

Machine learning–driven flaw detection for ultrasonic pipe inspections with acoustic lens

Thiago E. Kalid^{1,*}, André E. Lazzaretti¹, Tatiana A. Prado¹, Gustavo P. Pires², Daniel R. Pipa¹ and Thiago A. R. Passarin¹.

¹ Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brasil. ² Centro de Pesquisa Leopoldo Américo Miguez de Mello (CENPES), Rio de Janeiro, Brasil.

*Corresponding author, E-Mail: passarin@utfpr.edu.br

Abstract

Acoustic lenses have recently been employed during ultrasonic pipe inspection to increase the phased-array transducer's coverage area. However, the lens also introduces geometric artifacts that affect the flaw detection process. Classic flaw detection methodologies commonly rely on thresholding, which becomes problematic when geometric artifacts closely match the desired signal amplitude. We propose an adaptive method for flaw detection in ultrasonic inspections of pipes using acoustic lenses, based on anomaly detection techniques. A machine-learning model was trained on data from flawless pipe specimens using an acoustic lens. During training, the model captures patterns associated with the flawless pipe and lens. Then, during testing, it analyzes newly observed data and classifies it either as a flaw (anomaly) or a normal response (as in the training set). Our model, based on the Local Outlier Factor, achieved an area under the receiver operating characteristic curve of 0.96, a global accuracy of 0.93, a recall of 0.93, and an F2-score of 0.82. These results, obtained across a wide range of flaw geometries, suggest that our model could be part of an automated flaw detection system for pipeline inspections.

Keywords: Non-destructive testing (NDT); Acoustic lens; Ultrasonic testing (UT); Machine Learning (ML); Automated defect detection.

1 Introduction

Pipes are an essential infrastructure of the Oil & Gas industry for transporting fluids. Internal corrosion is among the leading factors that reduce their lifespan [1]. For subsea pipes, ultrasonic non-destructive testing (NDT) is commonly used to externally assess corrosion levels. When the inner surface of the pipe exhibits abrupt thickness changes, such as pitting corrosion, the wall thickness must be measured over a large portion of the pipe with a high angular density (Fig. 1a). These requirements significantly increase the duration and complexity of field inspections [2].

Acoustic lenses have been recently proposed to increase the coverage area of phased-array transducers in pipe inspections [3]. It enables the static inspection of a large section of the pipe (i.e., 90 degrees, shown in Fig. 1b) while projecting the sectorial scan (S-scan) into cylindrical coordinates (Fig. 1c). This device ultimately makes the inspection faster and easier to interpret. However, the acoustic lens also introduces geometric artifacts from internal reflections (indicated in Fig. 1c).

Geometrical artifacts exhibit a consistent response across multiple inspections because they mainly depend on the acoustic lens geometry. When the amplitude of these artifacts is comparable to the signal of interest (such as a flaw), flaw detectability is compromised [4]. This issue is especially problematic for classical flaw detection methods that rely on local thresholding [5].

This paper presents a method for flaw detection based on anomaly detection and machine learning. The model is trained on flawless pipe samples and associates these patterns

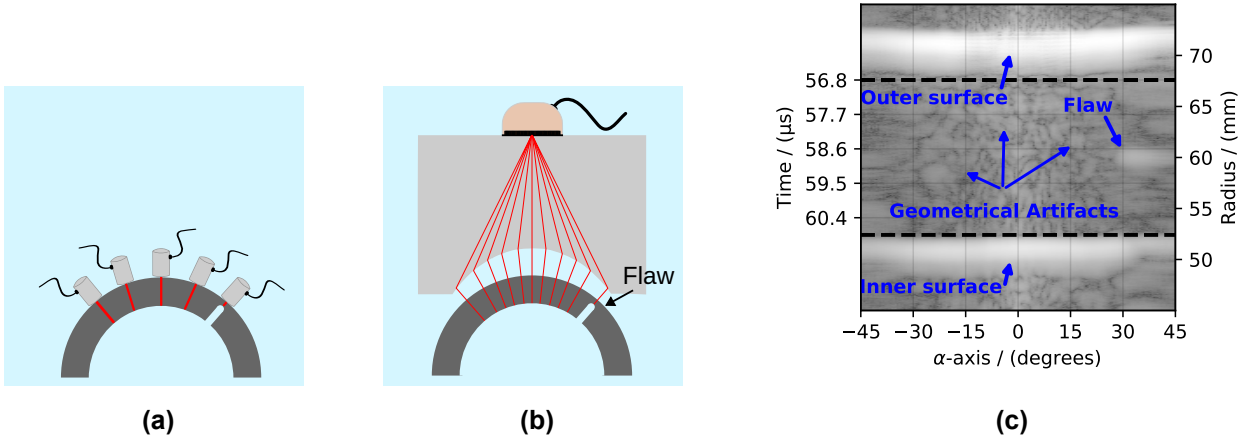


Fig. 1. Thickness measurement by different ultrasonic immersion testing. (a) Spot measurement using a single-element transducer. The ability to measure the correct remaining thickness with this method depends heavily on the inspection position, hence it is very inefficient at detecting localized flaws. (b) Sectorial scanning using an acoustic lens coupled to a phased-array transducer. The red lines represent the paths of the ultrasonic rays. (c) Typical sectorial scan projected to cylindrical coordinates by the acoustic lens. The pipe’s inner and outer surfaces appear as constant horizontal echoes due to their constant radius.

with a normal response. Then, during testing, it analyzes the new point and, using a novelty detection algorithm, classifies it as an anomaly or a normal response (as in the training set).

The rest of this paper is organized as follows. In Section 2, we explain the steps involved in building our dataset, as well as the modeling approach for flaw detection. In Section 3, we quantitatively compare results from multiple classification algorithms and discuss their improvements over the classical method. Finally, in Section 4 we summarize our findings, and briefly discuss improvements to the present work.

2 Methods

2.1 Dataset

We created a dataset by inspecting several pipe and half-pipe samples using a single acoustic lens model. Since the acoustic response is affected by geometric artifacts (which ultimately depend on the specific acoustic lens geometry), the dataset was tailored to support the detection of flaws with one particular geometry, namely that used in the seminal paper [3]. The methodology, on the other hand, is applicable to any similar acoustic lens geometry but requires retraining.

The flaw detection was based on S-scan images reconstructed from refraction-corrected beamforming. Since the acoustic lens projects the S-scan into a cylindrical coordinate system, each column of the image corresponds to a pipe sector, while each row corresponds to a radial location. The delay law required to properly shape the wavefront and reconstruct the S-scan is based on convergent waves. They were aimed at 181 evenly spaced foci within the interval $\alpha \in [-\pi/4, +\pi/4]$, located at the inner surface of the pipe.

We collected and annotated 1136 S-scans from ultrasonic mechanical sweeps of four pipe specimens. Each specimen could contain machined flaws of different shapes and sizes (e.g., side-drilled holes, spherical-bottom holes), which were segmented and labeled as a single class “anomaly”. The pipes are made of carbon steel ($c_l = 5900$ m/s), with an outer diameter ranging from 139.6 mm to 141.0 mm, and a wall thickness ranging from 16.3 mm to 20.7 mm. The specimens were carefully selected to reflect common values found in pipes used in subsea Christmas trees for water reinjection [2], which is the final inspection application.

The ultrasonic data were collected using linear phased-array scanning with a 64-element

Olympus® 5 MHz phased-array probe. The acquisition system was Panther™, by Eddify Technologies, set to drive the probe at 100 V and a receiver gain of 35 dB at a sampling frequency of 125 MHz.

2.2 Classification Model

Our approach is based on classical machine learning models. This choice was driven mainly by the size of our dataset, which is relatively small compared to similar anomaly detection tasks [6]. If we instead employed high-capacity models such as those used in deep learning, we could end up with poor generalization due to overfitting. On top of that, effective data augmentation techniques in NDT ultrasonic inspections remain an open problem, as they are too complex and resource-intensive [7].

Fig. 2 summarizes the steps involved in extracting the feature vectors from the input raw ultrasonic data. The input is composed by the reception signals acquired during an electronic steering inspection, denoted by $I''_n \in \mathbb{R}^{N'_t \times N_\alpha \times N_{rx}}$, where N'_t is the number of time-domain samples, N_α is the number of steering angles (or pipe sectors, since we are using an acoustic lens), and N_{ch} is the number of active channels. The subscript denotes the n -th acquisition, where $n = 1, 2, \dots, N_{shots}$. From the previously computed reception delay-law, the channels are delayed and summed to generate the n -th S-scan, $I'_n \in \mathbb{R}^{N'_t \times N_\alpha \times 1}$, which is cropped to a smaller region-of-interest (ROI) $I_n \in \mathbb{R}^{N_t \times N_\alpha \times 1}$, where $N_t \leq N'_t$. The ROI is set between the pipe walls, where flaws are most likely to occur. We split I_n into evenly spaced tiles $T_j^{(n)} \in \mathbb{R}^{N_x \times N_y \times 1}$ to be able to identify the anomaly spatial position, where N_x and N_y are the tile vertical and horizontal size in number of pixels. There were N_h evenly spaced tiles along the horizontal direction, and N_v along the vertical direction, resulting in $j = 1, 2, \dots, (N_h \cdot N_v)$.

The image brightness feature space is inherently high-dimensional; thus, we used a set of image descriptors to reduce the dimensionality of the problem. The features were: time-domain brightness mean, standard deviation, maximum, minimum, peak-to-peak, and median; frequency-domain mean, median, maximum, and minimum location; and the first 50 components obtained by Principal Component Analysis (PCA). Statistical features (in the time and frequency domains) are useful for quantifying signal characterization and are commonly employed to identify material irregularities and properties in ultrasonic inspections [5]. In addition, PCA is a powerful technique for uncovering data's latent structure and has been shown to be effective when incorporated into machine-learning models applied to ultrasound data [5]. Finally, each tile position was assigned a unique label, which was one-hot encoded to ensure that the model accounted for regional differences in the geometrical artifacts.

Given our particular experimental setup, $N_t = 675$ samples, $N_{ch} = 64$ channels, $N_\alpha = 181$ angles, $N_h = 19$ tiles and $N_v = 9$ tiles and $N_{shots} = 1136$ S-scans, resulting in 194256 feature vectors $\mathbf{x} \in \mathbb{R}^{231 \times 1}$, where each vector represent an unique tile to be classified.

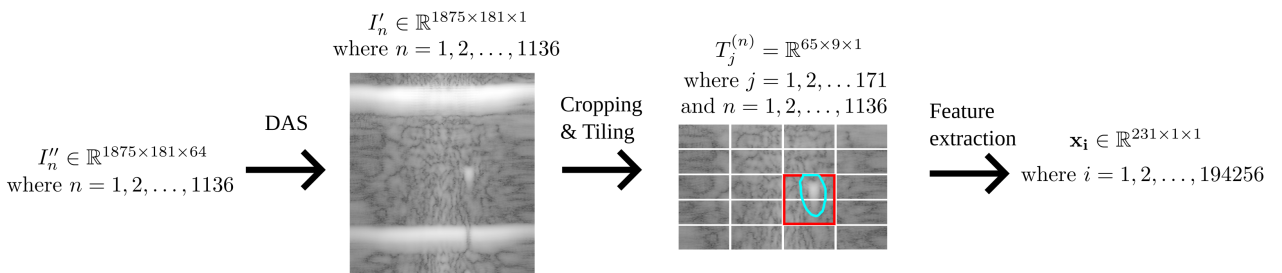


Fig. 2. Data pre-processing steps. The cyan closed-path represents the manually segmented ground truth. Based on the segmentation, red-framed tiles were labeled as anomalous. After cropping and tiling, the image is represented with fewer tiles just for illustration purposes.

In the inspection of pipes prone to pitting corrosion, the size and shape may assume very

distinct values [1]. For this reason, it is impossible to create a dataset with every possible shape and size of the expected flaw. Instead, it is easier for us to create a dataset from pristine pipe specimens and train a model to detect any deviation from normality. Therefore, in this work, the flaw detection was performed through novelty-detection algorithms.

The original annotations were pixel-wise anomaly segmentation (cyan closed-path in Fig. 2). Our model is a single-class novelty-detection classifier; thus, we had to convert these annotations into a binary "anomaly" label for each tile. For a given anomaly segmentation instance, if 1 % or more of its area overlapped a given tile $T_j^{(n)}$, the tile was labeled as anomalous (represented by red-framed tiles in Fig. 2), and normal otherwise.

Novelty detection models are trained on an anomaly-free dataset and capture patterns associated with normality; hence, they are classified as semi-supervised approaches [8]. During testing, the trained model will compare the newly observed feature with the learned normality pattern. If the mismatch is too large, the data point will be classified as anomalous; otherwise, it will be classified as normal. We used the implementation of novelty detection algorithms available in PyOD [8]—a Python library focused on outlier and novelty detection.

We adopted the following split of our dataset: 70 % for training, 15 % for testing, and 15 % for validation. The anomalous data points were proportionally and randomly split into validation and test sets, leaving the training set anomaly-free. The normal samples are randomly and proportionally split among all sets.

Novelty detection algorithms require a contamination rate to set the classification threshold. We performed hyperparameter tuning through a grid search with cross-validation, where the contamination was searched within the interval [1 %, 10 %], and the parameter that yielded the best F_2 -score was chosen. During inference, whenever the model yielded a low-confidence prediction, we opted for a conservative approach: assigning the data point as anomalous. Normally, in NDT field inspections, false negatives (FN , i.e., predicting that a sample is normal when, in fact, it is anomalous) pose a more dangerous problem than false positives (FP , i.e., the opposite mistake).

3 Results and Discussions

Multiple anomaly detection models based on machine learning were trained and tested on the same dataset. We computed five metrics to quantitatively compare them: accuracy, precision, recall, F_β -score, and area under the receiver operating characteristic curve (ROC AUC). The first four are defined as:

$$\text{Acc.} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Prec.} = \frac{TP}{TP + FP}, \quad \text{Rec.} = \frac{TP}{TP + FN}, \quad F_\beta = (1 + \beta^2) \frac{\text{Prec. Rec.}}{\beta^2 \text{Prec.} + \text{Rec.}},$$

in which TP stands for true positives (predicting a tile anomalous when in fact it is), TN is true negatives (predicting normal when in fact it is). In particular, we will assume $\beta = 1$ or $\beta = 2$ in our discussions. This choice was to evaluate our model under two scenarios: a conservative approach, where FN is more critical ($\beta = 2$); and a more balanced approach, where FP and FN are equally important ($\beta = 1$).

Additionally, a local threshold-based flaw detection method was implemented as a classical baseline. It detects the S-scan maximum pixel intensity in the region of interest and interprets anomalous tiles as those that contain a significant number of pixels above -6 dB of the maximum. Even though there have been significant technological advances in NDT ultrasound inspections, this method is still widely employed [9]. A particular inspection and its qualitative result are shown in Fig. 3a and 3b. The quantitative performance of the top three models (plus the classic threshold-based) is summarized in Table 1.

The best-performing model was based on the Local Outlier Factor (LOF). This model identified flaws across different sizes, shapes, and amplitudes. For instance, the two flaws shown in Fig. 3b were correctly identified by the model, while the threshold-based method

Table 1: Performance comparison of different methods.

Model	Accuracy	Precision	Recall	F_2 -Score	ROC AUC
Local Outlier Factor	0.934	0.567	0.927	0.823	0.960
K-Nearest Neighbors	0.918	0.510	0.745	0.682	0.935
Isolation Forest	0.887	0.343	0.369	0.363	0.768
Threshold	0.913	0.481	0.364	0.382	0.664

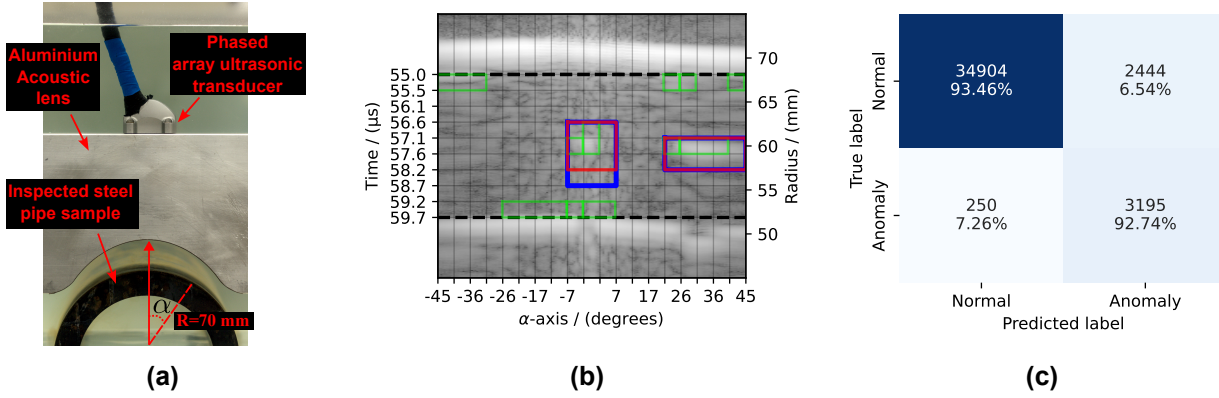


Fig. 3. A pipe inspection for flaw detection with an acoustic lens. (a) An example of the experimental setup; (b) Corresponding sectorial scan with proposed (blue), classical threshold-based (green) flaw detection techniques and ground-truth (red); (c) Confusion matrix of the best-performing method, based on Local Outlier Factor, tested on the entire test set. The confusion matrix data points refer to individual tiles.

misclassified normal tiles as anomalous. This performance gap is even larger for low-intensity flaws; in our approach, the model already knows how the geometrical artifacts should behave, thus it does not require local references to infer whether a given feature is anomalous. The threshold-based method, on the other hand, always relies on local comparison, therefore if the maximum local amplitude is low, more tiles will be above the -6 dB threshold and classified as anomalous.

As a side effect of the novelty detection modeling approach, any unseen behavior in the training set was classified as anomalous by our model. For instance, if poor positioning, air bubbles, or strong high-frequency noise result in unseen imaging artifacts within the ROI, they will be classified as anomalies even though they are not real flaws. This, along with conservative decisions regarding the model inference and hyperparameter tuning, led to high recall (0.93) but at a cost of poor precision (0.57). If we adopted less conservative decisions, such as adopting the F1-score on hyperparameter tuning, the LOF-based model would yield $Acc. = 0.968$, $Prec. = 0.806$, $Rec. = 0.820$, and $F_1 = 0.813$.

The training and testing were performed on an Intel(R) Core(TM) i7-14700 processor with 16 GB RAM. Although the LOF-based model yielded the best metrics, it was the slowest for inference (420 ms per S-scan, or 2.4 Hz) and required substantial memory (272 MB). The classic k-nearest-neighbor method yielded a faster inference time (95 ms per S-scan or 10.5 Hz) but at a cost of a higher model size (521 MB). The high memory requirement is a known issue with nearest-neighbor classification algorithms, since they classify a newly observed point by comparing it individually with the points stored in the training set. Isolation forests, on the other hand, classify new points by compactly representing the original dataset, yielding lower memory requirements (2.8 MB) and lower inference runtime (9 ms per S-scan or 110 Hz). The classic threshold-based method does not require training and relies on simple mathematical operations during inference; therefore, it requires virtually no memory storage and yields the fastest runtime (0.49 ms per S-scan or 2040 Hz).

The dataset and source code necessary to generate the results presented in this paper are openly available at <https://doi.org/10.5281/zenodo.19410832>.

4 Conclusions

In this work, we proposed a flaw detection method for pipe ultrasonic inspections performed with an acoustic lens. The algorithm is a machine learning model based on novelty detection that, after being trained on pristine pipe specimens, successfully detected 93 % of flaws, regardless of their size and shape. Our method splits the region of interest into smaller tiles and determines whether each tile contains an anomaly. Our best-performing model, based on LOF, represents a significant 118 % improvement in F_2 -score over the classic threshold-based method.

Since the dataset presented contains a limited number of machine specimens, the model's generalization to real-world field inspections remains worth evaluating. Taking these considerations into account, our model achieved strong performance with our dataset, especially compared to the classic threshold-based method. We believe that, in the future, the model could be a significant part of an autonomous or semi-autonomous pipe inspection system, aiding in day-to-day operational NDT inspection decision-making.

Acknowledgements

This work was funded by Petrobras/CENPES through the R&D project AUSPEX2 (SIG-ITEC 2022/00436-2) and the Brazilian National Council for Scientific and Technological Development (CNPq) through grant number 305075/2025-0. The authors thank UTFPR, CAPES, Fundação Araucária, FINEP, and CNPq for the scholarships and additional funding.

References

1. Mubarak, G., Verma, C., Barsoum, I., et al.: Internal corrosion in oil and gas wells during casings and tubing: challenges and opportunities of corrosion inhibitors, *Journal of the Taiwan Institute of Chemical Engineers* **150**, 105027 (2023).
2. Urbano, J. P., Pires, G. P., Santos, M. D., et al.: Melhoria de medição de espessura submarina por ultrassom via ROV, pt, [Revista Técnica da Universidade Petrobras](#) (2), 1–9 (2025).
3. Pires, G. P., Kalid, T. E., Prado, T. d. A., et al.: An acoustic lens for displacement-free sectorial inspection of pipes with ultrasonic phased arrays, *NDT & E International*, 103459 (2025).
4. Xue, Y., Croxford, A. J., Wilcox, P. D.: Improvement in repeatability of ultrasonic array imaging in materials with high structural noise, *NDT & E International* **133**, 102732 (2023).
5. Barashok, K., Choi, Y., Choi, Y., et al.: Machine learning techniques in ultrasonics-based defect detection and material characterization: a comprehensive review. *Advances in Mechanical Engineering* (Sage Publications Inc.) (6) (2025).
6. Liu, J., Xie, G., Wang, J., et al.: Deep industrial image anomaly detection: a survey, *Machine Intelligence Research* **21**(1), 104–135 (2024).
7. Uhlig, S., Alkhasli, I., Schubert, F., et al.: A review of synthetic and augmented training data for machine learning in ultrasonic non-destructive evaluation, *Ultrasonics* **134**, 107041 (2023).
8. Zhao, Y., Nasrullah, Z., Li, Z.: Pyod: a python toolbox for scalable outlier detection, [Journal of Machine Learning Research](#) **20**(96), 1–7 (2019).
9. Pérez, E., Ardiç, C. E., Çakıroğlu, O., et al.: Integrating ai in nde: techniques, trends, and further directions, *NDT & E International* **156**, 103442 (2025).