A MACHINE LEARNING APPROACH TO NON-DESTRUCTIVE ULTRASONIC TESTING OF INFRASTRUCTURE BOLTS

Abdul Azziz Bin Abd Talib, Lim Chun Yee and Liew Chin Kian

Singapore Institute of Technology, Singapore

ABSTRACT

Anchor bolts are critical connectors that maintain structural alignment and load-bearing capacity in civil infrastructures. Defects in these bolts can compromise safety and operational reliability, yet traditional visual inspections are often insufficient for detecting internal damage. This study presents a feature-based machine learning framework integrated with ultrasonic testing to enhance defect detection in anchor bolts. A fabricated bolt inspection system with a 10 MHz ultrasonic transducer was used to acquire 477 ultrasonic signals from pristine, straight thinning, and tapered thinning bolts, with additional validation signals collected from field-installed bolts. Three dimensionless features were engineered to capture signal clarity and defect-related scattering. Six machine learning classifiers were evaluated using stratified cross-validation, with Gradient Boosting achieving the highest accuracy of 93%, outperforming other classifiers, including ensemble methods. The model demonstrated strong robustness in distinguishing between non-defective (Green), monitoring required (Yellow), and defective (Red) bolts even under practical variability. Field deployment in train tunnels further validated the model's reliability with no false Red classifications. The results confirm the viability of integrating ultrasonic testing with machine learning for automated anchor bolt inspection, enabling accurate, data-driven infrastructure maintenance and predictive safety strategies.

KEYWORDS

Anchor bolts, Ultrasonic testing, Machine learning, Structural health monitoring, Non-destructive testing, Infrastructure safety

1. INTRODUCTION

Anchor bolts are essential load-bearing connectors that maintain structural alignment and integrity across a wide range of civil infrastructures, including bridges, tunnels, and rail systems. Failure of these critical components can lead to catastrophic structural collapse, service disruptions, and significant risks to public safety. Consequently, regular and reliable inspection of anchor bolts is a vital component of infrastructure maintenance and safety protocols. Traditional visual inspection methods have long been the conventional approach for assessing anchor bolts embedded within civil structures. However, these methods are fundamentally limited in their ability to detect internal defects such as sub-surface corrosion or material thinning, as they depend heavily on the expertise and subjective judgment of individual inspectors. This reliance introduces inconsistencies in assessment outcomes and raises the risk of undetected defects, which may ultimately result in hazardous failures and costly emergency repairs.

To overcome these limitations, recent developments in non-destructive testing (NDT) have focused on the integration of ultrasonic testing with machine learning. Ultrasonic testing works by transmitting high-frequency waves into materials and interpreting the reflected signals to

identify internal anomalies such as cracks, voids, or corrosion. Machine learning algorithms can analyse these signals more efficiently and consistently compared to manual approaches by learning from patterns in collected data, which reduces human error and enhances defect detection accuracy. By learning from collected datasets, these models can detect subtle anomalies and classify bolt conditions with high precision. This integration enhances reliability, reduces human error, and enables real-time decision-making during inspections.

In a comprehensive review of defect detection for civil infrastructures, Zhang et al. [2] highlighted the limitations of traditional NDT and showcased how Artificial Intelligence (AI) can automate inspection processes, improve detection reliability, and facilitate predictive maintenance. Similarly, Abdullah et al. [3] proposed an advanced signal processing approach for structural health monitoring of bridges, demonstrating successful extraction of damage-sensitive features. For bolted joints, Soleimanpour et al.[4] utilized contact acoustic nonlinearity, a specialized ultrasonic technique, to identify imperfections. Guided ultrasonic waves were transmitted through the bolts and nonlinear acoustic responses were observed at damaged interfaces, revealing high sensitivity in detecting and localizing loosened joints. Meanwhile, Smagulova et al. [5] employed the ultrasonic pulse-echo method to collect signal responses from adhesive joints. By extracting key ultrasonic features from these signals and applying machine learning classifiers, detection and measurement of defect depth in multilayered structures were obtained with high accuracy. In railway infrastructure, Meixedo et al. [6] developed a wavelet-based clustering approach that detects minor stiffness reductions while proving the effectiveness of unsupervised learning contexts. These studies consistently support the integration of machine learning and signal processing as a robust framework for identifying early-stage damage identification in various materials.

Despite these advancements, the application of machine learning enhanced ultrasonic testing for embedded anchor bolts remains underexplored. Avci et al. [7] reviewed vibration-based damage detection techniques that are advancing towards more data-driven machine learning solutions. Daghigh et al. [8] emphasized the importance of interpretable and physics-informed machine learning approaches in defect analysis across engineering materials and composites. These principles are increasingly valuable for modelling ultrasonic responses in structural bolts. Additionally, Al Lahham et al. [9] provided a comprehensive perspective on online condition monitoring in wind turbines, illustrating the importance of continuous diagnostics using sensor data and advanced signal processing, which aligns closely with goals in infrastructure health monitoring.

Building on these insights, this study makes three primary contributions: (1) it introduces a novel set of physically interpretable, dimensionless ultrasonic features (Pre-SNR, Post-SNR, and Echo Decay Ratio) specifically engineered for anchor bolt defect classification; (2) it presents a systematic benchmarking of six supervised classifiers, identifying Gradient Boosting as the optimal model for this application; and (3) it provides end-to-end validation spanning laboratory-controlled fabricated defects through to live field deployment in an operational train tunnel, a combination not previously demonstrated for embedded anchor bolts. Unlike prior works that apply raw signal processing or evaluate single-model pipelines, this framework directly addresses the gap in automated, multi-class defect severity classification for safety-critical embedded bolts. The framework is validated not only on laboratory-fabricated defects but also on field-acquired signals from operational train tunnel infrastructure, demonstrating its practical applicability for real-world deployment.

2. METHODOLOGY

2.1. Experimental Setup

Ultrasonic testing has been widely adopted for evaluating internal flaws in embedded or inaccessible structural components due to its sensitivity and non-invasiveness [4]. In this study, a fabricated bolt inspection system equipped with a 10MHz ultrasonic transducer was deployed. Signal acquisition was carried out using the transducer with bolt end contact surfaces which were cleaned to improve coupling and ultrasonic testing pulse echo signal quality [5].

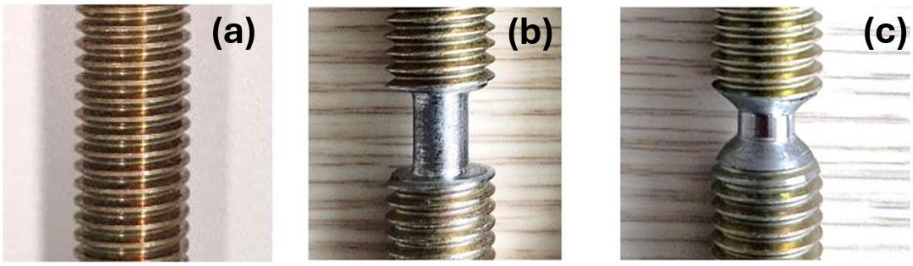

Figure 1: Fabricated bolt samples used: (a) Pristine Bolt, (b) Straight Thinning Bolt, and (c) Tapered Thinning Bolt

140mm length M12 galvanised steel anchor bolts threaded end-to-end were considered. The dataset included ultrasonic measurements from three bolt conditions:

- **Pristine Bolt**: An undamaged bolt with uniform diameter (Figure 1a)
- **Straight Thinning Bolt**: A bolt with a uniformly reduced section, simulating mid-span material loss (Figure 1b)
- **Tapered Thinning Bolt**: A bolt with a gradually thinned section, simulating progressive wear (Figure 1c)

Thinning defects were fabricated in a range of lengths (3mm, 6mm, and 9mm)and depths (1mm, 2mm, and 3mm) to simulate various bolt degradations with diverse signal responses for aiding in machine learning model training. The thinning defects were fabricated at a bolt location interfacing with a steel plate secured to concrete, which is a region prone to damage due to lateral vibration loads (Figure 2).

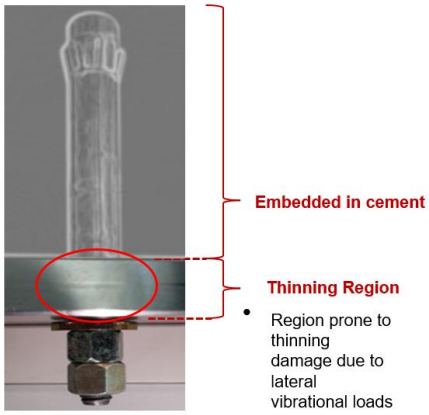

Figure 2: Anchor bolt embedded in concrete showing the thinning-prone region due to lateral vibrational loads

A total of 477 ultrasonic signals were collected from these bolts, comprising 165 Green, 132 Yellow, and 180 Red samples. Each signal was labelled based on defect severity according to measured thinning depth: Green (0–1mm, non-defective), Yellow (1–2.5mm, monitoring required), and Red (>2.5mm, defective). The class distribution (Green: 34.6%, Yellow: 27.7%, Red: 37.7%) exhibits a moderate imbalance, with the Yellow class being the smallest. To mitigate potential bias from this imbalance, stratified sampling was maintained throughout the 80/20 train-test split and all cross-validation folds, ensuring proportional class representation at each stage. In addition, weighted evaluation metrics (precision and recall) were computed to give appropriate weight to each class during performance assessment. No synthetic oversampling (e.g., SMOTE) was applied, as the natural class proportions reflect realistic field inspection scenarios and the dataset size was deemed sufficient for stable model training across all three classes. Ultrasonic signals were also collected from similar anchor bolts in good condition installed in overhead structures across different field locations, which were then used to validate the developed models in detecting Green bolts in real operating environments.

2.2. Feature Extraction

The data acquired from the bolts were in the form of raw radio frequency A-scan signals. These A-scans illustrate the propagation of ultrasonic waves through the bolt material. When ultrasonic waves encounter a discontinuity or change in material property such as a defect, part of the wave is reflected as an echo pulse. These reflected echoes were analysed in terms of their amplitude and time-of-flight to infer internal structural conditions. A sample of the ultrasonic A-scan signal from a thinned bolt is shown in Figure 3.

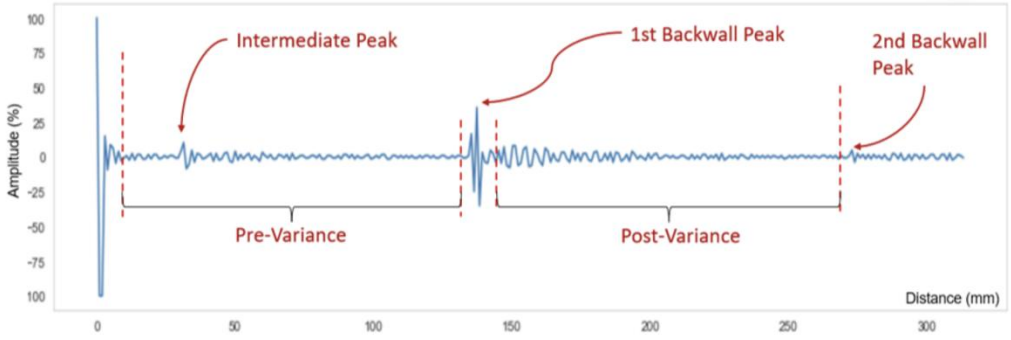

Figure 3: Characteristics of an ultrasonic A-scan signal acquired from a thinned bolt.

When collecting ultrasonic signals, the recorded amplitude could vary due to several factors such as slight differences in transducer contact pressure, surface coupling conditions, and cleanliness of the probe–bolt interface, even when the same gain setting was used. Such variations meant that identical bolts in the same condition might produce different signal amplitudes. To eliminate this inconsistency, normalization was performed to adjust all signal amplitudes to a comparable scale, allowing the focus to remain on defect-related signal features instead of acquisition artefacts.

Following normalization, an analytical process was performed to create dimensionless features that were scale-invariant and robust for training the machine learning models. The rationale for deriving dimensionless features was to ensure that differences in measurement units and scale did not affect how the model learned to distinguish between defective and non-defective bolts. Additionally, non-dimensionalization helps reduce the influence of measurement noise by focusing on relative differences within the signal shape and pattern rather than on raw values, making the features more robust and reliable for classification.

Three key dimensionless features were engineered to characterize the signal morphology based on the A-scan shown in Figure 3. Unlike raw amplitude values, these features are scale-invariant and robust against acquisition noise.

$$Feature\ 1 = \frac{1st\ Backwall\ Peak}{\sqrt{Pre - Variance}}$$

Feature 1 (Pre-SNR): Defined as the ratio of the first backwall echo amplitude to the square root of the signal variance in the region preceding the echo. This reflects the signal-to-noise ratio, quantifying how prominent the first echo is above the noise floor.

$$Feature\ 2 = \frac{1st\ Backwall\ Peak}{\sqrt{Post - Variance}}$$

Feature 2 (Post-SNR): Defined as the ratio of the first backwall echo amplitude to the square root of signal variance in the post-echo region, offering an alternate view of signal clarity.

$$Feature\ 3 = \frac{2nd\ Backwall\ Peak}{1st\ Backwall\ Peak}$$

Feature 3 (Echo Decay Ratio): Defined as the ratio of the second backwall peak to the first, indicating the rate of amplitude decay through the bolt.

Features 1 and 2 are inspired by the concept of Signal-to-Noise Ratio (SNR), which is commonly used to describe how clearly a signal stands out from background noise. In ultrasonic testing, the "signal" refers to the back wall echo amplitude, while the "noise" refers to the variability in the waveform in regions where no clear echoes are present.

The square root of the signal variance is used as the denominator based on three key reasons. Firstly, variance serves as a measure of the signal's fluctuations around its mean, effectively capturing the strength of background noise in regions without prominent echoes. Secondly, taking the square root of this variance ensures that the resulting value has the same unit as the signal amplitude, which allows for the formation of a dimensionless and interpretable ratio. Thirdly, a higher ratio indicates that the echo stands out more clearly above the noise floor, which is essential for accurate and reliable defect detection in ultrasonic testing.

Feature 1 (Pre-SNR) uses the variance before the echo, capturing noise in the region before the first back wall reflection, while Feature 2 (Post-SNR) uses the variance after the echo to evaluate signal clarity under different noise environments (e.g. reflections and scatterings occurring later in the signal).For Feature 3 (Echo Decay Ratio), the rationale was that as ultrasonic waves travelled through a material, defects such as thinning or cracks caused increased scattering and energy loss, leading to a faster decay in echo amplitudes. By taking the ratio of the second back wall peak amplitude to the first, the signal decay rate could be measured through the bolt. A lower decay ratio often indicates greater internal scattering, which could signal defect presence or severity. These features, when combined, were selected to highlight signal-to-noise characteristics and echo decay behaviour, which were strongly linked to internal bolt conditions. Together, they provided the model with robust, comparable indicators for defect classification.

To validate the discriminative power of these features prior to model training, a one-way ANOVA was conducted across the three defect classes (Green, Yellow, Red) for each feature (Table 1). All three features yielded statistically significant between-class differences: Feature 1

(Pre-SNR) achieved $F = 57.94$, $p = 3.09 \times 10^{-23}$; Feature 2 (Post-SNR) achieved $F = 4.12$, $p = 0.017$; and Feature 3 (Echo Decay Ratio) achieved the strongest separation with $F = 161.69$, $p = 2.91 \times 10^{-54}$. Notably, while all three features are statistically significant, Feature 2 shows a weaker effect compared to Features 1 and 3, which is consistent with the greater overlap observed for Post-SNR values in the 3D feature space. Pairwise Cohen's d effect sizes (Table 2) further confirm that the largest separations occur between the Green and Red classes for Feature 1 ($d = 0.97$) and Feature 3 ($d = 1.89$), both classified as large effects, while Green vs. Yellow separations are smaller ($d = 0.07$ and $d = 0.15$ respectively), consistent with the cluster overlaps observed in Figure 4. Feature 2 shows small to negligible effect sizes across all pairs, suggesting it contributes supplementary rather than primary discriminative information. All three features were retained as they each capture a distinct physical characteristic of the ultrasonic signal and together improve the overall separability of the classifier.

To complement the effect size analysis, pairwise point-biserial correlation coefficients, $|r|$ were calculated to assess the discriminative linearity of each feature. Feature 3 demonstrated a large correlation ($|r| = 0.69$) for the Green vs Red pair, indicating a strong, predictable linear trend as bolts transition from healthy to severely defective states. Similarly, Feature 1showed a large correlation ($|r| = 0.53$) for the Yellow vs Red distinction, confirming its role in identifying critical thinning. However, for the more challenging Green vs Yellow transition, Feature 2 achieved the highest correlation ($|r| = 0.19$), significantly outperforming Feature 1 ($|r| = 0.04$) and Feature 3 ($|r| = 0.07$) in this specific domain. This suggests that while Feature 2 is less effective at identifying severe defects, it provides critical, unique information for resolving the boundary between healthy and early-stage thinning. These results highlight that individual features possess varying localized strengths, reinforcing the need for an ensemble-based learner like Gradient Boosting to adaptively integrate these disparate signals into a unified, high-accuracy classification framework.

Table 1: One-way ANOVA results for each engineered feature across three bolt defect classes

Feature	***F*-statistics**	***p*-value**	**Sig.**	**Mean ± SD (Green)**	**Mean ± SD (Yellow)**	**Mean ± SD (Red)**
Feature 1(Pre-SNR)	57.94	3.09×10^{-23}	Yes	241.52 ± 189.27	230.03 ± 112.54	92.75 ± 103.26
Feature 2 (Post-SNR)	4.12	0.017	Yes	29.59 ± 15.48	37.19 ± 23.62	31.50 ± 28.37
Feature 3 (Echo Decay Ratio)	161.69	2.91×10^{-54}	Yes	0.41 ± 0.15	0.386 ± 0.15	0.18 ± 0.07

Table 2: Pairwise Cohen's d effect sizes and Point-biserial correlation coefficients between features

Feature	**Class Pair**	**Cohen's *d***	**$\|d\|$ Magnitude**	**Point-biserial $\|r\|$**	**$\|r\|$ Magnitude**
Feature 1 (Pre-SNR)	Green vs Yellow	0.07	Negligible	0.04	Negligible
Feature 1 (Pre-SNR)	Green vs Red	0.97	Large	0.44	Medium
Feature 1 (Pre-SNR)	Yellow vs Red	1.27	Large	0.53	Large
Feature 2 (Post-SNR)	Green vs Yellow	−0.38	Small	0.19	Small
Feature 2 (Post-SNR)	Green vs Red	−0.08	Negligible	0.04	Negligible

Feature	Class Pair	Cohen's *d*	|*d*| Magnitude	Point-biserial |*r*|	|*r*| Magnitude
Feature 2 (Post-SNR)	Yellow vs Red	0.22	Small	0.11	Small
Feature 3 (Echo Decay Ratio)	Green vs Yellow	0.15	Negligible	0.07	Negligible
Feature 3 (Echo Decay Ratio)	Green vs Red	1.89	Large	0.69	Large
Feature 3 (Echo Decay Ratio)	Yellow vs Red	1.74	Large	0.67	Large

A 3D feature space diagram (Figure 4) was constructed using the engineered dimensionless features to visualize the distribution and separability of ultrasonic signals across different bolt conditions like Green, Yellow, and Red colour codes presented previously. These categories align with practical infrastructure inspection standards, where early-stage thinning (Green) indicates that the bolts are still in good condition, moderate thinning (Yellow) suggests that closer monitoring of the bolt is required, and severe thinning (Red) indicates a defective bolt that requires prompt remedial action.

To generate the feature space, standardization (i.e. *Z*-score normalization) was applied to the extracted features by subtracting the mean and dividing by the standard deviation for each feature across the dataset. This process ensures that all features are on the same scale, allowing the machine learning models to interpret and compare features accurately without bias toward those with larger numerical ranges. The Z-score normalization formula is expressed as:

$$z = \frac{(\mathrm{X} - \mu)}{\sigma}$$

Where X is the original feature value, μ is the mean, and σ is the standard deviation of this feature.

The 3D feature space visualization revealed distinct clustering among the Green, Yellow, and Red categories, demonstrating the effectiveness of the selected features in distinguishing defect severities. However, overlaps between clusters, particularly between Green and Yellow, were observed. These overlaps are expected in practical ultrasonic testing due to slight measurement inconsistencies, material variability, and the gradual nature of thinning transitions, which can cause signals from bolts with near-boundary thinning depths to exhibit similar feature patterns.

As a result, it is not feasible to achieve 100% classification accuracy across all classes, particularly in borderline cases. The practical goal is to minimize critical misclassifications, specifically cases in which a Red (defective) bolt is incorrectly classified as Green (non-defective), while maintaining high overall accuracy and precision. This reflects a realistic understanding of ultrasonic testing limitations and aligns with the use of machine learning in defect detection, where the focus is on robust, consistent detection of defect presence and severity despite inherent uncertainties. In particular, the Red category (defective bolts) consistently exhibited lower Feature 1 and 2 values within this 3D space, reflecting reduced echo strength and higher signal variance due to increased scattering from internal defects (see Figure 4).This observation supports the physical interpretation of the features and reinforces their relevance for detecting severe bolt degradation.

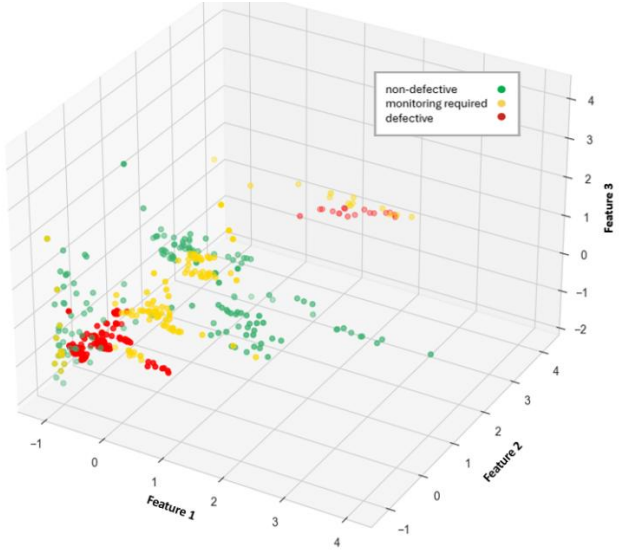

Figure 4: 3D feature space diagram for ultrasound signals collected from 3 classes of bolt conditions (Green, Yellow, and Red)

2.3. Model Development

To evaluate the classification performance, six machine learning models were selected, following recent reviews highlighting their use in structural health monitoring and NDT [8]. The machine learning models developed and trained using field and lab-acquired ultrasonic signals included support vector classifier (SVC), logistic regression, k-nearest neighbour, random forest, bagged decision trees, and gradient boosting, all implemented using the Scikit-learn Python library, which provides a consistent and reliable framework for model development, evaluation, and hyperparameter tuning. To ensure reproducibility, a fixed random seed (random_state = 223) was set for all stochastic components, including the train-test split, cross-validation fold assignment, and ensemble model initialization. Preprocessing consisted of amplitude normalization followed by Z-score standardization, applied independently to the training set with the derived parameters (mean and standard deviation) subsequently applied to the test set to prevent data leakage. The dataset was stratified prior to all splitting and folding operations to preserve class proportions. A stratified 80/20 training-test split was first used to partition the dataset. Subsequently, 5-fold stratified cross-validation was applied exclusively on the training set for hyperparameter tuning via grid search, ensuring no test-set information was used during model selection. To provide statistical confidence in the reported accuracy, a 95% confidence interval was computed for the Gradient Boosting test accuracy using the Wilson score method: 92.71% (95% CI: 85.7%–96.4%),confirming the reliability of the result given the test set size of 96 samples.

Hyperparameters were optimized via grid search, and the final selected values are summarized in Table 3. Detailed descriptions of each hyperparameter and the corresponding search ranges are provided in the Supplementary Material. This aligns with best practices for data-driven diagnostics in structural health monitoring and fault classification [7][9].

Table 3: Best Hyperparameters from Grid Search

Model	Hyperparameters Identified from Grid Search
Support Vector Classifier	{ ‘C’:100, ‘gamma’:’auto’, ‘kernel’:’rbf’ }
Logistic Regression	{ ‘C’:0.1, ‘penalty’:’l1’, ‘solver’:’saga’ }

K-Nearest Neighbour	{ 'metric':'euclidean', 'n_neighbours':3, 'weights':'uniform' }
Random Forest	{ 'max_features':'sqrt', 'n_estimators':1000 }
Bagged Decision Tree	{ 'n_estimators':1000 }
Gradient Boosting	{ 'learning_rate':0.1, 'max_depth':9, 'n_estimators':100, 'subsample':0.7 }

3. RESULTS AND DISCUSSION

3.1. Model Performance

Performance was evaluated using **accuracy**, **precision**, and **recall** giving a balanced view of each classifier's capabilities. Accuracy is the proportion of correctly classified instances; precision is the ratio of true positives to all predicted positives; and recall is the ratio of true positives to all actual positives. Weighted averages were used to account for class imbalance. The use of these metrics follows standard recommendations for ML in NDT, particularly when class imbalance is present [8].

Table 4: Performance metrics for each machine learning model (test set).

Model	Accuracy	Precision (Weighted Average)	Recall (Weighted Average)
Support Vector Classifier	89%	89%	89%
K-Nearest Neighbour	90%	90%	90%
Logistic Regression	67%	68%	67%
Random Forest	92%	92%	92%
Bagged Decision Tree	91%	91%	91%
Gradient Boosting	93%	93%	93%

The performance of each classifier on the **test set** is summarized in Table 4, including accuracy, precision, and recall metrics. These results provide a comprehensive evaluation of each model's ability to classify bolt conditions based on ultrasonic signal features.

Among the six classifiers evaluated, **Logistic Regression** achieved the lowest accuracy at 67%, which reflects the limitations of a linear decision boundary in separating the overlapping Green and Yellow clusters present in the feature space. The **Support Vector Classifier** (rbf kernel, C=100) performed slightly below the ensemble methods at 89%, likely due to the kernel's sensitivity to the noisy separability between adjacent severity classes. **K-Nearest Neighbours** (n_neighbors=3, euclidean metric) achieved 90% accuracy, performing reasonably well given its instance-based approach, though its sensitivity to local noise makes it less robust for borderline cases. **Bagged Decision Trees** and **Random Forest** both outperformed these models at 91% and 92% respectively, leveraging ensemble learning to reduce variance and to better handle the structured ultrasonic data. However, neither method adaptively emphasizes hard-to-classify samples during training, which limits their refinement of decision boundaries in challenging regions of the feature space.

Among all the classifiers considered, **Gradient Boosting** achieved the highest accuracy of 93%, with strong precision and recall across all classes. Its superior performance stems from its sequential boosting strategy, where each weak learner corrects the errors of the previous one, allowing the model to focus on hard-to-classify borderline samples and adapt to the overlapping Green-Yellow region more effectively than other methods. These advantages align well with the safety-critical demands of automated ultrasonic bolt inspection, reinforcing Gradient Boosting as the most suitable model for deployment in this study.

These results align with the goals of achieving reliable, accurate, and data-driven defect detection in infrastructure maintenance, positioning Gradient Boosting as the preferred choice for integration into the automated inspection process. This is due to its ensemble approach which has been proven effective in handling structured ultrasonic data [5][8].

To contextualise these results within the broader literature, Table 5 presents a qualitative comparison with selected related works in NDT-based defect classification. Smagulova et al. [5] applied ultrasonic pulse-echo features with an SVM classifier to detect and classify defect depth in adhesive joints, achieving over 90% accuracy on train/test data and 83% on unseen data for binary defect classification. However, their work targeted planar laminated structures rather than embedded cylindrical bolts, used binary rather than multi-class severity grading, and did not include field validation. Meixedo et al. [6] demonstrated an unsupervised wavelet and clustering approach for early damage detection in railway bridges using traffic-induced responses, with high sensitivity to small stiffness reductions; however, no classification accuracy metric was reported and the method does not support multi-class defect severity grading. Daghigh et al. [8] provide a comprehensive review of ML applications for defect detection in composite materials, discussing a range of supervised and deep learning approaches; as a review, it does not report a single experimental accuracy figure but highlights that deep learning models generally require large labelled datasets and offer limited interpretability. The proposed framework is distinctive in combining interpretable, physics-derived features with a compact dataset (477 samples), three-class defect severity classification, and direct field deployment in an operational train tunnel, a combination not demonstrated by any of the above works.

Table 5: Qualitative comparison of the proposed framework with selected related works in NDT-based defect classification

Study	**NDT Method**	**ML Approach**	**Target Structure**	**Reported Performance**	**Multi-class Severity**	**Field Validation**
Smagulova et al. [5]	Ultrasonic pulse-echo	SVM (feature-based)	Adhesive joints (planar)	>90% (train/test); 83% (unseen)	No (binary only)	No
Meixedo et al. [6]	Vibration / traffic response	Unsupervised wavelet + clustering	Railway bridge	High sensitivity to 5% stiffness reduction; no accuracy % reported	No (anomaly detection only)	Simulated (digital twin)
Daghigh et al. [8]	Various (IR, ultrasonic, etc.)	Review (SVM, CNN, DL)	Composite materials	N/A (review paper)	Varies by study reviewed	N/A
Present Study	**Ultrasonic A-scan (pulse-echo)**	**Gradient Boosting (feature-based)**	**Anchor bolts (embedded)**	**92.71% (test set); 95% CI: 85.7%–96.4%**	**Yes (3 classes: Green / Yellow / Red)**	**Yes (operational train tunnel)**

To further evaluate the best-performing model, the confusion matrix for the Gradient Boosting classifier is presented in Table 6. The matrix reveals that all 36 Red (defective) bolts in the test set were correctly classified, yielding a perfect recall of 1.00 for the most safety-critical class. Of the 33 Green samples, 32 were correctly identified, with only one misclassified as Yellow. Six of the 27 Yellow samples were misclassified as Green. Notably, there were zero misclassifications

involving the Red category, meaning no defective bolt was missed and no non-defective bolt was falsely flagged as defective. All classification errors occurred exclusively between the Green and Yellow categories, which represent adjacent severity levels with inherently similar signal characteristics. This error pattern is acceptable from a safety perspective, as it does not compromise the detection of critical defects.

Table 6: Confusion matrix for Gradient Boosting classifier on the test set (80/20 stratified split)

Actual \ Predicted	Green	Yellow	Red
Green (33)	32	1	0
Yellow (27)	6	21	0
Red (36)	0	0	36

3.2. Case Study

The performance of these models was further validated with on-site signals collected from 10 similar overhead structure anchor bolts of good condition in a train tunnel, which were not part of the training or test dataset. The results are summarized in Table 7, highlighting their false positive rate (e.g. the chance of classifying Green or Yellow bolt as Red) when deployed on real on-site anchor bolts. The anchor bolts were all undamaged, serving as the baseline for evaluating the precision of each classifier's predictions for the Green category.

In Table 7, the Gradient Boosting Classifier again emerged as the most reliable model for field deployment. Of the 10 on-site bolts, 8 were classified as Green and 2 as Yellow, with no false Red classifications observed across any of the tested locations. This is particularly significant given that all bolts were known to be in good condition, as it confirms the model's ability to avoid critical false alarms in a real operating environment. In contrast, the Support Vector Classifier produced one false Red classification (Bolt 5) and two Yellow classifications, while Logistic Regression also flagged one bolt as Red (Bolt 3). These false Red results are undesirable in practice as they would trigger unnecessary remedial action. K-Nearest Neighbours, Random Forest, and Bagged Decision Trees each produced two Yellow classifications with no false Red results, performing comparably to Gradient Boosting in avoiding critical misclassifications in the field. However, when considered alongside the laboratory test set results in Table 4, Gradient Boosting remains the preferred model as it achieved the highest overall accuracy of 93%, outperforming all other classifiers across both controlled and real-world conditions. This consistency across laboratory and field evaluations reinforces Gradient Boosting as the most suitable model for integration into an automated anchor bolt inspection system.

Table 7: Predicted class for 10 overhead structure anchor bolts at on-site train tunnel locations.

Bolt Number	Support Vector Classifier	Logistic Regression	K-Nearest Neighbour	Random Forest	Bagged Decision Tree	Gradient Boosting Classifier
1	Green	Green	Green	Green	Green	Green
2	Green	Green	Green	Green	Green	Green
3	Green	Red	Green	Green	Green	Green
4	Yellow	Green	Green	Green	Green	Green
5	Red	Red	Green	Green	Green	Green
6	Green	Green	Green	Green	Green	Green
7	Yellow	Green	Yellow	Yellow	Yellow	Yellow

8	Green	Green	Green	Green	Green	Green
9	Green	Green	Green	Green	Green	Green
10	Yellow	Green	Yellow	Yellow	Yellow	Yellow

4. CONCLUSION

This study presented and validated a feature-based machine learning framework for automated, multi-class defect severity classification in infrastructure anchor bolts using ultrasonic A-scan signals. The work addresses a recognised gap in the NDT literature, where machine learning approaches for embedded bolts with real field validation remain scarce. By combining physically interpretable dimensionless features with a systematic classifier comparison and end-to-end validation spanning controlled laboratory conditions to an operational train tunnel, the study advances the state-of-the-art in automated structural bolt inspection.

A feature extraction strategy utilizing three dimensionless features (Pre-SNR, Post-SNR, and Echo Decay Ratio) provided compact, physics-grounded signal descriptors. Statistical validation confirmed that all three features exhibited significant between-class discrimination ($p < 0.001$), and their combination produced clear 3D feature space clustering, with predictable Green-Yellow boundary overlap reflecting the gradual nature of thinning defects.

Among the six machine learning classifiers evaluated, Gradient Boosting emerged as the best-performing model, achieving the highest test accuracy of 93% and demonstrating superior precision and recall across all defect classes. Critically, the model achieved perfect classification of Red (defective) bolts with zero false negatives, meaning no defective bolt was missed during evaluation. All misclassifications occurred exclusively between the Green and Yellow categories, which represent adjacent severity levels with inherently overlapping signal characteristics. This error pattern is acceptable from a safety standpoint, as it does not compromise the detection of structurally compromised bolts. The model further demonstrated strong generalization to unseen field data, maintaining zero false Red classifications during on-site deployment, which is essential for safety-critical inspection applications.

The results validate the feasibility and effectiveness of machine learning-enhanced ultrasonic testing as a viable alternative to manual inspection methods for automated bolt defect detection. The framework enables reliable, data-driven infrastructure monitoring with the potential for integration into routine maintenance workflows. Nevertheless, several limitations of the present study should be acknowledged. First, the training dataset was derived from a single bolt geometry (140 mm M12 galvanised steel) with two specific defect morphologies (straight and tapered thinning). The generalisability of the trained model to other bolt sizes, materials, or defect types (e.g. corrosion pitting, thread damage) has not yet been established. Second, the field validation set comprised only 10 bolts from a single infrastructure site, which, while encouraging, is insufficient to draw statistically robust conclusions about real-world performance across diverse deployment environments. Third, the current feature engineering approach relies on manually identified signal windows and peak amplitudes; any systematic variation in waveform morphology caused by extreme temperature, coupling pressure, or bolt geometry could shift the computed feature values and potentially degrade classification performance. Fourth, while the selected features carry physical meaning, the internal decision logic of the Gradient Boosting ensemble is not directly transparent. This is a secondary concern given that the framework is intended to supplement rather than replace inspector judgement, providing an objective, data-driven confidence indicator alongside conventional visual inspection rather than acting as a standalone decision system. Future work may nonetheless explore explainability tools such as SHAP values to further strengthen inspector trust in the model outputs. Addressing these

limitations through expanded field studies and physics-informed modelling represents an important direction for future research. Future work will focus on expanding the dataset with additional field samples from diverse infrastructure environments, investigating the use of physics-informed machine learning approaches for improved model interpretability, and developing a real-time inspection system that combines the trained classifiers with portable ultrasonic hardware for on-site predictive maintenance of critical infrastructure systems.

REFERENCES

[1] C. Lu and Y. Sonoda, "An analytical study on the pull-out strength of anchor bolts embedded in concrete members by sph method," Applied Sciences (Switzerland), vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188526.

[2] Y. Zhang, C. L. Chow, and D. Lau, "Artificial intelligence-enhanced non-destructive defect detection for civil infrastructure," Mar. 01, 2025, Elsevier B.V.doi: 10.1016/j.autcon.2025.105996.

[3] H. A. Abdullah, M. U. Hanif, M. U. Hassan, J. M. Shahid, S. A. Khan, and A. Ali, "Improved damage assessment of bridges using advanced signal processing techniques of CEEMDAN-EWT and Kernal PCA," Eng Struct, vol. 329, Apr. 2025, doi: 10.1016/j.engstruct.2025.119774.

[4] R. Soleimanpour, S. M. Soleimani, and N. K. Mohammad, "Damage detection and localization in loose bolted joints," in Procedia Structural Integrity, Elsevier B.V., 2021, pp. 956–963. doi: 10.1016/j.prostr.2022.02.031.

[5] D. Smagulova, V. Samaitis, and E. Jasiuniene, "Machine learning based approach for automatic defect detection and classification in adhesive joints," NDT and E International, vol. 148, Dec. 2024, doi: 10.1016/j.ndteint.2024.103221.

[6] A. Meixedo, D. Ribeiro, J. Santos, R. Calçada, and M. Todd, "Automatic wavelet-based clustering approach for damage detection on railway bridges," in Transportation Research Procedia, Elsevier B.V., 2023, pp. 4287–4294. doi: 10.1016/j.trpro.2023.11.629.

[7] O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, and D. J. Inman, "A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications," Jan. 15, 2021, Academic Press. doi: 10.1016/j.ymssp.2020.107077.

[8] V. Daghigh, H. Daghigh, T. E. Lacy, and M. Naraghi, "Review of machine learning applications for defect detection in composite materials," Machine Learning with Applications, vol. 18, p. 100600, Dec. 2024, doi: 10.1016/j.mlwa.2024.100600.

[9] E. Al Lahham, L. Kanaan, Z. Murad, H. M. Khalid, G. A. Hussain, and S. M. Muyeen, "Online condition monitoring and fault diagnosis in wind turbines: A comprehensive review on structure, failures, health monitoring techniques, and signal processing methods," Apr. 01, 2025, KeAi Communications Co.doi: 10.1016/j.grets.2024.100153.