific content, analyses, interpretations, and conclusions.

Ethics Approval: The scientific content of this article is the result of the authors' research and has not been published in any Iranian or international journal.

Acknowledgements: The authors have no acknowledgements to declare.

Funding: This research received no external funding.

CRediT Authorship Contribution: Mohammad Khodadosti: Conceptualization, Methodology, Programming and code implementation, Formal analysis, Investigation, Data curation, Visualization, and Writing - original draft. Abdolrahim Taheri: Validation and manuscript review. All authors read and approved the final manuscript.

Conflict of Interest: The authors declare no conflict of interest.

References

[1] M. S. B. Reddy, D. Ponnamma, K. K. Sadasivuni, S. Aich, S. Kailasa, H. Parangusan, M. Ibrahim, S. Eldeib, O. Shehata, M. Ismail, et al., "Sensors in advancing the capabilities of corrosion detection: A review," Sensors and Actuators A: Physical, vol. 332, Art. no. 113086, 2021, doi: 10.1016/j.sna.2021.113086.

[2] V. Vasagar, M. K. Hassan, A. M. Abdullah, A. V. Karre, B. Chen, K. Kim, N. Al-Qahtani, and T. Cai, "Non-destructive techniques for corrosion detection: A review," Corrosion Engineering, Science and Technology, vol. 59, no. 1, pp. 56–85, 2024, doi: 10.1177/1478422X241229621.

[3] C. R. Farrar and K. Worden, "An introduction to structural health monitoring," Philosophical Transactions of the Royal Society A, vol. 365, no. 1851, pp. 303–315, 2007, doi:10.1098/rsta.2006.1928.

[4] M. Mitra and S. Gopalakrishnan, "Guided wave based structural health monitoring: A review," Smart Materials and Structures, vol. 25, no. 5, Art. no. 053001, 2016, doi: 10.1088/0964-1726/25/5/053001.

[5] P. Cawley, "Guided waves in long range nondestructive testing and structural health monitoring: Principles, history of applications and prospects," NDT & E International, vol. 142, Art. no. 103026, 2024, doi: 10.1016/j.ndteint.2023.103026.

[6] Z. Su, L. Ye, and Y. Lu, "Guided Lamb waves for identification of damage in composite structures: A review," Journal of Sound and Vibration, vol. 295, nos. 3–5, pp. 753–780, 2006, doi: 10.1016/j.jsv.2006.01.020.

[7] R. Gorgin, Y. Luo, and Z. Wu, "Environmental and operational conditions effects on Lamb wave based structural health monitoring systems: A review," Ultrasonics, vol. 105, Art. no. 106114, 2020, doi: 10.1016/j.ultras.2020.106114.
[8] A. J. Croxford, J. Moll, P. D. Wilcox, and J. E. Michaels, "Efficient temperature compensation strategies for guided wave structural health monitoring," Ultrasonics, vol. 50, nos. 4–5, pp. 517–528, 2010, doi: 10.1016/j.ultras.2009.11.002.
[9] P. B. Nagy, F. Simonetti, and G. Instanes, "Corrosion and erosion monitoring in plates and pipes using constant group velocity Lamb wave inspection," Ultrasonics, vol. 54, no. 7, pp. 1832–1841, 2014, doi: 10.1016/j.ultras.2014.01.017.
[10] T. R. Hay, R. L. Royer, H. Gao, X. Zhao, and J. L. Rose, "A comparison of embedded sensor Lamb wave ultrasonic tomography approaches for material loss detection," Smart Materials and Structures, vol. 15, no. 4, p. 946, 2006, doi: 10.1088/0964-1726/15/4/007.
[11] P. Belanger and P. Cawley, "Feasibility of low frequency straight-ray guided wave tomography," NDT & E International, vol. 42, no. 2, pp. 113–119, 2009,doi: 10.1016/j.ndteint.2008.10.006.
[12] W. Li and Y. Cho, "Quantification and imaging of corrosion wall thinning using shear horizontal guided waves generated by magnetostrictive sensors," Sensors and Actuators A: Physical, vol. 232, pp.251–258,2015,doi: 10.1016/j.sna.2015.06.008.
[13] Z. Zhang, H. Pan, X. Wang, and Z. Lin, "Machine Learning-Enriched Lamb Wave Approaches for Automated Damage Detection," Sensors, vol. 20, no. 6, Art. no. 1790, 2020, doi: 10.3390/s20061790.
[14] C. Schnur, P. Goodarzi, Y. Lugovtsova, J. Bulling, J. Prager, K. Tschöke, J. Moll, A. Schütze, and T. Schneider, "Towards interpretable machine learning for automated damage detection based on ultrasonic guided waves," Sensors, vol. 22, no. 1, Art. no. 406, 2022, doi: 10.3390/s22010406.
[15] A. Abbassi, N. Römgens, F. F. Tritschel, N. Penner, and R. Rolfes, "Evaluation of machine learning techniques for structural health monitoring using ultrasonic guided waves under varying temperature conditions," Structural Health Monitoring, vol. 22, no. 2, pp. 1308–1325, 2023, doi:10.1177/14759217221107566.
[16] M. Rautela, J. Senthilnath, J. Moll, and S. Gopalakrishnan, "Combined two-level damage identification strategy using ultrasonic guided waves and physical knowledge assisted machine learning," Ultrasonics, vol. 115, Art. no. 106451, 2021, doi:10.1016/j.ultras.2021.106451.
[17] X. Wang, M. Lin, J. Li, J. Tong, X. Huang, L. Liang, Z. Fan, and Y. Liu, "Ultrasonic guided wave imaging with deep learning: Applications in corrosion mapping," Mechanical Systems and Signal Processing, vol. 169, Art. no. 108761, 2022, doi:10.1016/j.ymssp.2021.108761.
[18] Y. Ran, C. Qian, X. Wang, W. Zhang, and R. Wang, "Corrosion Monitoring in Automotive Lap Joints Based on Imaging Methods of Lamb Waves," Sensors, vol. 24, no. 24, Art. no. 8092, 2024, doi: 10.3390/s24248092.
[19] L. Huang, Z. Luo, L. Zeng, and J. Lin, "Detection and localization of corrosion using the combination information of multiple Lamb wave modes," Ultrasonics, vol. 138, Art. no. 107246, 2024, doi: 10.1016/j.ultras.2024.107246.
[20] H. Sun, W. Shao, J. Song, X. Yang, Y. Wang, and X. Qing, "Corrosion quantification of plate-type structures using Lamb wavefield and monogenic signal processing," Ultrasonics, vol. 130, Art.no.106935,2023, doi: 10.1016/j.ultras.2023.106935.
[21] S. Sawant, A. Sethi, S. Banerjee, and S. Tallur, "Unsupervised learning framework for temperature compensated damage identification and localization in ultrasonic guided wave SHM with transfer learning," Ultrasonics, vol. 130, Art. no. 106931, 2023, doi:10.1016/j.ultras.2023.106931.
[22] V. Nerlikar, O. Mesnil, R. Miorelli, and O. D'Almeida, "Damage detection with ultrasonic guided waves using machine learning and aggregated baselines," Structural Health Monitoring, vol. 23, no. 1, pp. 443–462, 2024, doi: 10.1177/14759217231169719.
[23] J. Li, M. Wang, H. Chen, P. Wang, and W. Wang, "Investigation on characteristics of tensile damage and microstructure evolution of steel AISI 316L by nonlinear ultrasonic Lamb wave," International Journal of Pressure Vessels and Piping, vol. 198, Art. no. 104680, 2022, doi: 10.1016/j.ijpvp.2022.104680.
[24] J. Hua, X. Cao, Y. Yi, and J. Lin, "Time-frequency damage index of Broadband Lamb wave for corrosion inspection," Journal of Sound and Vibration, vol. 464, Art. no. 114985, 2020, doi: 10.1016/j.jsv.2019.114985.
[25] V. T. Rathod and D. Roy Mahapatra, "Ultrasonic Lamb wave based monitoring of corrosion type of damage in plate using a circular array of piezoelectric transducers," NDT & E International, vol. 44, no. 7, pp. 628–636, 2011, doi: 10.1016/j.ndteint.2011.07.002.
[26] D. Wang, W. Zhang, X. Wang, and B. Sun, "Lamb-Wave-Based Tomographic Imaging Techniques for Hole-Edge Corrosion Monitoring in Plate Structures," Materials, vol. 9, no. 11, Art. no. 916, 2016, doi: 10.3390/ma9110916.
[27] Z. Guo, R. Zhou, Y. Gao, W. Fu, and Q. Yu, "Towards ultrasonic guided wave fine-grained damage detection on hierarchical multi-label classification network," Mechanical Systems and Signal Processing, vol. 218, Art. no. 111582, 2024, doi: 10.1016/j.ymssp.2024.111582.
[28] Y. Shen, J. Wu, J. Chen, W. Zhang, X. Yang, and H. Ma, "Quantitative Detection of Pipeline Cracks Based on Ultrasonic Guided Waves and Convolutional Neural Network," Sensors, vol. 24, no. 4, Art. no. 1204, 2024, doi: 10.3390/s24041204.
[29] C. Nicard, M. Rébillat, O. Devos, M. El May, F. Letellier, S. Dubent, M. Thomachot, M. Fournier, P. Masse, and N. Mechbal, "In-situ monitoring of µm-sized electrochemically generated corrosion pits using Lamb waves managed by a sparse array of piezoelectric transducers," Ultrasonics, vol. 147, Art. no. 107527, 2025, doi: 10.1016/j.ultras.2024.107527.
[30] H. Lu, S. Cantero-Chinchilla, B. Rotsaert, A. Croxford, K. Gryllias, and D. Chronopoulos, "Damage identification using ultrasonic Lamb waves with multi-scale adaptive attention Transformer-based unsupervised domain adaptation," Mechanical Systems and Signal Processing, vol. 234, Art. no. 112772, 2025, doi:10.1016/j.ymssp.2025.112772.
[31] C. Nicard, M. Rébillat, O. Devos, M. El May, F. Letellier, S. Dubent, M. Thomachot, M. Fournier, P. Masse, and N. Mechbal, "COQTEL project dataset: Corrosion quantification trough extended use of Lamb waves," Data in Brief, vol. 59, Art. no. 111393, 2025, doi: 10.1016/j.dib.2025.111393.
[32] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016, doi:10.1145/2939672.2939785.
[34] R. Caruana, "Multitask learning," Machine Learning, vol. 28, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.
[35] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in Computer Vision – ECCV 2016 Workshops, pp. 443–450, 2016, doi:10.1007/978-3-319-49409-8_35.
[36] H. R. Künsch, "The jackknife and the bootstrap for general stationary observations," The Annals of Statistics, vol. 17, no. 3, pp. 1217–1241, 1989, doi:10.1214/aos/1176347265.
[37] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340–1347, 2010, doi:10.1093/bioinformatics/btq134.
[38] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023, doi:10.1561/2200000101.