

INVATOR

SAMHÄLLE. TEKNIK. INNOVATION.



Report

Machine learning-based decision support tool for damage assessment in concrete structures

Swedish

Beslutsstödsverktyg baserat på maskininlärning för skadekontroll i betongkonstruktioner

Coordinator:

Björn Täljsten

Compiled by:

Ali Mirzazade, Cosmin Popescu & Björn Täljsten

Client:

Smart Built Environment – Innovationsidén 7

INVATOR Sverige AB

Postadress:

Box 1093

181 22 Lidingö

Tfn: 070-537 43 70

info@invator.se

www.invator.se

Org.nummer:

556680-3200

Report

Machine learning-based decision support tool for damage assessment in concrete structures

KEYWORDS

NDT, Machine Learning,
Concrete bridges

VERSION

1

DATE

2026-02-06

CLIENT

Smart Built Environment – Innovationsidén 7

OUR REFERENCE

PROJECT NO.

2436

NUMBER OF PAGES

128

SUMMARY

Click or tap here to enter text.

COMPILED BY

Ali Mirzazade

SIGNATUR

APPROVED BY

Björn Täljsten

SIGNATUR

History

VERSION	DATE	Version description
1	2024-11-30	Click or tap here to enter text.
C:\SynologyDrive\Projects 2024\2436_FORMAS\Report\Rapport_2025.01.19_MH.docx		
Click or tap here to enter text.	Click or tap to enter a date.	Click or tap here to enter text.

Table of content

Summary	7
Sammanfattning	8
1 Introduction	9
1.1 Background	9
1.2 Objectives	10
2 Literature review and existing practices	11
2.1 Importance of grouting and risks of voids in PT systems	11
2.1.1 Segregated or defective grout	12
2.1.2 Voids due to grouting errors	13
2.2 Detection of voids in tendon ducts: Challenges of detecting internal defects	14
2.2.1 Duct material and geometry	15
2.2.2 Access limitations	16
2.2.3 Signal interference	17
2.2.4 Resolution and sensitivity	17
2.2.5 Human factors and operator dependence	17
2.3 Basis for current study: findings	18
3 NDT methods: Physical principles and data structure	20
3.1 Ground-penetration radar (GPR)	20
3.1.1 Physical Principle	20
3.1.2 Data type and structure – SEG-Y format	21
3.1.3 Essential signal processing steps	21
3.2 Ultrasonic Pulse Echo (UPE)	23
3.2.1 Physical Principle	23
3.2.2 RAW data structure	24
3.2.3 Mathematical basis of UPE and SAFT processing	24
3.3 Impact-Echo (IE)	25
3.3.1 Physical Principle	25
3.3.2 RAW data structure	26
3.3.3 Mathematical basis of Impact–Echo processing	26
4 Mock-Up specimens in NDT verification: Insights from past studies and basis for the present work	28
4.1 Test 1: (Wiggenhauser et al., 2015)	28
4.2 Test 2: (Giannini et al., 2020)	29
4.3 Test 3: (Dethof & Keßler, 2024)	29
4.4 Test 4: (Oh et al., 2024)	30

4.5	Test 5: (Krause et al., 2008)	30
4.6	Test 6: (Kessler & and Grosse, 2023)	31
4.7	Test 7: (Popescu & Täljsten, 2020).....	32
5	Methodology	34
5.1	General.....	34
5.2	Mock-Up Specimen Design.....	34
5.2.1	Specimen 1	34
5.2.2	Specimen 2	36
5.2.3	Specimen 3	37
5.2.4	Specimen 4	38
5.3	Fabrication process of mock-up specimens.....	38
5.4	Testing objectives.....	40
5.4.1	Overall Testing Strategy	40
5.4.2	Data Collection Protocols	41
6	Current NDT Inspection Practice	42
6.1	Human inspector work flow.....	43
6.1.1	Phase 1: Instrument Setup and Calibration.....	43
6.1.2	Phase 2: Field Data Acquisition	44
6.1.3	Phase 3: Automatic processing by commercial software	46
6.1.4	Phase 4: Visual interpretation and expert judgment.....	49
6.1.5	Phase 5: Decision-making and reporting	50
6.2	Limitations of reliance on processed data.....	50
6.2.1	Subjectivity and operator dependency	50
6.2.2	Device-specific processing biases	51
6.2.3	Probability of missing anomalies	51
6.3	Motivation for machine learning on raw data	52
6.3.1	Fundamental premise: Raw data preserves complete physical information	52
6.3.2	Automatic learning of subtle, multi-dimensional patterns	52
6.3.3	Objectivity, consistency, and quantified confidence.....	53
6.3.4	Scalability and real-time deployment potential.....	54
6.3.5	Hybrid workflow: ML as a decision-support, not replacement	54
7	Machine learning applications for automated NDT analysis.....	55
7.1	Overview of three NDT modalities	55
7.2	Recent advances in ML for NDT applications	56
7.2.1	Ground-Penetrating Radar (GPR).....	56
7.2.2	Ultrasonic Pulse Echo (UPE).....	57
7.2.3	Impact Echo (IE)	57
7.2.4	Multi-Modal integration	58
7.3	Machine learning architectures and model development	58
7.3.1	Model selection criteria and architectures	58

7.3.2	Support Vector Machine (SVM).....	58
7.3.3	Random Forest (RF)	60
7.3.4	1D-Convolutional Neural Network (1D-CNN) – Spectral analysis	62
7.4	Feature extraction and selection	64
7.4.1	Philosophy: Physics-Informed feature engineering.....	64
7.4.2	GPR raw signal features.....	65
7.4.2.1	Extraction Methodology	65
7.4.2.2	Feature categories and physical interpretation	66
7.4.3	UPE Features	69
7.4.3.1	Unique aspects of MIRA data	69
7.4.3.2	Raw A-scan features.....	70
7.4.4	Impact Echo (IE) acoustic features.....	73
7.4.4.1	IE fundamentals	73
7.4.4.2	Feature extraction from a single measurement point	73
7.4.4.3	Spatial context features (Unique to IE)	75
7.4.4.4	Summary of IE Feature Set.....	76
7.4.5	Feature selection and dimensionality reduction	77
7.4.5.1	Feature Importance Analysis (Random Forest)	77
7.4.5.2	Correlation Analysis	78
7.5	Data Augmentation.....	79
7.5.1	GPR augmentation	79
7.5.2	UPE augmentation	80
7.5.3	Impact Echo (IE) augmentation	81
8	Model training and hyperparameter tuning.....	82
8.1	Training configuration and hyperparameter tuning.....	82
8.1.1.1	Support Vector Machine (SVM).....	82
8.1.1.2	Random Forest (RF)	83
8.1.1.3	1D-Convolutional Neural Network (1D-CNN)	83
8.1.1.4	U-Net (GPR Segmentation Baseline)	84
8.1.2	Cross-Validation strategy	84
8.1.3	Performance evaluation metrics	84
8.1.3.1	Understanding the confusion matrix	84
8.1.3.2	Quantitative performance metrics	85
8.1.3.3	Results and model comparison.....	86
8.1.4	Model selection recommendations	88
8.2	Integration and deployment	88
8.2.1	3D visualization system and multi-modal data fusion	88
8.2.2	Confidence-weighted decision fusion.....	96
8.2.3	Benefits of multi-modal approach.....	97
8.2.4	Quality assurance and system validation.....	97

9	Technology validation; Pilot Case: Gisund Bridge	99
9.1	Introduction	99
9.2	Condition assessment of post-tensioned concrete bridges	100
9.3	Results from NDT investigation	101
9.3.1	Identification of tendon ducts	101
9.3.2	Grouting status along tendon number 31_North side	102
9.3.2.1	UPE measurements	102
9.3.2.2	Machine Learning Analysis and Classification	103
9.3.3	Verification through Destructive Testing	108
9.4	Key Validation Outcomes	112
10	Conclusions and future development	114
10.1	Discussion and limitations	114
10.1.1	Advantages of the raw data approach	114
10.1.2	Limitations and challenges	115
10.2	Opportunities for future development	116
10.2.1	Immediate priority: Invator implementation and inspector training	116
10.2.2	Machine learning model improvements	118
10.2.3	Automated data collection (longer-term development)	119
10.3	Market possibilities	119
10.4	Closing remarks	120
	References	121
	Appendix - A	124

BILAGOR

Click or tap here to enter text.

Click or tap here to enter text.

Click or tap here to enter text.

Click or tap here to enter text.

Summary

This project developed and validated a machine learning–based decision-support framework for condition assessment of concrete structures, with specific focus on detecting internal defects in post-tensioned tendon ducts. The work addresses a central limitation in current inspection practice: although non-destructive testing (NDT) methods generate large volumes of subsurface data, interpretation remains largely manual, experience-based, and subject to variability. The project therefore aimed to transform raw NDT measurements into structured, data-driven damage assessments.

The methodology integrates three complementary NDT techniques. Ground Penetrating Radar (GPR) was used for locating and mapping tendon duct geometry, Ultrasonic Pulse Echo (UPE) for detecting voids and material discontinuities through shear-wave propagation, and Impact-Echo (IE) for resonance-based identification of internal anomalies. Because the methods rely on different physical principles, they provide distinct but complementary information. Their combined use improves detection reliability compared with single-method inspections.

A key methodological choice was to analyse raw signal data rather than relying solely on processed imagery from commercial equipment software. Raw data retains the full physical signal content, allowing machine learning models to detect subtle patterns associated with voids, incomplete grouting, and internal deterioration. This approach reduces dependence on proprietary processing algorithms and supports more standardized and transferable interpretation workflows.

To enable controlled validation, dedicated mock-up specimens were designed and fabricated with known internal conditions. These included fully grouted ducts, partially grouted ducts, and air-filled void scenarios. The specimens allowed repeatable testing, calibration of analysis procedures, and benchmarking of model performance under verified conditions. Laboratory validation was complemented by field investigations on real bridge structures to ensure practical applicability.

Machine learning development included physics-informed feature extraction to ensure interpretability of signal attributes. Multiple model architectures were evaluated, including Support Vector Machines, Random Forest classifiers, and 1D Convolutional Neural Networks for spectral analysis. Structured training, hyperparameter tuning, and cross-validation workflows were implemented to ensure robust classification performance.

Beyond single-modality analysis, the project implemented multi-modal data fusion. Outputs from GPR, UPE, and IE were combined through confidence-weighted decision logic, producing integrated defect probability assessments. This improves reliability in areas where individual methods face access limitations, signal interference, or material constraints. Visualization and deployment aspects were also addressed. A 3D visualization environment was developed to spatially map detected anomalies within structural geometry. This supports clearer communication between inspectors, engineers, and asset owners and creates a foundation for integration into digital asset management and structural assessment platforms.

Field validation was carried out on operational infrastructure, where machine learning classifications were compared with conventional NDT interpretations and selectively verified through destructive testing. The results demonstrate that the framework functions effectively as a decision-support tool, improving interpretation consistency and reducing subjectivity while maintaining the role of expert engineering judgement. Identified limitations include the need for larger labelled datasets, dependence on high-quality raw data acquisition, and the requirement for broader validation across additional structural typologies. Future development pathways include dataset expansion, increased automation in data acquisition, and deeper integration of analytics within inspection hardware ecosystems.

In summary, the project delivers a validated, physics-informed machine learning framework capable of analysing multi-modal NDT raw data to support objective and repeatable condition assessment of post-tensioned concrete structures.

Sammanfattning

Projektet har utvecklat och verifierat ett maskininlärningsbaserat beslutsstödsverktyg för tillståndsbedömning av spännarmerade betongkonstruktioner, med särskilt fokus på att identifiera interna defekter i foderrör hos efterspända system. Bakgrunden är en central begränsning i dagens inspektionspraxis: även om oförstörande provningsmetoder (NDT) kan generera stora mängder mätdata från konstruktionens inre, baseras tolkningen fortfarande till stor del på manuell bedömning och individuell erfarenhet. Projektets huvudsakliga ambition har därför varit att skapa ett datadrivet analysstöd som ökar objektivitet och repeterbarhet i skadebedömningar.

Arbetet bygger på en kombination av tre etablerade NDT-metoder: georadar (GPR) för lokalisering och kartläggning av spännkanaler, ultraljud (UPE) för detektering av hålrum och diskontinuiteter baserat på skjuvvågsutbredning samt Impact-Echo (IE) för resonansbaserad identifiering av interna avvikelser. Metoderna kompletterar varandra genom att de bygger på olika fysikaliska principer och därmed ger olika informationsinnehåll. Genom att integrera resultaten kan tillförlitligheten i skadeidentifieringen ökas jämfört med enskilda metoder.

En viktig metodologisk utgångspunkt har varit analys av rådata i stället för enbart processade bilder från kommersiella mjukvaror. Rådata innehåller fullständig fysikalisk signalinformation, vilket möjliggör identifiering av subtila mönster kopplade till hålrum, bristfällig injektering och andra interna defekter. Detta minskar beroendet av utrustningsspecifika algoritmer och skapar bättre förutsättningar för standardiserad och jämförbar analys.

För verifiering utvecklades särskilda provkroppar med kända defektförhållanden, inklusive fullt injekterade kanaler, delvis fyllda kanaler och luftfyllda hålrum. Dessa mock-ups möjliggjorde kontrollerad datainsamling, modellträning och prestandautvärdering under kända randvillkor. Laboratorieresultaten kompletterades med fältundersökningar på verkliga brokonstruktioner för att säkerställa metodens praktiska tillämpbarhet.

Maskininlärningsutvecklingen omfattade fysikbaserad feature-extraktion samt test av flera modellarkitekturer, däribland Support Vector Machines, Random Forest och konvolutionsneuronnätverk. Systematiska tränings- och valideringsstrategier implementerades för att säkerställa robust klassificering. Utöver enskild metodanalys utvecklades även multimodal datafusion, där resultat från olika NDT-metoder vägdes samman genom konfidensbaserad beslutslogik för att generera en samlad skadebedömning.

Projektet inkluderade också utveckling av visualiseringsverktyg. En tredimensionell presentationsmiljö möjliggör rumslig kartläggning av identifierade avvikelser, vilket underlättar kommunikation mellan inspektörer, konstruktörer och beställare samt skapar förutsättningar för integration i digitala förvaltningssystem.

Fältvalidering genomfördes på befintlig infrastruktur där maskininlärningsbaserade klassificeringar jämfördes med traditionell tolkning och, där möjligt, verifierades genom destruktiva kontroller. Resultaten visar att tekniken fungerar som ett effektivt beslutsstöd som ökar tolkningskonsistens och minskar subjektivitet, utan att ersätta inspektörens expertbedömning.

Identifierade begränsningar rör främst behovet av större träningsdatamängder, krav på högkvalitativ rådatainsamling samt vidare verifiering på fler konstruktionstyper. Framtida utveckling omfattar utökade databaser, ökad automatisering i datainsamling och djupare integration mellan analysverktyg och mätsystem.

Sammanfattningsvis har projektet etablerat en verifierad metodik där maskininläring kombineras med multimodal NDT-data för att möjliggöra mer objektiv, repeterbar och datadriven tillståndsbedömning av efterspända betongkonstruktioner.

1 Introduction

The introduction sets the stage for understanding both the engineering challenge and the technological solution explored in the research presented in this report. Post-tensioned concrete bridges represent a crucial part of civil infrastructure, offering economic and structural advantages through steel tendons that compress the concrete and enable long bridge spans. However, these systems have a critical vulnerability: after tensioning, the steel tendons must be protected from corrosion by injecting cement grout into the ducts that house them. When grouting is incomplete, voids remain inside the ducts, creating pathways for moisture and eventual corrosion that can compromise structural safety.

The challenge lies in detecting these hidden voids without damaging the structure. Direct visual inspection is impossible since the ducts are embedded within the surrounded concrete. This is where non-destructive testing (NDT) methods become essential assessment techniques. Ground-penetrating radar, ultrasonic testing, and impact-echo can probe the interior without drilling or cutting. But interpreting the complex signals from these instruments typically requires experienced specialists and remains somewhat subjective. The current research addresses that limitation by combining NDT measurements with machine learning algorithms that can automatically interpret raw data from NDT devices and identify void signatures in the data. The introduction first explains the engineering background of post-tensioned systems and why void detection matters for safety. It then outlines the project's objectives: developing systematic procedures for ML-enhanced NDT inspection, creating practical guidelines for field inspectors, and bridging the gap between laboratory validation and real-world bridge assessment.

1.1 Background

Post-tensioned (PT) concrete systems have been widely used in bridge construction since the late 1950s (Tinkey & Olson, 2007). These systems are favoured for their economic efficiency, ability to span long distances, increased structural capacity, and relatively straightforward construction processes (Im & Hurlbaas, 2012). The technique involves installing steel tendons or steel rods inside ducts—either internal or external to the concrete section—which are tensioned after the concrete has hardened, introducing a compressive force that counteracts tensile stresses. The underlying concept of prestressing concrete is to overcome its inherent weakness in tension preventing it to crack under service loads, particularly due to bending moments caused by dead and live loads. These cracks reduce stiffness, increase long-term deformations, and expose reinforcement to corrosive environments. To mitigate these effects, compressive stresses are intentionally introduced into concrete through prestressing techniques. Prestressing can be implemented in two main ways:

1. **Pre-tensioning**, where the tendons are tensioned before casting the concrete.
2. **Post-tensioning**, where the tendons are tensioned after the concrete has cured.

This report focuses on the latter—post-tensioning—where tendons are threaded through ducts (typically made of metal or plastic), and stressed only after the concrete has gained sufficient strength. A key advantage of this method is the flexibility it offers during on-site construction, particularly the ability to implement curved tendon profiles. Curved tendons are essential in multi-span continuous bridges, where the cable path is typically elevated over supports (to resist negative moments) and lowered mid-span (to counteract positive moments), thereby optimizing internal force distribution. However, a major vulnerability of post-tensioned systems arises after stressing: the exposed tendons must be protected against corrosion by injecting grout—typically a cement-based slurry—into the ducts. The purpose of the grouting is to bond the tendons to the ducts, increase composite action between ducts and concrete, prevent corrosion, and ensure long-term durability. Incomplete grouting and voids in tendon ducts have been identified as a major durability risk in post-tensioned structures (Sharma et al., 2024), i.e. breakage of the ducts and possible future failure of the cable/rod which in worse cases could lead to a collapse of a part of or the entire structure.

This issue is at the heart of the current research effort. Incomplete grouting often leads to voids inside the tendon ducts, which in turn become potential initiation sites for corrosion. Detecting such internal voids non-destructively is a significant technical challenge. Direct visual inspection is typically impossible without invasive measures. However, recent advances in Non-Destructive Testing (NDT) offer potential for detecting and mapping these voids indirectly, providing valuable input for structural condition assessment. There is also a concern that extended lengths of ducts may be left ungrouted, which not only increases the risk of corrosion but may negatively affect load transfer mechanisms. The detailed structural implications of such unfilled ducts—especially on stress development, tendon anchorage, and duct-concrete interaction—will be explored in a subsequent chapter of this report.

The primary objective of the current work is to develop and verify NDT procedures for detecting voids in post-tensioned tendon ducts, with the goal of supporting practical and reliable condition assessments to the inspectors. To this end, the project employs specially designed mock-up specimens with known void configurations, enabling repeatable testing and robust verification of NDT methods under controlled conditions. Three NDT techniques, Ground Penetration Radar (GPR), Ultrasound (MIRA), and Impact Echo (IE), performed on the mock-up samples and RAW data from these analysed by Machine Learning (ML) methods to develop a decision-support for human inspectors at field.

1.2 Objectives

The main objective of this work is to develop and validate a framework for applying ML to NDT data in order to improve the accuracy, efficiency, and reliability of structural condition assessment. More specific to create a decision support tool for the inspectors. The project addresses limitations in current practice, which relies heavily on manual interpretation, fragmented datasets, and operator-dependent evaluations. To support this goal, the following sub-objectives have been established:

- **Develop systematic, verifiable workflows**
The project aims to establish procedures for acquiring, processing, and interpreting NDT data in combination with ML techniques. Controlled laboratory tests and selected field case studies provide a basis for objective verification of model performance and method reliability.
- **Define guidelines for data-driven defect detection and assessment**
By comparing conventional inspection outcomes with ML-assisted analysis, the project seeks to produce step-by-step guidelines for inspectors. These guidelines will help identify and interpret signals associated with typical damage scenarios (e.g., voids, cracks, delamination, and corrosion indicators).
- **Enable training and knowledge transfer through data-driven tools**
The project contributes to capacity-building by creating training material and demonstrators that show how ML-enhanced NDT can support inspectors. This will promote more consistent interpretation across teams, reduce subjectivity, and improve decision support.
- **Bridge the gap between laboratory validation and field implementation**
While initial validation is carried out on mock-ups and controlled specimens, the methods are further tested in real-world inspections of bridges and other concrete infrastructure. This ensures a clear transition from proof-of-concept to practical application.
- **Provide data-driven decision support to strengthen inspector assessments**
The project aims to develop machine-learning-based tools that offer a consistent, objective second opinion during condition assessment. These tools will help reduce human error, highlight overlooked indicators, and enhance the reliability of inspection outcomes.

2 Literature review and existing practices

This section explores why incomplete grouting in post-tensioned bridges is such a serious problem and why detecting hidden voids is so technically challenging. When cement grout isn't properly injected into the ducts that house the prestressing tendons, the resulting voids expose steel to moisture and oxygen, leading to corrosion that can cause sudden structural failures—sometimes within just 10-20 years after construction rather than the intended decades of service life. The literature documents two main culprits: segregated or defective grout that itself becomes corrosive, and air voids created by improper grouting procedures that leave sections of duct completely empty.

But finding these hidden defects is harder than it sounds. The section walks through the practical obstacles that make void detection difficult: metallic ducts that block radar signals, curved tendon paths requiring multiple scan angles, ducts buried deep inside massive concrete sections or hidden behind asphalt overlays, closely-spaced tendons whose signals overlap and confuse interpretation, and the physical challenges of working inside cramped bridge box girders while holding equipment overhead for hours. These aren't just technical problems—they're human work environmental factors that affect inspection quality and reliability.

The chapter concludes by explaining how these challenges shaped the current research approach, leading to the selection of three complementary NDT methods—Ground Penetrating Radar for mapping duct locations, ultrasonic testing for detecting internal voids, and Impact-Echo for confirmation—that together provide more reliable results than any single technique alone.

2.1 Importance of grouting and risks of voids in PT systems

In post-tensioned concrete structures, grout serves three critical functions: it protects prestressing tendons from corrosion, provides bond between the tendon and surrounding concrete, and contributes to the system's stiffness and effective force transfer. However, construction records and field inspections from various bridge inventories have consistently revealed instances of poor or incomplete grouting, resulting in voids within the tendon ducts. These voids—filled with air, bleed water, or moisture—undermine all three functions and significantly compromise structural integrity.

Voids expose prestressing strands to oxygen and chlorides, creating an environment conducive to corrosion. This is particularly hazardous in aggressive environments such as coastal regions or areas with frequent use of de-icing salts. Studies and documented failures in both Europe and the U.S. have linked voids directly to cases of tendon rupture and sudden structural degradation. According to (Menga et al., 2023), incomplete grouting is the predominant root cause of corrosion in post-tensioned systems. The problem is most acute at typical weak points in the tendon layout—such as high points, curved ducts, and anchorage zones—where air naturally collects and grout flow is more difficult to control.

The corrosion that initiates in these regions often leads to severe steel pitting and localized section loss in the steel strands. Because this type of degradation progresses inside the structure and without surface indicators, it often goes undetected until a sudden and brittle rupture occurs. When a single strand fails in an ungrouted region, force redistribution becomes highly unreliable, potentially triggering progressive failure of neighbouring tendons. This mode of failure has been observed in structures as early as 10–20 years into service—far earlier than the intended design life—demonstrating the severe durability risks posed by ungrouted voids. Characteristics of strand corrosion are provided in Figure 1. In the next section two primary mechanisms of corruptions from the literature will be discussed.

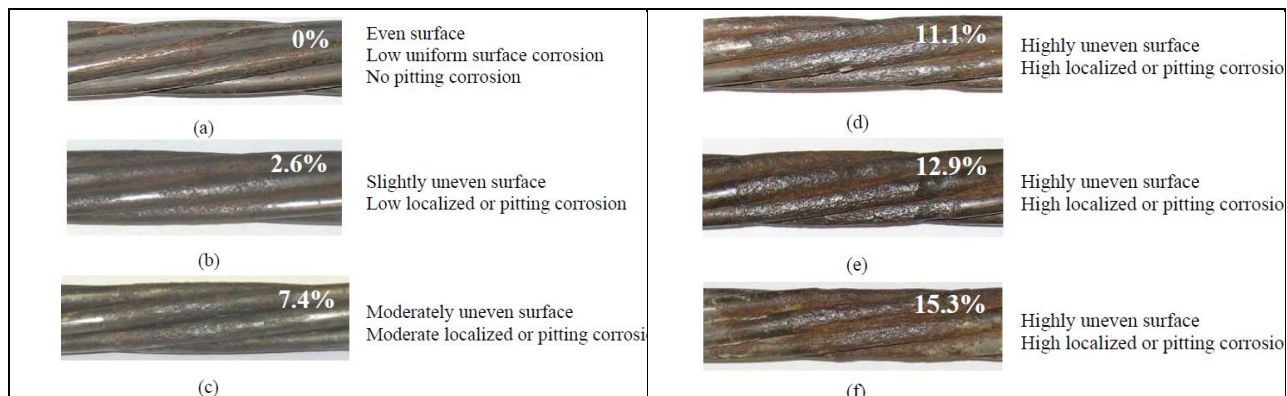


Figure 1 Photographs showing corrosion level and surface characteristics of strands. (Trejo et al., 2009)

2.1.1 Segregated or defective grout

This phenomenon is characterized by the appearance of a white, soft, unhardened paste—an exudation product that separates from the intended grout mix. Such exudate is highly enriched in water and sulphates and exhibits an extremely high pH, often between 13 and 14 (Carsana & Bertolini, 2015). Though cementitious, this segregated material fails to protect the strands and instead becomes an aggressive corrosive medium. Laboratory and field investigations have shown that strands in direct contact with this material suffer from localized corrosion, particularly pitting (Bore et al., 2010; Carsana & Bertolini, 2015). The combination of high alkalinity and ionic content fosters deep corrosion pits, acting as notches that concentrate stress.

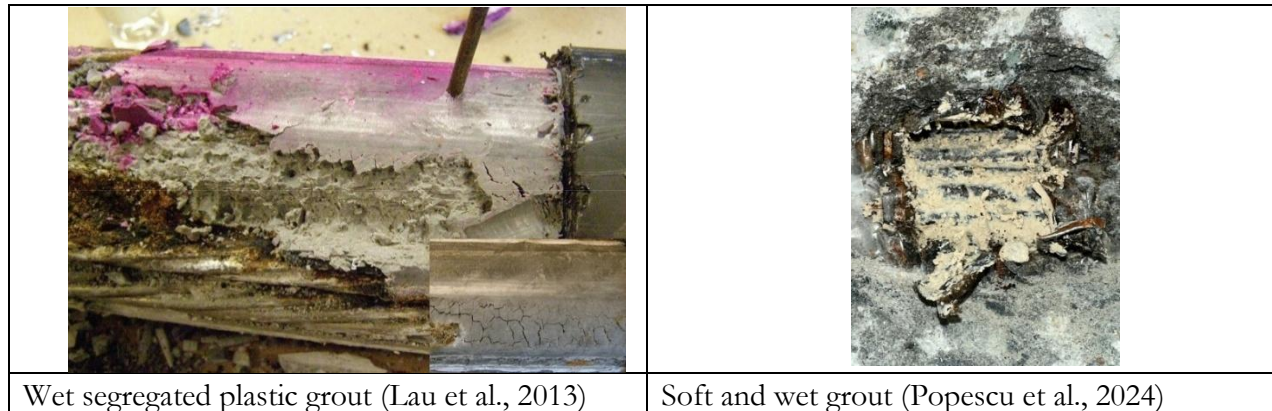


Figure 2 Examples of segregated and defective grout observed in post-tensioned tendon ducts

This localized degradation significantly weakens the steel, increasing the likelihood of brittle fracture without prior visible damage. In one documented case, a tendon failed in under two years due to such conditions (Bore et al., 2010; Carsana & Bertolini, 2015). The corrosive action is exacerbated in high-strength prestressing steels, which are notably more sensitive to pitting and hydrogen-related cracking than conventional reinforcement. As explained by (Nürnberg, 2002), such strands under tension are vulnerable to stress-corrosion cracking, hydrogen embrittlement, and corrosion fatigue, especially when exposed to aggressive bleed water environments. This bleed water, often trapped during construction, may contain high concentrations of sulphates, chlorides, and potassium ions, all of which contribute to the corrosive potential. Once corrosion initiates in these strands, even small pit depths (e.g., 0.2–0.6 mm) can dramatically reduce their mechanical performance, making rupture under service loads much more likely (Nürnberg, 2002).

2.1.2 Voids due to grouting errors

The second, and more prominent, cause of corrosion in post-tensioned tendons is the presence of voids in the grout. These voids can form due to several issues such as entrapped air pockets, grout bleeding, or improper grouting procedures (Im & Hurlbaas, 2012). Air voids may be trapped if the grouting vents are closed too early—before the ducts are completely filled—or as a result of air naturally migrating toward the high points in the tendon duct system, often located near anchorages (Cavell & Waldron, 2001). Additionally, grout bleeding allows water to separate from the grout mixture. This water can move along the interface between wire strands and eventually accumulate near anchorages where it may evaporate, leading to increased risk of corrosion (Hansen, 2007).

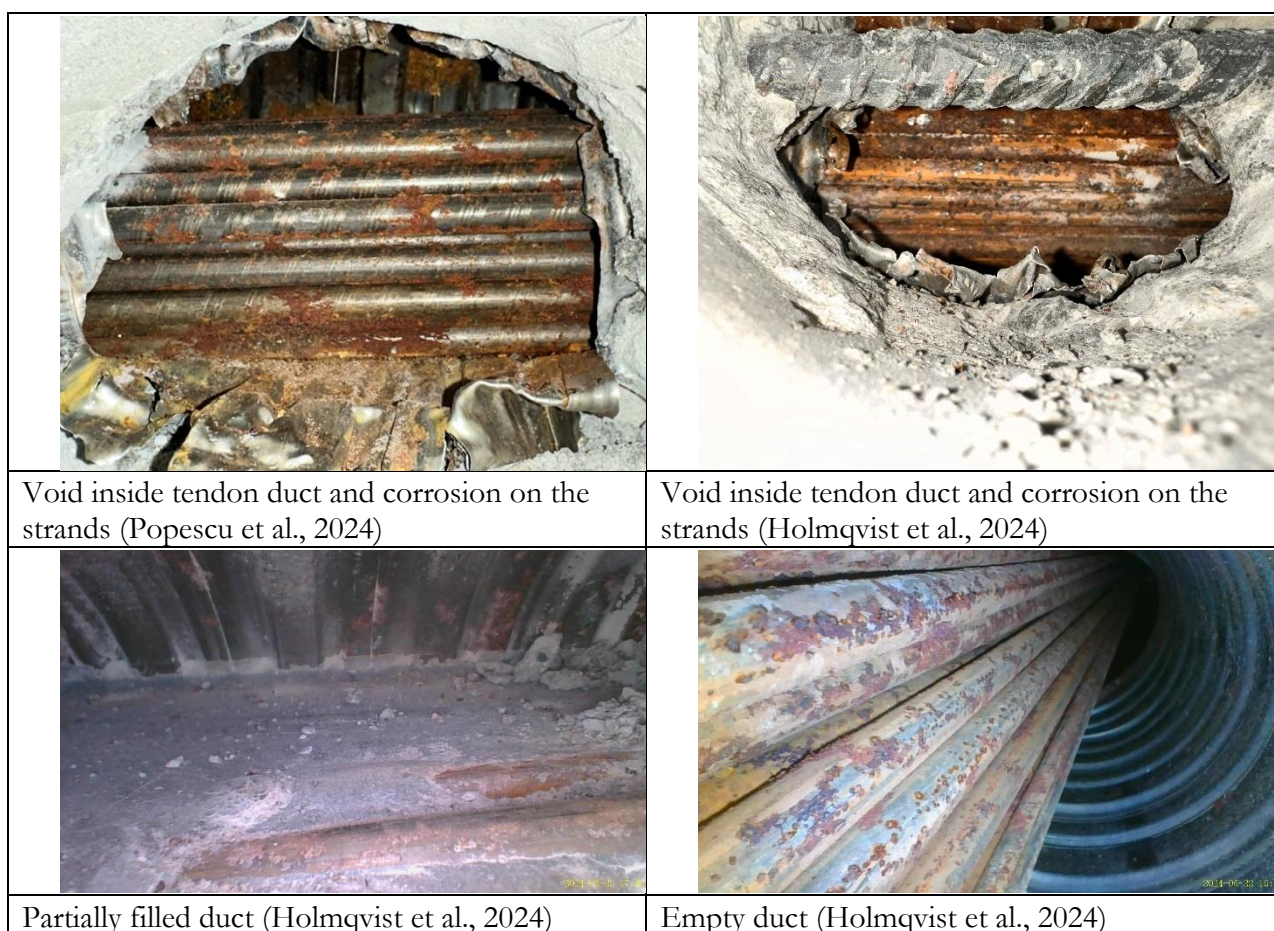


Figure 3 Examples of voids and missing grout in post-tensioned tendon ducts

Regardless of the specific cause, the presence of these voids leads to two primary concerns. First, the prestressing strands within these voids remain ungrouted and might thereby be exposed to environmental aggressors, in particularly chlorides. These chlorides can penetrate from external sources such as de-icing salts or seawater spray and also from airborne marine aerosols or chloride-rich construction materials like sand used in concrete production (Cavell & Waldron, 2001; Minh et al., 2007; Nürnberger, 2002). Once these aggressive agents reach the exposed steel surface, localized corrosion (especially pitting) may occur, which drastically reduces the fatigue resistance and load-bearing capacity of prestressing steel—even at small corrosion pit depths of 0.2–0.6 mm (Nürnberger, 2002). Moreover, in crevice-like environments such as those formed by voids or poor grout contact, the risk of hydrogen-induced stress corrosion cracking (H-SCC) increases, especially under tensile stress conditions typical in post-tensioned systems (Nürnberger, 2002).

Second, the presence of voids prevents the intended mechanical interaction between the grout and the prestressing strands. If a strand breaks, proper stress redistribution to neighbouring strands cannot occur unless the grout is continuous and in full contact with the steel. This lack of redundancy in the load path may lead to disproportionate failures, especially near anchorage zones where the stresses are highest (Martin et al., 2001).

Although limited research exists that directly quantifies the effect of grout voids alone on the ultimate load-carrying capacity of post-tensioned members, most studies suggest that voids primarily become critical when combined with corrosion. In such cases, the resulting loss of cross-sectional area and prestressing force in the tendons significantly compromises structural integrity. In otherwise intact, grouted systems, even if a strand fails, the surrounding grout may enable a re-anchoring mechanism—an effect observed in controlled demolitions. This re-anchoring allows the fractured tendon to regain some anchorage in the grout and partially restore the prestressing force.

Experimental observations from controlled demolition of post-tensioned bridges (Buchner & Lindsell, 2004) and tests involving deliberate tendon cuts (Belhadj & Waldron, 1993; Dai et al., 2021; Felstead & Lindsell, 1981) confirm that such re-anchoring behaviour can occur in fully grouted ducts. However, the effectiveness of this mechanism strongly depends on the grouting condition: full grouting can facilitate re-anchoring, whereas completely ungrouted or partially grouted ducts inhibit it.

(Wang et al., 2014) also found that insufficient grouting has minimal influence on the initial, linear-elastic response before cracking. Once cracking initiates, however, grout deficiencies lead to reduced section stiffness, lower ductility, and a decrease in ultimate flexural strength. Therefore, while grout voids alone may not drastically affect the structure in the uncracked state, their presence in combination with tendon rupture represents a dangerous scenario. Such cases must be carefully considered when evaluating the residual capacity of post-tensioned members.

Given the central role of grout in ensuring both the durability and mechanical performance of post-tensioned systems, the presence of voids within tendon ducts emerges as a critical vulnerability—effectively the root cause behind many long-term service issues. Voids not only negate corrosion protection but also disrupt force transfer mechanisms and compromise redundancy in the event of tendon rupture. These problems are particularly concerning because they develop invisibly, without external indications, and are often discovered only after substantial deterioration has occurred. Traditional inspection techniques fall short in addressing this hidden threat, leaving NDT as the only viable solution for timely identification and assessment of internal voids. However, the effectiveness of NDT depends not just on the hardware used, but equally on the post-processing methods and the interpretive skill of the human operator. Variability in detection accuracy can result from inconsistent application, lack of training, or insufficient calibration against known conditions. *Therefore, the present study is both timely and necessary—it aims to validate NDT techniques under controlled conditions, develop repeatable testing procedures, and create step-by-step guidance for interpreting results.* By doing so, this work addresses a major gap in current inspection practices and contributes toward safer, more reliable condition assessments of post-tensioned concrete infrastructure.

2.2 Detection of voids in tendon ducts: Challenges of detecting internal defects

Post-tensioned (PT) concrete systems present unique challenges for NDT, particularly in detecting voids within grouted ducts. The presence of tendons inside metallic ducts, often embedded in large cross-sections or covered with overlays, limits the effectiveness of surface-based techniques. Despite these difficulties, the literature has explored and applied a variety of NDT methods for internal void detection. Common approaches include GPR, Ultrasonic Pulse Echo (UPE), IE, Radiography, and newer hybrid or tomographic methods. Each method offers certain advantages and limitations depending on duct accessibility, materials, geometry, and signal clarity.

2.2.1 Duct material and geometry

Metallic ducts attenuate or reflect signals, reducing effectiveness of some techniques (e.g., GPR). GPR relies on the propagation of high-frequency electromagnetic waves, which reflect at boundaries where there is a significant change in dielectric properties. When these signals encounter a metallic duct, the contrast in electrical conductivity between concrete and steel is so high that nearly 100% of the electromagnetic energy is reflected at the duct's outer surface. This reflective behaviour causes the duct to act like a mirror, preventing the waves from penetrating into their interior. Consequently, while GPR can accurately detect the location and alignment of metallic ducts, it cannot reveal the internal condition—whether the duct is fully grouted, partially voided, water-filled, or affected by internal corrosion.

Regarding duct geometry, while it does not inherently prevent detection, it introduces practical complexities. Curved tendon profiles, which are common in multi-span post-tensioned structures, require multiple scans to accurately trace the tendon path. As shown in Figure 4, each scan provides only a sectional view (2D-view) of the duct layout. Without sufficient spatial coverage, it becomes difficult to reconstruct the full tendon profile, especially when they deviate from straight lines or interact with surface reinforcement.

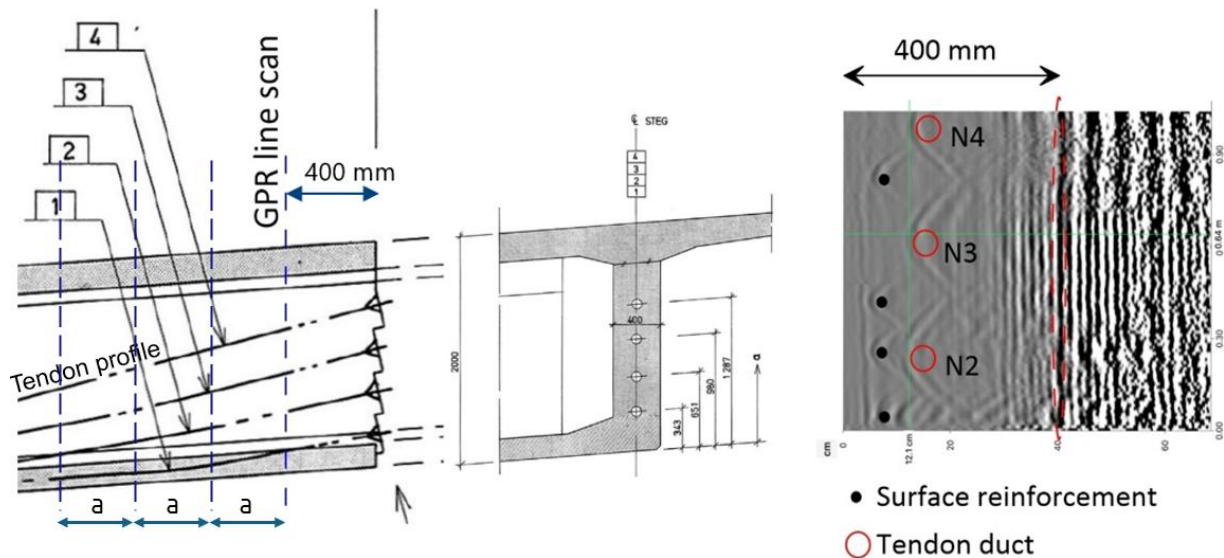


Figure 4 Example of GPR scan slices spaced (a) used to reconstruct curved tendon profiles

Interpreting these results demands experience and careful alignment of signal reflections across several scan slices or views. However, accurate reconstruction of the tendon profile requires more than just correlating signal patterns across multiple scans. Reflections detected by GPR may correspond to either tendon ducts or conventional reinforcement (see Figure 4) and distinguishing between the two can be complicated. To resolve these ambiguities, a third critical input is often necessary: reference to original or idealized as-built drawings. These drawings help validate the expected tendon layout and improve the accuracy of interpretation. This introduces added complexity, as correct analysis requires not only NDT expertise but also a solid understanding of post-tensioning design and the ability to interpret structural drawings accurately.

Failure to correctly identify tendon locations may result in an incorrect tendon profile, which can have implications in subsequent steps. For instance, if UPE or other focused testing is applied based on a misidentified duct location, it may yield irrelevant results. Worse, if invasive verification such as coring is attempted at the wrong location, there is a real risk of unintentionally intersecting an active tendon. Thus, while not a fundamental barrier, it does demand a more rigorous and structured scanning strategy to ensure reliable results.

2.2.2 Access limitations

Post-tensioned ducts are frequently embedded deep within massive concrete structure or situated in locations that are physically difficult to access – such as the corners of box girders or zones where ducts pass over supports. These areas are often further obstructed by concrete overlays, asphalt surfacing, or other structural components. Additionally, a shadowing effect can occur when multiple ducts are installed in parallel: signals may be partially or fully blocked by the first duct, rendering it difficult or impossible to assess ducts positioned directly behind.

As illustrated in Figure 5, the left side shows ducts embedded in the upper deck slab – the traffic-bearing surface – while the right side shows ducts in the bottom slab, which is accessible from inside the box girder. Typically, inspectors have physical access only from the inside of the box-girder, unless external scaffolding or lifting platforms are used.

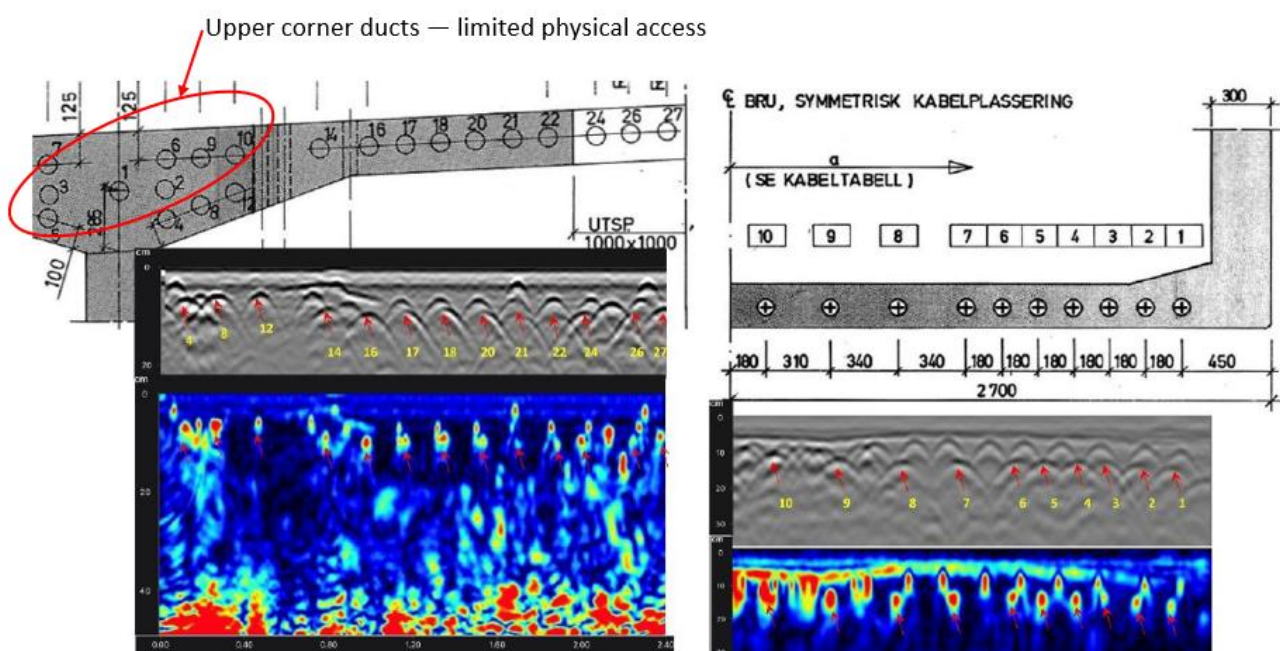


Figure 5 Cross-section of a post-tensioned box girder segment (top) with corresponding GPR scan results (bottom)

Access to the upper tendon zones, especially those placed near the deck-web intersection (corners of the box girder), is severely constrained. These areas are covered by asphalt surfacing, which limit physical access. Ducts positioned near the deck–web intersection are inaccessible from within the box girder and typically cannot be inspected from above without traffic disruption, making them effectively uninspectable without major intervention.

Moreover, as seen in the GPR scans in Figure 5, even when scanning is possible, signal strength is highest from ducts closest to the surface. Ducts positioned deeper, especially those obscured by other layers or reinforcement or ducts, suffer from attenuation or signal shadowing. This limits the ability to assess all tendon paths reliably.

2.2.3 Signal interference

In regions where tendon ducts are closely spaced – particularly around corners or congested zones – the ability to distinguish individual ducts using NDT methods becomes significantly reduced. Reflections from adjacent tendons may overlap, producing signal clusters that are difficult to interpret with confidence (see Figure 5). Moreover, not all anomalies detected in the scan are directly related to tendon defects. Signal patterns can also be influenced by previous repair zones, delamination, or defects in the parent concrete, such as honeycombing or segregation pockets near the duct. These features can mimic the signal response of voids or grouting deficiencies, leading to false-positive indications.

2.2.4 Resolution and sensitivity

Given the current state of practice, it remains difficult to reliably distinguish between poor-quality grout, partially filled ducts, and completely voided sections using available NDT methods. While both ultrasonic and electromagnetic techniques are generally effective in detecting the presence of anomalies, they often lack the resolution or interpretive confidence needed to assess the size, type, or severity of the defect. As a result, a signal anomaly may indicate anything from minor grout segregation to a fully empty duct, but this cannot be determined with certainty. This limitation increases the need for exploratory verification, making it almost mandatory to confirm NDT findings through targeted coring and endoscopic inspection. Without such follow-up, it remains difficult to confidently assess the true condition of the tendon duct.

In addition to the general limits of spatial resolution, there is also uncertainty in the precise location of detected anomalies within the duct cross-section. Tendons are not always centred within their ducts; they may rest eccentrically, pressed to one side. In such cases, if a void is located behind the tendon (relative to the accessible face), it may still be detected by ultrasonic methods. However, confirmation using coring and videoscope becomes extremely difficult. Coring from the same side will typically strike the tendon bundle before reaching the void, making it nearly impossible to visually validate the anomaly. This limitation is not a failure of detection per se, but a limitation in verification due to restricted access and spatial interpretation. It underscores the challenge of relying on one-sided access in box girders and highlights the need for caution when dismissing NDT findings that cannot be intrusively confirmed.

2.2.5 Human factors and operator dependence

Inspections of post-tensioned structures are often carried out in difficult environments—at height, sometimes over open water, and frequently under harsh weather conditions. These structures typically require scaffolding (Figure 6), which limits both working time and physical comfort for proper data collection and review. Inside box girders (Figure 6), inspectors may need to operate in cramped positions, kneeling or sitting for prolonged periods in order to scan all relevant areas. In larger sections, mobile scaffolding is frequently used to reach high-placed ducts. Inspectors often have to operate with equipment held overhead, arms fully extended above shoulder height, which introduces significant physical strain and limits both precision and endurance. This setup not only slows down the inspection process but also increases the risk of incomplete data collection due to operator fatigue or limited manoeuvrability.



Figure 6 Practical challenges during NDT inspections of post-tensioned bridges, including access via scaffolding above water, confined interior spaces, and overhead scanning in constrained positions

Under such constraints, inspectors may experience fatigue and time pressure, increasing the likelihood of rushed decisions and reduced accuracy, particularly when evaluating unclear or ambiguous signal patterns.

In addition, modern NDT equipment often depends on tablets or smart devices for data acquisition and visualization. While convenient, these devices are prone to rapid battery drain in cold conditions, which can interrupt inspections or result in loss of real-time feedback.

Given these factors, it is evident that NDT results are highly dependent on the operator's skill and experience. To improve consistency and reliability, there is a growing need for structured inspection protocols, hands-on training with realistic mock-ups, and robust post-processing tools that can assist or validate field interpretations.

2.3 Basis for current study: findings

A detailed review of NDT methods was conducted by the authors. The evaluation included a wide range of techniques with respect to parameters such as field readiness, sensitivity to voids, portability, safety, and ease of interpretation.

Based on that review, a shortlist of practical and technically suitable methods was identified for further testing and verification within this project. The selection of NDT methods prioritized for further development in this project was guided by several practical and technical criteria. First, field applicability was considered essential; all selected techniques needed to function effectively under typical site constraints, such as limited access inside box girders, work at height, or time-restricted inspections performed from scaffolding. Portability and ease of setup were thus important factors. Equally important was the sensitivity of the method to internal duct anomalies. The ability to detect voids – whether air-filled, water-filled, or associated with poor grout quality – was prioritized. Since no single method is flawless, the value of combining complementary techniques was also emphasized. For example, GPR may be used for duct location and profiling, while UPE or IE methods provide a more direct indication of void presence. This multi-method approach allows cross-validation of findings, improving reliability and reducing the risk of false positives or negatives.

Another critical criterion was the interpretability of the collected data. Methods that provide clear signal outputs, imaging, or patterns that can be consistently interpreted were favoured over those requiring overly subjective analysis. However, the role of human operators remains vital. Therefore, techniques were preferred to support structured training and repeatable workflows, minimizing dependency on individual intuition or highly specialized expertise.

Safety and intrusiveness were also factored into the selection process. Preference was given to techniques that are non-invasive, pose minimal health risks, and ideally require access from only one side of the structure—a common limitation in bridge inspections. Finally, the level of technological maturity and commercial availability equipment/methods was also considered. The selected methods had to be sufficiently developed and available for practical deployment, even if further procedural refinement is needed for specific applications.

These include:

- **Ground Penetrating Radar (GPR):** Used to locate the tendon ducts and reconstruct their profile prior to further testing. Although limited in its ability to detect internal duct conditions in metallic sheaths, it provides critical spatial orientation and mapping, especially in curved or congested zones.
- **Ultrasonic Pulse Velocity (UPV) and Ultrasonic Tomography (UPE) (both transverse and longitudinal waves):** These were identified as the primary candidates for detecting voids and irregularities inside ducts. Their ability to detect changes in material stiffness and discontinuities made them promise for initial screening.
- **Impact Echo (IE):** Considered a complementary method to help narrow down the areas of interest identified by ultrasonic methods.

While other methods such as radiography, muon tomography, and acoustic emission, were reviewed, they were deemed less feasible due to safety concerns or complex interpretation requirements. These techniques remain of academic interest but are not prioritized for field deployment in this phase of the project.

3 NDT methods: Physical principles and data structure

Before diving into how machine learning analyses NDT data, it's essential to understand what that data actually represents and where it comes from. This chapter provides a technical reference for the three NDT methods used throughout this research: Ground Penetrating Radar (GPR), ultrasonic pulse echo (UPE) with the MIRA system and Impact-Echo (IE). Each method works on fundamentally different physics—GPR sends electromagnetic waves through concrete, ultrasonic testing uses mechanical sound waves, and Impact-Echo creates low-frequency stress waves with a mechanical impact—but all share the same goal of revealing what's hidden inside concrete without drilling or cutting.

For each method, the chapter explains the underlying physical principle that makes detection possible, describes the raw data format and structure that comes from the instruments, details the essential processing steps needed to transform raw signals into interpretable information, and provides the mathematical formulations that govern wave propagation and signal analysis. This technical foundation is important because it shows what information each method can actually measure and what its limitations are. Understanding how these instruments work and what their signals represent helps make sense of the machine learning analysis presented later—when models identify potential defects, it's useful to know what physical measurements they're responding to and why those measurements indicate problems.

3.1 Ground-penetration radar (GPR)

3.1.1 Physical Principle

GPR is an electromagnetic method that transmits short pulses of high-frequency radio waves—typically ranging from 200 MHz to 2.6 GHz—into concrete structures. The system consists of one or two antennas that move across the concrete surface. The transmitting antenna emits short electromagnetic pulses, while the receiving antenna records reflected energy from interfaces where the dielectric permittivity (ϵ_r) of materials changes, such as at concrete-air, concrete-steel, or concrete-water boundaries.

The fundamental principle relies on electromagnetic wave reflection and transmission at material interfaces. When waves encounter a boundary between materials with different dielectric properties, part of the energy reflects back toward the antenna while the remainder transmits into the second material. The amplitude and polarity of reflected waves depend on the dielectric contrast between materials: when this contrast is low, reflected energy is weak and detection becomes difficult. In the presence of metallic layers or dense reinforcing bars, wave penetration becomes significantly limited due to strong reflections and signal attenuation. The two-way travel time of reflected waves is governed by the electromagnetic wave velocity in the medium:

$$v = \frac{c}{\sqrt{\epsilon_r}}$$

where $c \approx 3 \times 10^8 \text{ m/s}$ is the speed of light in vacuum. Typical concrete has $\epsilon_r = 6\text{--}9$, yielding wave velocities around $v \approx 0.1 \text{ m/ns}$. The amplitude and polarity of reflections depend on the dielectric contrast at material boundaries. Metallic layers produce strong reflections and limit penetration depth due to high electrical conductivity.

GPR is commonly used to evaluate reinforced or prestressed concrete and masonry elements, detecting internal anomalies such as voids, delaminations, embedded metal or wood inclusions, tendon ducts, and thickness variations in structures accessible from only one side. However, determining precise geometric characteristics—such as the diameter of individual reinforcing bars—remains challenging due to limited resolution and complex signal scattering.

3.1.2 Data type and structure – SEG-Y format

Modern GPR systems often store the collected measurements in **SEG-Y (SEG-Y)** format, originally developed for seismic exploration but now widely adopted in high-resolution radar applications. SEG-Y provides a standardized structure that enables efficient handling, processing, and archiving large radar datasets.

A SEG-Y file typically contains:

- **Textual header (3200 bytes):** project information, date, client, notes
- **Binary header (400 bytes):** global parameters (sampling interval Δt , samples per trace, data format, antenna frequency)
- **Trace headers (240 bytes per trace):** position (x, y or distance along scanline), time-zero, gain, trigger type, GPS coordinates, line ID
- **Trace data:** raw A-scans as 32-bit float amplitudes. Each trace is a discrete time signal:

$$A(t) = [a_1, a_2, \dots, a_N]$$

where a_i is the recorded amplitude at sample i .

A single survey line produces a **B-scan**, which is a sequence of traces along a line as a 2D matrix $A(x, t)$. As you see on Figure 7, each trace represents a single time series data measurement at a specific location, typically acquired every 5–20 mm.

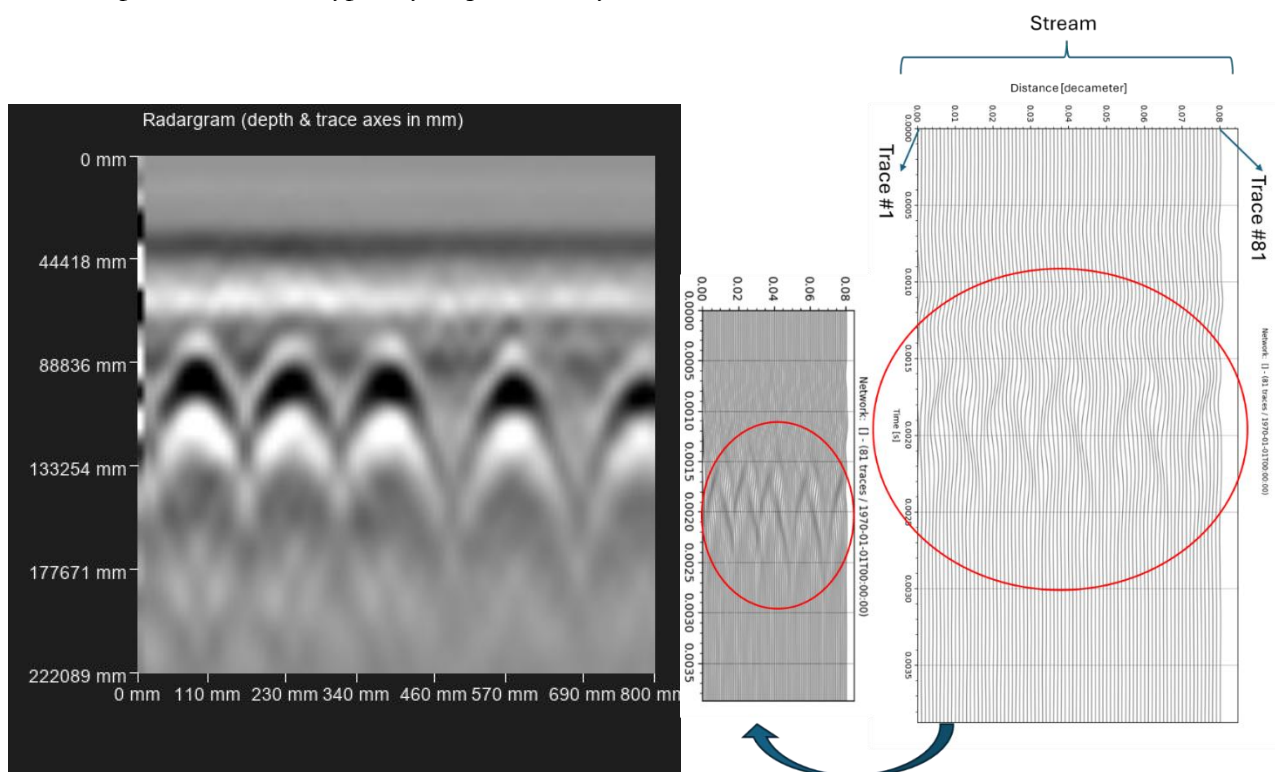


Figure 7 Example B-scan (radargram) formed by stacking consecutive A-scans. Hyperbolic signatures indicate point reflectors (rebars, ducts)

3.1.3 Essential signal processing steps

Time-Zero Correction: The recorded trace contains early-time system noise and antenna coupling effects. A time shift t_0 is applied so that the first break corresponds to actual wave emission:

$$A_{\text{corr}}(t) = A(t + t_0)$$

This ensures consistent depth interpretation across traces.

Filtering (Noise and Clutter Removal)

i. Bandpass Filtering

Radar systems often record noise outside the antenna's effective bandwidth. A frequency-domain bandpass filter is applied:

$$A_{\text{filtered}}(t) = \mathcal{F}^{-1}[H(f) \mathcal{F}[A_{\text{corr}}(t)]]$$

where:

$\mathcal{F}$ is the Fourier transform

$H(f)$ is a bandpass window centered on antenna frequency

ii. Background Subtraction

Removes horizontal banding by subtracting mean trace:

$$A_{\text{bg}}(x, t) = A(x, t) - \frac{1}{M} \sum_{i=1}^M A(i, t)$$

Migration (Hyperbola Collapse): Migration transforms blurred hyperbolas into points or coherent reflectors. Point reflectors appear as hyperbolas in B-scans. Migration focuses energy back to the true location. A common method is **Kirchhoff migration**, which applies:

$$A_{\text{mig}}(x, z) = \sum_i A(x_i, t = 2R_i/v)$$

where:

$R_i = \sqrt{(x - x_i)^2 + z^2}$ is the distance from trace i to point (x, z) ,
 v is EM wave velocity in concrete.

Figure 8 provides an example of GPR usage. The evaluation of results heavily relies on existing algorithms and user interfaces. The possibilities and limitations of GPR inspection technology are discussed in Table 10, outlining GPR's capabilities and constraints.

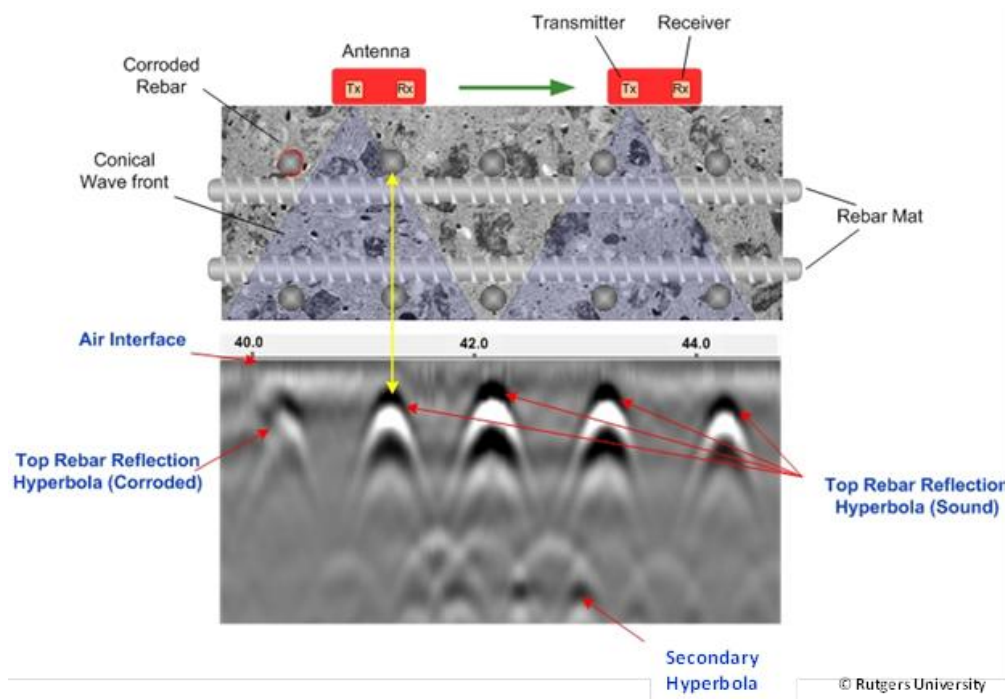


Figure 8 Ground-penetrating radar (GPR) (<https://fhwaapps.fhwa.dot.gov>)

3.2 Ultrasonic Pulse Echo (UPE)

3.2.1 Physical Principle

UPE is based on the propagation of elastic mechanical waves through concrete. When a transducer excites the concrete surface, it generates two primary wave types: longitudinal waves (P-waves) where particle motion occurs parallel to the propagation direction, and transverse waves (S-waves or shear waves) where particle motion is perpendicular to the propagation direction.

In concrete inspection, shear waves are strongly preferred due to their superior characteristics for defect detection. Shear waves experience lower attenuation in concrete compared to longitudinal waves, cannot propagate through air or fluids (making them highly sensitive to voids), and provide better sensitivity to material discontinuities. The shear wave velocity in concrete is governed by:

$$v_s = \sqrt{G/\rho}$$

where G is the shear modulus (elastic property) and ρ is the material density. Typical concrete exhibits shear wave velocities of $v_s = 2,000\text{--}2,800$ m/s, substantially slower than P-wave velocities of $3,500\text{--}4,500$ m/s.

The MIRA system, Figure 9-a, employs 48 dry point-contact transducers arranged in 12 rows of 4 transducers each, with approximately 25 mm spacing between rows. During operation, Figure 9-b, each row sequentially acts as a transmitter while the remaining rows serve as receivers, creating a dense array of ray paths through the concrete volume. This pitch-catch tomography configuration enables comprehensive sampling of the internal structure, Figure 9-c.

When shear waves encounter a concrete-air interface (such as a void), a strong reflection occurs due to the dramatic acoustic impedance mismatch. Acoustic impedance $Z = \rho \cdot v$ represents the material's resistance to wave propagation. Concrete has $Z \approx 6\text{--}8$ MRayl (Mega-Rayleigh), while air has $Z \approx 0.0004$ MRayl—a contrast of approximately 15,000:1. This enormous mismatch causes nearly complete reflection at void boundaries, making voids highly detectable through amplitude loss in transmitted signals.

The method provides significant penetration depth—up to 1 m in heavily reinforced concrete and 2 m in lightly reinforced elements, depending on concrete quality and reinforcement density.

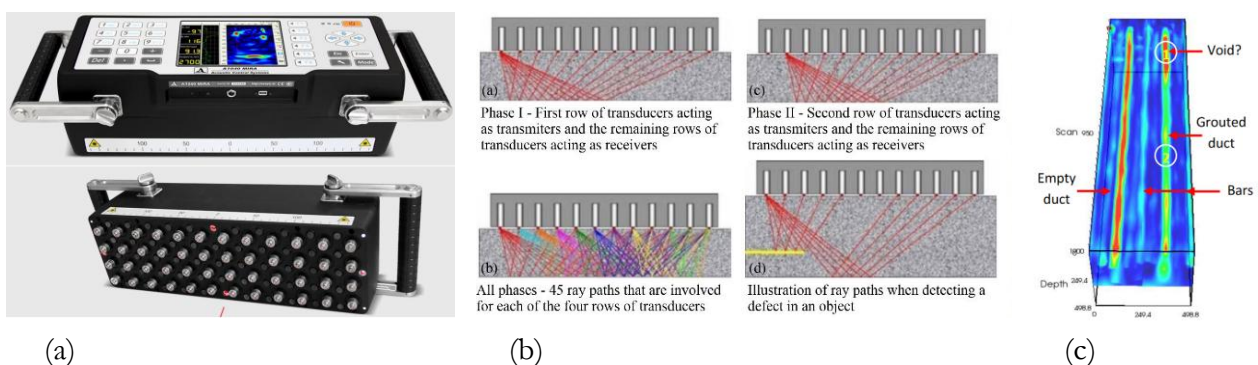


Figure 9 Shear wave tomography: (a) MIRA equipment; (b) working principle; and (c) MIRA 3D image grouted vs un-grouted tendon ducts (<https://www.germanninstruments.com>)

3.2.2 RAW data structure

UPE devices such as MIRA output a structured dataset that includes:

Raw Ultrasonic Traces (A-scans): Each transmit–receive pair generates a **time-domain signal**:

$$U(t) = [u_1, u_2, \dots, u_N]$$

where u_i is the measured wave amplitude at time sample i .

A single activation of a transmitter row produces multiple A-scans—one for each receiving transducer.

Data Arrangement: The full dataset can be represented as a 3D array:

$$U(x_r, x_s, t)$$

where:

x_s = transmitter position (row index)

x_r = receiver position (row index)

t = time samples

This dense matrix of traces forms the input to the tomographic reconstruction.

Each trace is accompanied by metadata that describes acquisition parameters, including:

- sampling interval Δt
- number of samples per trace N
- transducer array geometry (spacing, row number, layout)
- material coupling mode (dry point contact characteristics)
- excitation settings
- expected wave velocity range
- gain and filtering parameters applied by the device

Metadata is essential for the computational reconstruction, as the geometry determines the ray paths used in back-projection or SAFT.

3.2.3 Mathematical basis of UPE and SAFT processing

Ultrasonic tomography reconstruction relies on wave propagation physics. The recorded signal for each transmitter–receiver pair includes reflections, refractions, and scatter from internal boundaries.

Travel-Time and Amplitude Formation: The received waveform can be modeled as:

$$U(x_r, x_s, t) = \sum_i A_i s(t - \tau_i)$$

where:

A_i = amplitude of reflection from interface i ,

$s(t)$ = source wavelet,

$\tau_i = \frac{L_i}{v}$ = travel time,

L_i = path length from transmitter → interface → receiver,

v = shear-wave velocity in concrete.

Defects change the apparent travel time and reflection strength.

Time-Zero Correction: As in GPR, signals contain device-dependent time delays. A time shift t_0 is subtracted:

$$U_{\text{corr}}(t) = U(t + t_0)$$

ensuring correct depth estimates.

Filtering: Bandpass filters remove noise outside the expected shear-wave frequency band (typically 25–75 kHz):

$$U_{\text{filt}}(t) = \mathcal{F}^{-1}[H(f) \mathcal{F}[U_{\text{corr}}(t)]]$$

with $H(f)$ representing the bandpass window.

SAFT – Synthetic Aperture Focusing Technique (Core of MIRA Reconstruction)

SAFT reconstructs a focused 2D/3D image by summing energy along calculated wavefronts. For each image point (x, z) , energy from all transmitter–receiver pairs have integrated according to their travel paths:

$$I(x, z) = \sum_s \sum_r U_{\text{filt}} \left(t = \frac{d(x_s, x) + d(x_r, x)}{v} \right)$$

where:

- $I(x, z)$ = reconstructed intensity at point (x, z) ,
- $d(x_s, x)$ = distance from transmitter x_s to point x ,
- $d(x_r, x)$ = distance from point x to receiver x_r ,
- v = shear-wave velocity.

Interpretation:

- Defects (voids, delaminations) produce strong reflections → high $I(x, z)$.
- Homogeneous concrete produces weak or diffuse reflectivity.

Depth Conversion: Using the shear-wave velocity v_s , depth is computed as:

$$z = v_s t$$

(Shear waves travel one-way in tomography; reflection-based methods use $z = v_s t/2$).

The shear-wave velocity may be estimated through calibration blocks or picked first-arrival times.

3.3 Impact-Echo (IE)

3.3.1 Physical Principle

Impact-Echo operates on the principle of transient stress wave propagation and resonance in concrete elements. A mechanical impact—delivered by a spring-loaded impactor or instrumented hammer—generates three types of elastic waves that propagate through the concrete:

Longitudinal waves (P-waves): Fastest wave type with velocities $C_P = 3,500\text{--}4,500$ m/s in concrete, where particle motion is parallel to wave propagation direction

Shear waves (S-waves): Slower waves with velocities around 2,000–2,800 m/s, where particle motion is perpendicular to propagation direction

Rayleigh waves (surface waves): Propagate along the surface with velocities slightly less than S-waves, decaying rapidly with depth. The P-wave velocity in concrete is determined by:

$$C_P = \sqrt{[(K + 4G/3)/\rho]}$$

where K is the bulk modulus, G is the shear modulus, and ρ is density.

The key diagnostic principle underlying IE is stress wave resonance. When waves reflect from internal interfaces—such as the bottom surface of a slab, a delamination, or a void—they travel back to the surface and reflect again, creating standing wave patterns. For a solid plate of thickness H , the round-trip travel time is $T = 2H/C_P$, producing a resonance at the fundamental thickness frequency:

$$f_{\text{thickness}} = C_P/(2H)$$

Internal defects such as voids or delaminations create additional shallow reflectors that produce higher resonant frequencies, enabling depth estimation. The lower frequencies (2-20 kHz) used in IE provide good penetration for detecting large-scale anomalies while remaining sensitive to impedance contrasts at void boundaries.

3.3.2 RAW data structure

Impact-Echo data acquisition is simple in structure but requires accurate timing and high-resolution digitization.

Raw Time-Domain Signals (A-scans): Each test point produces a single **time-domain waveform**:

$$u(t) = [u_1, u_2, \dots, u_N]$$

where:

u_i = recorded particle displacement or velocity at time t_i ,
 sampling frequency = typically 40–200 kHz,
 record length = short (2–20 ms).

These waveforms contain reflections from internal boundaries and defects.

Data Structure (Arrays of Impact Points): For a survey with multiple impact locations, data may be organized as:

$$u(x, t)$$

where x indexes the test point on the surface grid.

More advanced systems organize this in 2D or 3D grids:

$$u(x, y, t)$$

allowing the creation of thickness maps and defect maps.

Metadata Stored with Each Trace: Metadata is essential for accurate frequency interpretation and velocity estimation. Metadata typically includes:

- sampling rate Δt
- sensor type (velocity/displacement)
- impactor characteristics (steel ball size, impact energy)
- calibration constants
- test location coordinates
- pre-trigger and post-trigger durations
- automatic gain settings
- environmental conditions

Unlike GPR and UPE, IE data are usually stored as single-trace files per point.

3.3.3 Mathematical basis of Impact–Echo processing

IE processing relies heavily on spectral analysis, boundary reflections, and the relationship between wave velocity and structural thickness.

Wave Propagation and Reflection: The stress waves generated by the impact reflect from boundaries or defects. For a plate of thickness H , the travel time for a round-trip reflection is:

$$T = \frac{2H}{C_p}$$

where C_p is the P-wave velocity in concrete (typically 3,500–4,500 m/s).

These reflections appear as periodic oscillations in the recorded signal.

Frequency-Domain Transformation: The recorded signal is transformed using the FFT:

$$U(f) = \mathcal{F}[u(t)]$$

The amplitude spectrum highlights resonant frequencies associated with internal reflections.

The **fundamental thickness frequency**, f_{IE} , is the dominant peak in the spectrum.

Thickness Calculation: The thickness of the element is estimated using:

$$H = \frac{C_p}{2f_{IE}}$$

If the detected frequency is **higher** than expected, it suggests a shallow reflector (e.g., delamination).

A **lower** frequency indicates a deeper boundary or greater member thickness.

4 Mock-Up specimens in NDT verification: Insights from past studies and basis for the present work

Testing NDT methods on real bridges presents a fundamental problem: you can never be completely sure what's actually inside the concrete unless you drill holes or cut sections open, which defeats the purpose of non-destructive testing. This is where mock-up specimens become invaluable—concrete blocks built in laboratories with known, documented internal conditions. By creating specimens where the exact locations of voids, defects, and embedded features are controlled and recorded during construction, researchers can test whether NDT instruments actually detect what they're supposed to detect.

This chapter reviews seven previous research projects from different institutions that built mock-up specimens to validate NDT methods for concrete inspection. Each project offers valuable lessons about what works and what doesn't. Some mock-ups were too simple and didn't realistically simulate field conditions. Others were so complex that signals from multiple defects interfered with each other, making results hard to interpret clearly. Some specimens deteriorated or changed over time in unexpected ways, compromising the controlled conditions they were meant to provide.

The chapter walks through these case studies chronologically, explaining the design choices each research team made, what problems they encountered, and what lessons emerged. These insights directly influenced how the mock-up for the current project was designed—learning from past mistakes to create a specimen that balances realism with clarity, provides reliable ground truth, and serves as an effective platform for validating the machine learning approach developed in this research.

4.1 Test 1: (Wiggenhauser et al., 2015)

One example is the mock-up campaign carried out by (Wiggenhauser et al., 2015) on steel-concrete modules intended for nuclear power plant applications (Figure 10). The aim was to evaluate acoustic-based NDT methods for detecting embedded voids and assessing the feasibility of inspecting such composite elements, where concrete is enclosed between steel liners.

Without going into detail about the specific test results, which fall outside the scope of the current project, it is important to highlight the key lesson: while the small mock-up fulfilled its role as a testing ground for defect detection in bare concrete, it did not fully deliver on its purpose of simulating the more challenging steel-concrete interface conditions encountered in real structures; the limitations of wave propagation through steel and the dominance of reflections from welds and edges significantly reduced measurement reliability.

From a broader perspective, the campaign demonstrated that mock-up design needs to be not only *realistic* but also *constructively honest* about the conditions it aims to simulate. Several recommendations were made: larger mock-ups to reduce surface interference; better control of defect types, sizes, and positions; inclusion of complex interfaces such as delaminated zones; and new ultrasonic systems specifically developed for testing through steel.



Figure 10 Fabrication of mock up specimens (from (Wiggenhauser et al., 2015))

4.2 Test 2: (Giannini et al., 2020)

Another mock-up campaign was led by (Giannini et al., 2020) focusing on alkali-silica reaction (ASR) damage in plain concrete. Unlike earlier studies limited to small-scale mortar bars, this effort aimed to create larger, standardized concrete slabs with known levels of ASR-induced deterioration. The purpose was not to validate a specific method, but to provide a shared platform for assessing the viability of different NDT techniques under conditions more representative of field structures.

Without going into test results, which are beyond the scope of this work, the main takeaway is that the mock-ups largely achieved their goal. Each slab was conditioned to a target expansion (0.05%, 0.10%, 0.20%) and stored in controlled environments, providing a solid basis for inter-laboratory comparison.

Nonetheless, the campaign exposed key challenges. The larger slabs continued expanding after shipment, while smaller prisms shrank due to different storage conditions, complicating efforts to maintain distinct damage states. Delays in testing access also meant some measurements were taken well after the intended conditioning phase.

The broader lesson is the difficulty of balancing realism and control in mock-up design. While the specimens successfully simulated various ASR stages, sustaining stable damage levels over time proved difficult. Future efforts should account for how deterioration may evolve both during and after conditioning, especially when aiming to simulate progressive degradation.

4.3 Test 3: (Dethof & Keßler, 2024)

A third example comes from a study authored by (Dethof & Keßler, 2024) which focused on the simulation-based design and ultrasonic testing of concrete mock-ups with embedded defects. The aim was to explore how numerical modelling can guide the design of diagnostically useful specimens, particularly for evaluating the detectability of honeycombs.

The mock-up included embedded reinforcement, steel pipes, and cast-in honeycombs. Finite element simulations were used beforehand to predict how ultrasonic waves would interact with these features, helping to optimize their placement and spacing. This approach minimized signal interference and enhanced interpretability during testing.

Importantly, the study also included ultrasonic inspections, with results presented and evaluated. A key finding was that even well-designed honeycombs could be difficult to detect in practice, depending on their geometry and surrounding material. This highlighted the need for mock-ups that not only simulate realistic flaws but also provide meaningful diagnostic challenges.

The main lesson is that combining simulation-driven design with physical testing improves both the quality and relevance of mock-ups. These insights were used in designing the present mock-up to ensure that the embedded features are both representative and easy to interpret during testing.

4.4 Test 4: (Oh et al., 2024)

In another campaign by (Oh et al., 2024) focused on evaluating the performance of different NDT methods in detecting internal delaminations in reinforced concrete. The specimen, representing a typical concrete bridge deck with an asphalt overlay. Delaminations were simulated by embedding inserts made of plastic film and mesh at varying depths — between the overlay and deck, above the top reinforcement, and below the bottom reinforcement — reflecting locations where such defects commonly occur in practice. The mock-up was designed with long-term use in mind, stored outdoors to allow natural exposure to freeze–thaw cycles. This added a layer of surface deterioration, providing a semi-realistic context for testing. Importantly, the design allowed for controlled and repeatable delamination simulation, offering a platform for method comparison.

However, the approach also highlighted some inherent trade-offs. While the artificial inserts allowed for precise control over location and geometry, they remained idealized and may not fully capture the material complexity or irregularity of naturally developed delaminations. Additionally, the use of plastic materials to simulate air gaps simplifies the condition but does not replicate mechanical debonding or moisture influence present in actual field cases. The key takeaway from this study is the value of deliberately simplified yet well-documented specimens for benchmarking.

4.5 Test 5: (Krause et al., 2008)

Another significant mock-up campaign, reported by (Krause et al., 2008) aimed to evaluate the ability of ultrasonic methods to detect voids in grouted post-tensioning ducts embedded in concrete. The specimen included metallic ducts with known internal conditions: fully grouted, partially grouted, and ungrouted segments (Figure 11). These ducts were embedded in a reinforced concrete block to realistically reflect the geometry and material constraints of in-situ bridge components.

Several practical and conceptual lessons emerged from the mock-up effort. First, the study demonstrated the critical importance of embedding ducts within concrete, rather than testing them in isolation. The steel-concrete interface had a clear influence on wave propagation and signal interpretation, and without this structural context, test results would have lacked field relevance.

Second, the campaign highlighted the benefit of having well-documented internal ground truth, allowing the researchers to evaluate method sensitivity and reliability under controlled conditions. However, it also showed that the simplified void geometries used in the mock-up — often clean and uniform air gaps — did not fully capture the irregular shapes, moisture presence, or partial bonding often observed in real voids. This limited the ability to generalize the detection thresholds to actual structures.



Figure 11 Left: Tendons being installed in the Large Concrete Slab (LCS) prior to grouting. Right: Close-up of a tendon duct containing prestressing strands with an intentionally created artificial void, prepared before placement in the slab

4.6 Test 6: (Kessler & and Grosse, 2023)

This study, presented by (Kessler & and Grosse, 2023) focused on integrating a mock-up into civil engineering education to provide students with hands-on experience in NDT. Unlike mock-ups designed primarily for research or method benchmarking, this initiative prioritized training, accessibility, and realism in an academic setting. A reinforced concrete slab (Figure 12) was constructed with deliberately embedded features – including plastic and steel elements – to mimic internal heterogeneities and allow testing with methods such as UPE, GPR, and IE.

The mock-up served as a versatile educational tool. It was designed for repeated use across multiple training sessions, and its layout was well-documented to enable effective interpretation and comparison of results. One of the key outcomes was the recognition that physical mock-ups enhance conceptual understanding of wave propagation, reflection, and practical testing constraints far more effectively than simulations or theoretical exercises alone.

The study underscored that even relatively simple specimens can offer substantial educational value, especially when their geometry and embedded features are thoughtfully planned to challenge interpretation without overwhelming students. Although the specimen was not built to represent deterioration in field structures, it succeeded in meeting its goal of teaching inspection techniques and reinforcing key diagnostic concepts in a controlled, repeatable environment.



Figure 12 Photos of the mock-up before concreting

4.7 Test 7: (Popescu & Täljsten, 2020)

(Popescu & Täljsten, 2020) developed a mock-up specimen to evaluate NDT techniques for detecting internal defects in tendon ducts, including grouting voids, water ingress, and corrosion indicators. The reinforced concrete specimen included a combination of steel and plastic ducts, tendons, and a series of engineered defects, such as water-filled bags, rubber inserts, ungrouted sections, and corrosion zones. The aim was to reflect the wide range of conditions encountered in real post-tensioned structures, offering a high level of realism and material diversity.

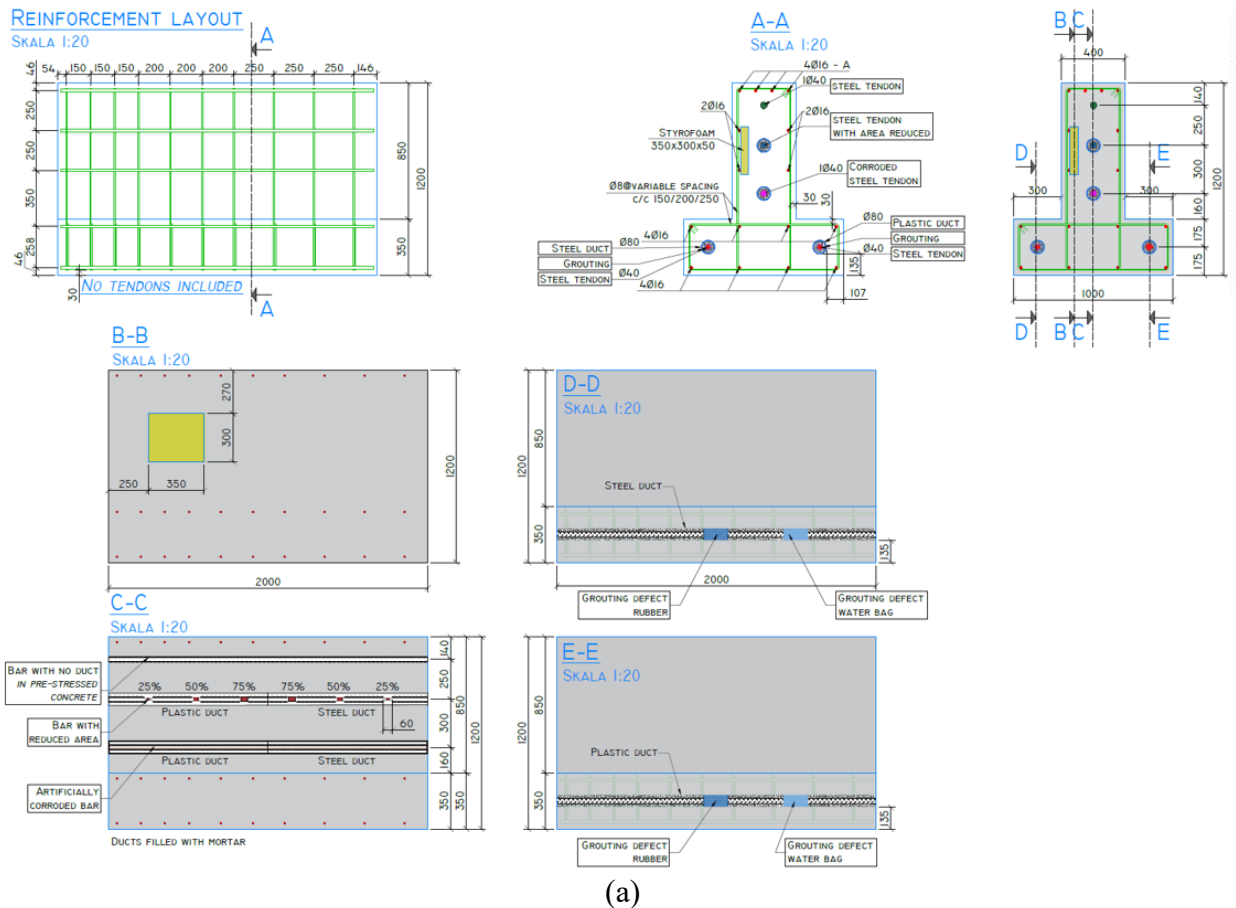


Figure 13 Design (a) and fabrication process: (b–c) Water bags and rubber inserts simulating water ingress and grouting defects inside tendon ducts; (d) grouting of the ducts; (e) placement of all ducts, reinforcement, and artificial defects before concreting; (f) completed mock-up specimen

While the mock-up succeeded in simulating realistic conditions, several important lessons emerged. One of the key takeaways was that increased complexity does not always lead to better outcomes – particularly when the mock-up is used for training personnel or evaluating specific NDT equipment. When too many defect types are combined in close proximity, interactions between defects can distort signals and make it difficult to isolate the performance of individual methods or interpret results clearly. In effect, the setup can become more complicated than what would typically be found in situ on a real bridge, undermining the clarity needed for controlled testing or instructional purposes.

This experience highlighted that while realistic features are essential for field-relevant validation, mock-up design should be guided by a clear purpose – whether it is to simulate field conditions, evaluate equipment sensitivity, or train operators. Each goal may demand a different balance between realism, simplicity, and control.

In planning the present mock-up specimen, these lessons were taken into careful consideration. The design phase emphasized clarity and traceability, with detailed documentation of the reinforcement layout, embedded duct configurations, and intended defects. Particular attention was given to ensure that known errors in earlier campaigns — such as overlapping reinforcement causing signal confusion or lack of ground truth — would be avoided. The goal was to create a specimen that is not only representative of field conditions but also reliable as a testbed for method verification and comparison.

5 Methodology

This chapter explains how the research was actually conducted, walking through the complete process from designing concrete specimens and fabrication in the lab, to testing the NDT devices on prepared specimens. The approach follows a logical progression: first, build carefully controlled mock-up specimens where everything inside the concrete is known and documented; second, use these specimens to collect high-quality NDT measurements and document the collected data to train machine learning models where the correct answers are already known.

Four different concrete specimens were built, each serving a specific purpose. Specimen 1 is a simple reference slab with just reinforcement bars and no defects—essentially a baseline to understand what normal concrete looks like to NDT instruments. Specimen 2 introduces post-tensioned ducts, some filled with grout and others left empty, testing whether voids can be detected. Specimen 3 increases complexity with ducts at two different depths and alternating void patterns, challenging the detection limits. Specimen 4 focuses on different types of defects—delamination and honeycombing—that aren't related to tendon ducts but commonly occur in deteriorated concrete.

The chapter describes exactly how these specimens were designed, fabricated, and tested, providing the foundation for all the machine learning analysis that follows in later sections.

5.1 General

The project follows a complete development chain, beginning with the fabrication of controlled mock-up specimens where representative defects—such as partially grouted ducts, voids, and delaminations—are intentionally introduced. These specimens enable systematic NDT data collection using GPR, UPE, and IE under well-defined conditions, forming a high-quality dataset for training and validating machine-learning models. The trained models are then integrated into a functional prototype and assessed through field demonstrations on real bridge structures. This stepwise process ensures that the methods are not only scientifically verified but also proven under practical inspection conditions, enabling a robust transition from laboratory validation to operational field implementation.

5.2 Mock-Up Specimen Design

To assess the applicability of NDT methods in detecting voids and anomalies typically encountered in post-tensioned concrete structures, four custom-designed concrete mock-up specimens were fabricated. Each specimen was tailored to reflect specific defect types under controlled conditions, allowing for repeatable validation of inspection techniques.

5.2.1 Specimen 1

The first mock-up specimen was designed as a baseline concrete element to support the calibration and verification of NDT methods aimed at detecting internal reinforcement and evaluating material properties.

The specimen measures $1600 \times 1200 \times 300$ mm and is reinforced with Ø16 mm bars placed in both the top and bottom layers. The reinforcement spacing is 100 mm and 200 mm centre-to-centre, arranged in orthogonal directions. The clear cover to the rebar is approximately 35 mm. No ducts or embedded defects were introduced in this specimen.

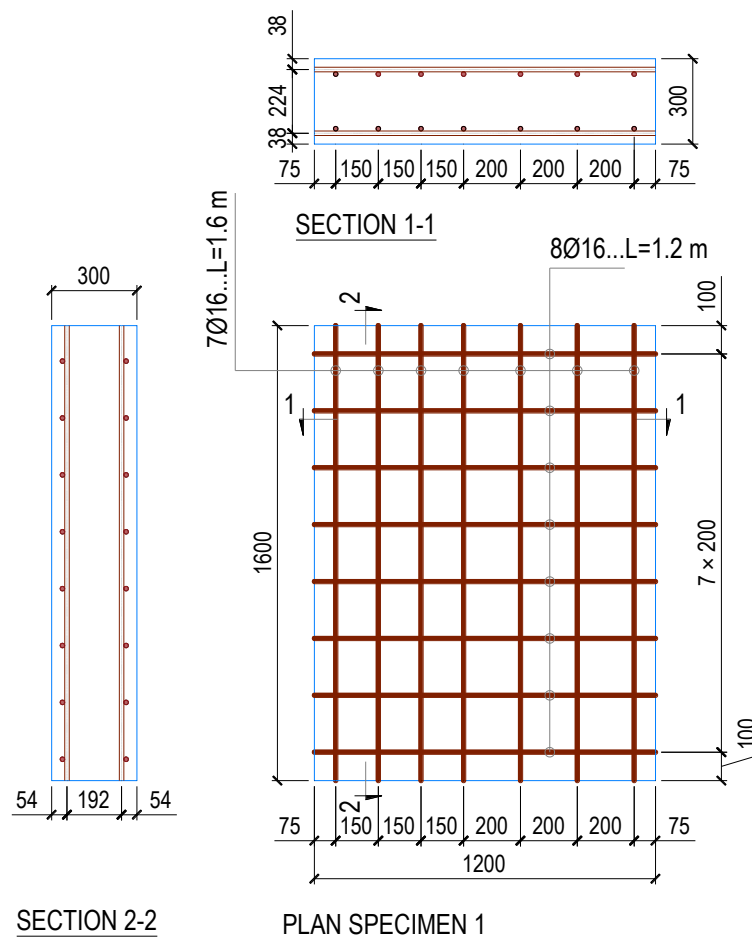


Figure 14 Design of specimen 1

The concrete used corresponds to C30/37, with specified values for compressive strength f_c and modulus of elasticity E_c . Material samples were taken during casting to support destructive testing for cross-validation.

This specimen served two primary purposes:

1. **Verification of material properties** using non-destructive and destructive methods.
2. **Assessment of NDT performance** in locating embedded reinforcement, identifying bar diameters, and confirming layer positions.

In addition, the slab was used to evaluate the applicability of basic NDT instruments, such as the combined Schmidt hammer and Ultrasonic Pulse Echo (UPE) method, for estimating in-situ concrete compressive strength.

The absence of ducts and defects made this specimen ideal as a reference slab, allowing evaluation of baseline measurement accuracy without interference from complex features. It also played a key role in distinguishing signal responses from intact rebar and concrete prior to introducing artificial defects in the remaining specimens.

5.2.2 Specimen 2

Specimen 2 was designed to introduce controlled complexity by simulating post-tensioned ducts with intentional voids. It shares the same outer geometry and reinforcement layout as Specimen 1—measuring $1600 \times 1200 \times 300$ mm with $\varnothing 16$ mm bars spaced at 100 mm and 200 mm on the top and bottom layers.

In this specimen, two vertical corrugated steel ducts were embedded: both half-filled with grout and other half left intentionally ungrouted. The aim was to evaluate the detectability and differentiation of grouted versus ungrouted tendon ducts using NDT. The ducts were placed at a 150 mm depth from the top surface and symmetrically aligned across the width.

The positioning of the ducts under zones of both dense and sparse reinforcement allowed for testing how overlying rebar density affects signal quality and interpretation. As with Specimen 1, the concrete was of class C30/37. Tendons were inserted to mimic real-world bridge post-tensioning scenarios.

This configuration supports training and method validation focused on locating ducts, estimating grouting extent, and understanding how tendon visibility is influenced by rebar congestion. However, unlike more intricate configurations, the relatively simple layout ensures that each feature can be clearly linked to the corresponding response in the NDT data—supporting both clarity in results and transferability to field conditions.

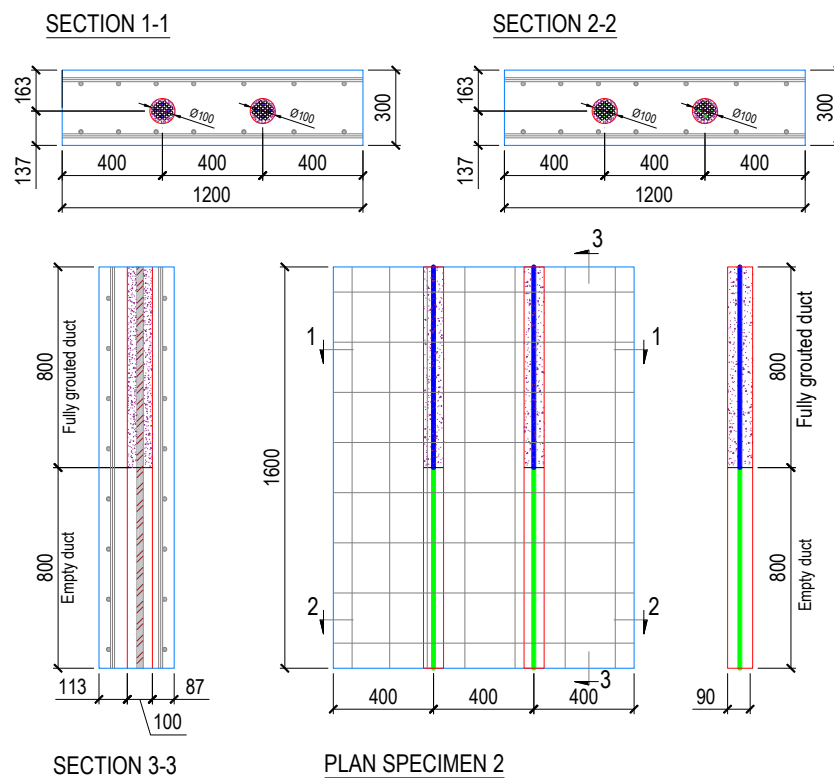


Figure 15 Design of specimen 2

5.2.3 Specimen 3

Specimen 3 is a $1600 \times 1200 \times 600$ mm concrete slab, sharing the same reinforcement layout as Specimen 1, with top and bottom layers of $\varnothing 16$ mm bars at 100/200 mm spacing. The specimen includes four vertical corrugated steel ducts embedded in two layers:

1. **Top layer:** Two ducts, each half fully grouted and half left empty (void).
2. **Bottom layer:** Two ducts, also alternating between grouted and empty segments, with the inverse pattern of the top layer to simulate alternating defect depth.

The vertical separation between layers is 170 mm, with the top ducts positioned 200 mm from the top surface and the bottom ones 200 mm from the bottom. Steel tendons were inserted in all ducts, consistent with post-tensioned bridge practice. The primary aim of this mock-up was to challenge NDT equipment and operators to:

- Locate tendons in both shallow and deep positions.
- Distinguish between fully grouted and voided segments within the same duct.
- Assess the influence of tendon depth and grouting pattern on NDT results.
- Investigate signal interference due to alternating voids across depth layers—an artificial but instructive complexity.

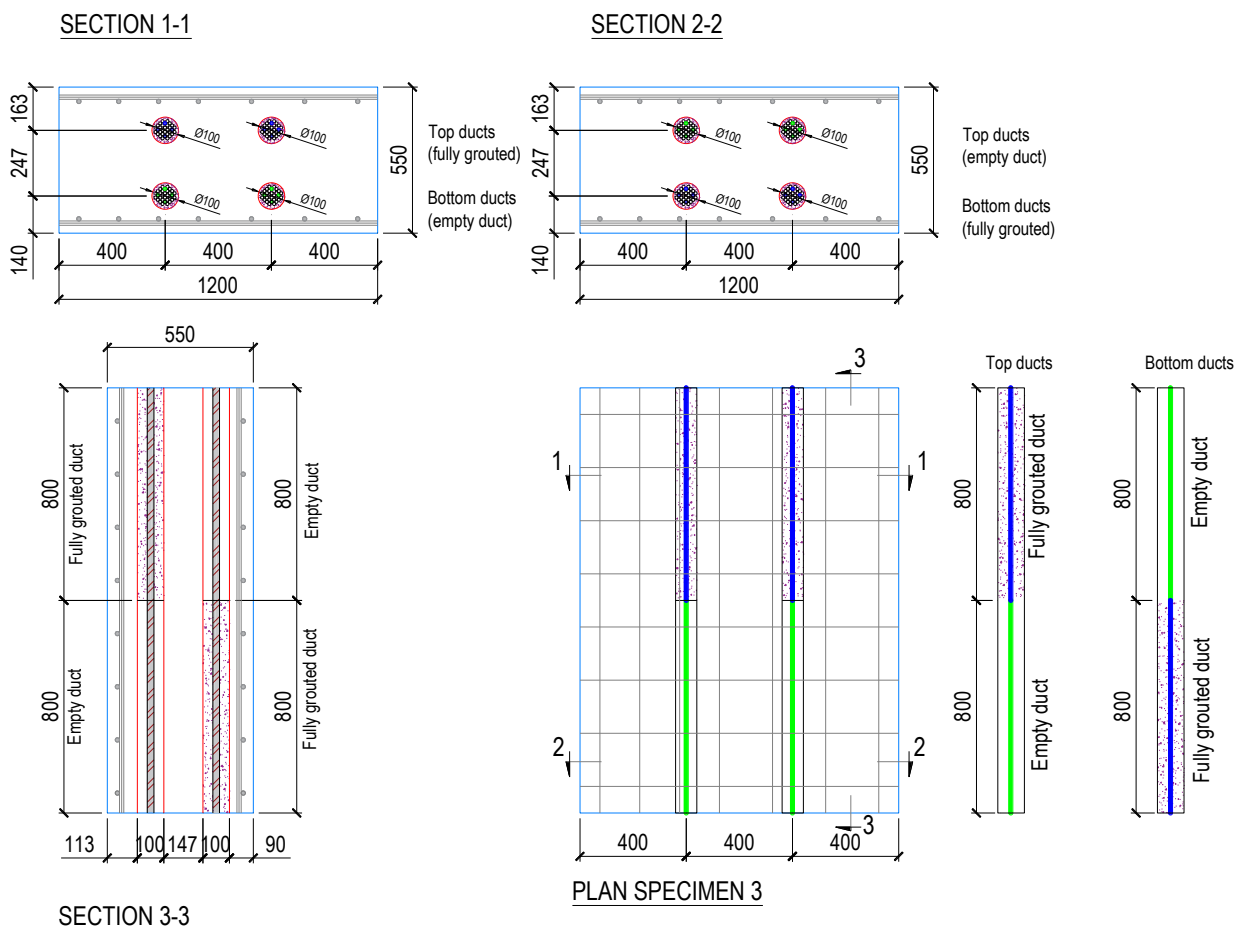


Figure 16 Design of specimen 3

5.2.4 Specimen 4

Specimen 4 was designed to simulate two common, but distinctly different defect types often encountered in reinforced concrete structures: delamination and honeycombing. The specimen measures $1600 \times 1200 \times 600$ mm and features the same reinforcement layout as Specimen 1, ensuring consistency in surface steel configuration across all mock-ups. No ducts or tendons were included, as the focus here was on material-related defects rather than tendon integrity.

To represent a delamination, a thin plastic foil (approximately 2–3 mm thick) was embedded horizontally below the top reinforcement layer, simulating a planar separation or poor bond between layers—something that may occur due to cold joints or improper casting. In addition to this, an irregular region of intentionally poorly compacted concrete was introduced near the bottom surface to mimic honeycombing, representing large internal voids caused by insufficient vibration or segregation during casting.

The purpose of this specimen was to evaluate how well NDT techniques can detect and differentiate between delaminations and volumetric defects such as honeycombs, particularly when they occur at different depths. It also served to challenge both the capabilities of the inspection equipment and the interpretative skills of operators.

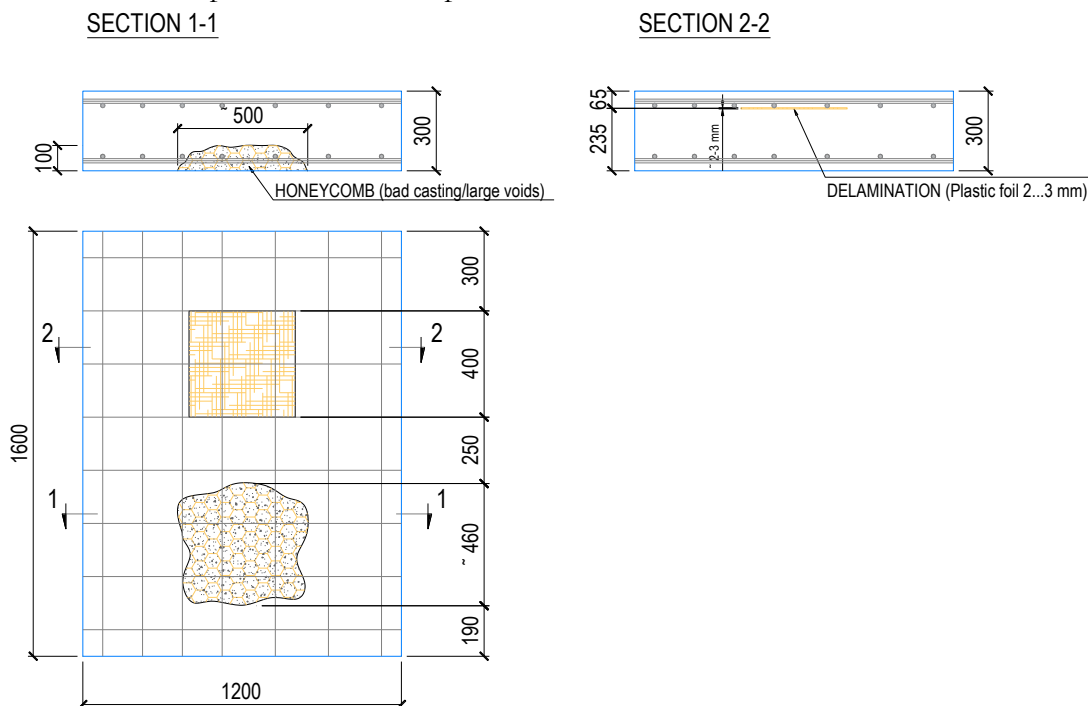


Figure 17 Design of specimen 4

5.3 Fabrication process of mock-up specimens

Before casting the final mock-up specimens, a prototype element was first constructed to test and refine the fabrication method. The aim of this prototype was to verify the feasibility of placing all reinforcement and artificial defects precisely according to the design drawings.

Given that each rebar had to be positioned with high accuracy to match the documented layout, the side edges of the formwork were specially modified. Custom protrusions were introduced along the inner face of the formwork walls to lock each bar into its correct position, thereby ensuring the intended concrete cover was achieved throughout the specimens.

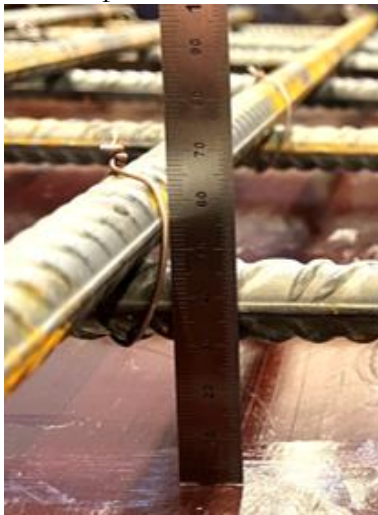
The prototype phase also served to test several critical aspects of the process:

- The grouting procedure for tendon ducts, especially to simulate both fully filled and empty conditions.
- The installation of artificial defects, such as delamination foils and honeycombing areas.
- The handling sequence for integrating tendons, ducts, and reinforcement within the formwork.

Following successful validation, small adjustments were made based on the prototype outcomes, after which the fabrication of all four specimens commenced.

The production process included the following steps:

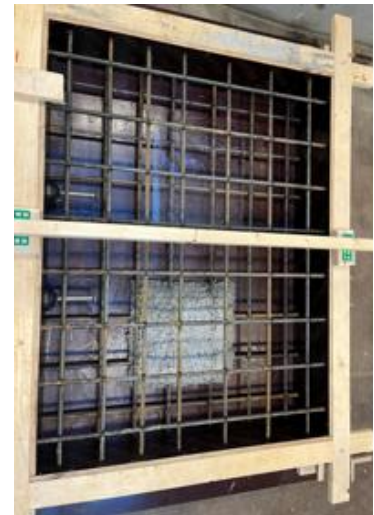
1. **Pre-fabrication of artificial defects** separately from the specimens.
2. **Grouting of ducts** (where applicable) before placing them into the formwork.
3. **Installation of surface reinforcement and tendon ducts** into the prepared formwork, using alignment guides and fixed spacers to maintain correct positioning.
4. **Concrete casting**, carried out after a final inspection of all embedded components to confirm compliance with the drawings and defect layout.



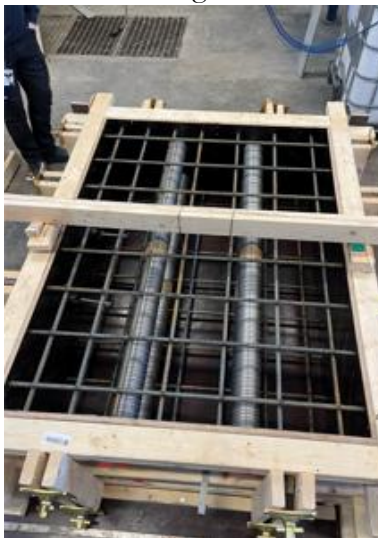
Rebar alignment



Duct preparation



Artificial defects



Duct placement in
formwork



Casting of specimens



Completed specimen

Figure 18 Fabrication process

5.4 Testing objectives

The primary objective of the testing program was to systematically evaluate the capabilities and limitations of multiple NDT techniques for detecting various defect types in concrete structures under controlled conditions. By testing specimens with known, documented internal features—ranging from baseline reinforcement to complex grouting voids and material defects—the study aimed to establish ground truth datasets suitable for training and validating machine learning models.

5.4.1 Overall Testing Strategy

The testing program employed a hierarchical approach, progressing from simple baseline measurements to increasingly complex defect scenarios:

1. **Baseline characterization** using Specimen 1 (reference slab with no defects)
2. **Single-layer void detection** using Specimen 2 (ducts at uniform depth)
3. **Multi-layer complexity** using Specimen 3 (ducts at two depths with alternating voids)
4. **Alternative defect types** using Specimen 4 (delamination and honeycombing)

This progression allowed systematic isolation of variables affecting NDT performance, including defect type, depth, size, and the influence of overlying reinforcement.

NDT Methods Employed

Each specimen was tested using a combination of complementary NDT techniques, selected to provide both rapid screening and detailed characterization:

Basic material characterization:

- **Schmidt hammer** combined with **ultrasonic pulse echo (UPE)** for estimating in-situ concrete compressive strength
- **Dry-coupled shear-wave transducers** for measuring dynamic elastic modulus

Structural element detection:

- **Cover meter** for rapid rebar depth verification
- **Ground-penetrating radar (GPR)** using two antenna frequencies:
 - 1.6 GHz high-frequency antenna for shallow, high-resolution imaging
 - 400 MHz low-frequency antenna for deeper penetration

Defect detection and characterization:

- **Ultrasonic pulse echo tomography (MIRA)** for high-resolution internal imaging and void detection through multi-channel coherence analysis
- **Impact Echo (IE)** for resonance-based thickness evaluation and localized void detection

Specimen-Specific Testing Objectives:

Table 5-1 summarizes the specific purpose and testing focus for each specimen.

Table 5-1 Test specimens and their evaluation objectives

Specimen	Dimensions (mm)	Embedded Features	Purpose
1	1600×1200×300	Ø16 rebar mesh only (no defects)	Testing material properties and steel mesh not affected by tendon ducts
2	1600×1200×300	Two Ø100mm vertical ducts at 150mm depth: • Half fully grouted and half empty	Investigating grouting defect's
3	1600×1200×600	Four Ø100mm ducts in two layers: • Top layer (200mm depth): alternating half-filled/half-empty • Bottom layer (370mm depth): inverse pattern	Investigating grouting defects and effects of shadowing
4	1600×1200×600	• Horizontal plastic foil (2-3mm) simulating delamination • Intentionally poorly compacted region (honeycombing)	investigating of void defects inside the concrete structure

5.4.2 Data Collection Protocols

For each specimen and NDT method, systematic data collection protocols were established to ensure repeatability and comprehensive coverage:

GPR measurements:

- Parallel scan lines with 5-10 cm spacing covering entire specimen surface
- Proceq GPR with SFCW (Stepped-Frequency Continuous Wave) technology operating across 0.2-4.0 GHz frequency range used on same grid
- Raw A-scan data saved in SEG-Y format with spatial metadata

MIRA ultrasonic tomography:

- 10 cm measurement spacing along identified duct centerlines
- Additional area scans (20 cm grid) for comprehensive defect mapping
- Raw multi-channel I/Q waveforms saved in proprietary .sin/.cos format
- Multiple scans at critical locations to assess repeatability

Impact Echo:

- Grid-based measurements with 10-15 cm point spacing
- Minimum 3 impacts per location to verify consistency
- Raw time-domain waveforms saved for spectral analysis
- Supplementary measurements at defect boundaries for spatial resolution assessment

The systematic, controlled nature of the mock-up testing phase established the foundation for all subsequent machine learning development, providing the labelled datasets and performance benchmarks necessary to transition from research prototype to field-deployable inspection tool.

6 Current NDT Inspection Practice

Before explaining the machine learning models themselves, it's important to understand the problem they're designed to solve. This chapter starts by walking through how bridge inspections currently work in practice—not the ideal theoretical process, but the actual day-to-day reality that inspectors face. The typical workflow involves setting up NDT equipment, collecting measurements, letting commercial software automatically process the raw signals into images and graphs, and then having human inspectors visually examine these processed outputs to identify potential problems. This approach works reasonably well for obvious, large defects, but it has serious limitations.

The chapter systematically identifies these limitations: different inspectors interpret the same data differently based on their experience and fatigue levels, commercial software from different manufacturers processes identical raw signals into noticeably different visualizations, and the preprocessing steps that create pretty pictures for human eyes often discard subtle information that could reveal small or ambiguous defects. This sets up the motivation for the machine learning approach: instead of having humans look at processed images, train algorithms to analyse the complete raw measurement data directly, preserving all the physical information. The chapter explains why this matters, what advantages it offers, and critically, how machine learning should work alongside human inspectors rather than replacing them—providing a quantitative "second opinion" that helps prioritize where experienced engineers should focus their attention.

In contemporary engineering practice, NDT in concrete structures still relies heavily on the expertise and subjective interpretation of human inspectors. Although modern instruments such as GPR, ultrasonic tomography systems (UPE), and IE generate large volumes of high-resolution raw data, inspectors rarely interact with these raw signals. Instead, their evaluations are based on pre-processed, highly structured visual outputs produced by each instrument's proprietary software.

The methodological framework developed in this project seeks to replicate, formalize, and ultimately automate the decision-making process traditionally performed by inspectors during on-site condition assessments. The complete workflow—summarized in Figure 19 (Methodological Workflow)—illustrates all operational steps from field acquisition to final reporting.

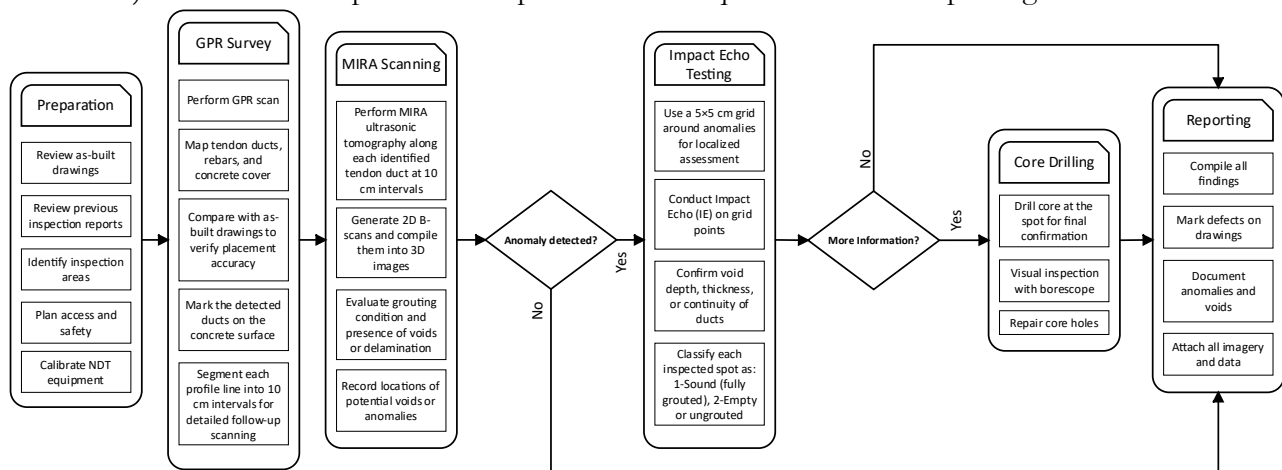


Figure 19 Methodological workflow

During on-site inspections, evaluators work exclusively with processed visualizations rather than raw measurement data. For GPR, inspectors interpret radargrams (B-scans and C-scans) automatically generated by the instrument's software. For ultrasonic testing, they examine tomographic images reconstructed by MIRA's proprietary SAFT algorithms. For Impact Echo, they analyse frequency spectra and dominant-frequency charts produced through automated FFT processing.

Each of these visualizations represents the endpoint of extensive signal processing. GPR radargrams undergo filtering, time-zero correction, windowing, and migration. MIRA tomograms result from SAFT-based reconstruction applied to multi-channel waveforms. IE charts emerge from FFT-based peak extraction applied to time-domain responses. While this processing creates images suitable for human visual interpretation, it introduces three significant limitations:

First, inspectors access only a fraction of the information present in the original raw waveforms. Much of the phase relationships, amplitude details, and high-frequency content gets discarded during visualization-oriented processing. **Second**, while preprocessing effectively reduces measurement noise, it may inadvertently eliminate subtle signatures characteristic of small or geometrically complex defects. **Third**, interpretation becomes strongly influenced by the specific visualization styles and algorithmic assumptions embedded in each manufacturer's commercial software—different software processing the same raw data can produce noticeably different images, leading to inconsistent interpretations.

6.1 Human inspector work flow

The conventional NDT workflow for concrete structures follows a well-defined sequence. While effective for detecting major defects, it is fundamentally limited by its dependence on human visual interpretation of processed outputs. Understanding this workflow is essential for identifying where ML can improve objectivity and repeatability.

A typical on-site workflow for an NDT assessment consists of five stages:

1. Instrument setup and calibration
2. Field data acquisition
3. Automatic data processing by commercial software
4. Visual interpretation and expert judgment
5. Decision-making and reporting

Each stage introduces uncertainty and operator-dependent variability that influence the final assessment.

6.1.1 Phase 1: Instrument Setup and Calibration

Before data collection, inspectors configure each instrument according to expected subsurface conditions, structural geometry, and inspection objectives. Although manufacturer guidelines exist, parameter selection is largely experienced-driven and varies between operators.

GPR configuration

For SFCW (Stepped-Frequency Continuous Wave) GPR systems like the Proceq unit used in this study, the system automatically sweeps across the full frequency range (0.2–4.0 GHz), capturing both deep penetration and high-resolution information simultaneously. Traditional impulse GPR systems (e.g., GSSI, Hilti) require the operator to select discrete antenna frequencies, trading off between:

- **400–900 MHz:** penetration up to ~1.5 m with moderate resolution
- **1.6–2.6 GHz:** millimetre-scale resolution with shallow penetration (0.3–0.5 m)

Additional configuration parameters include:

- **Trace spacing:** 1–5 cm, determining horizontal sampling density
- **Time window:** 10–50 ns, relating to maximum depth of investigation
- **Gain settings:** manual or automatic gain control (AGC)
- **Trigger mode:** distance-based or time-based acquisition

These settings significantly shape the resulting radargrams, yet are often selected without evaluating raw waveforms.

MIRA Ultrasonic Tomography Configuration

The MIRA device uses 12 pairs of dry-point-contact shear-wave transducers arranged in a linear array. Reliable coupling with the concrete surface is essential. The inspector must:

- Ensure consistent coupling pressure
- Verify surface conditions (clean, smooth, free of debris)
- Select the appropriate measurement mode (through-transmission or pulse-echo)
- Configure key sampling parameters, such as:
 - Sampling rate (~ 0.5 MHz)
 - Block pitch (25 mm spacing)
 - Imaging depth (0–500 mm)
 - Shear-wave velocity (commonly ~ 2300 m/s)

Impact Echo Configuration

IE is a point-based technique requiring careful positioning at each grid location. The inspector:

- Place and couple the accelerometer correctly
- Deliver repeatable impacts
- Set acquisition parameters (sampling rate ~ 50 kHz, record length)
- Calibrate P-wave velocity at a reference point

Because IE produces only discrete measurements, it is time-intensive but highly localized.

At this early stage, operator choices often differ—introducing variability even before data collection begins.

6.1.2 Phase 2: Field Data Acquisition

After configuration, instruments are used to collect spatially distributed measurements. Acquisition procedures differ significantly among GPR, MIRA, and IE.

GPR data acquisition

GPR surveys involve moving the antenna along predefined scan lines using a manual cart or robotic platform. For area surveys, the inspector establishes a regular grid (typically 5–10 cm spacing), producing a set of parallel B-scans, see Figure 20.

Each B-scan consists of hundreds to thousands of **A-scan traces**, see Figure 7, each containing a time-domain waveform. Raw GPR data—usually stored in SEG-Y format—contain:

- **Trace headers:** position, gain, acquisition time
- **Waveform samples:** typically, 512–1024 samples per trace
- **Global file metadata:** antenna frequency, dielectric assumptions, trace spacing



Figure 20 defined gridlines on the concrete surface and scanning along the scan lines

MIRA Ultrasonic Acquisition

MIRA operates in static mode. At each position (Figure 21), the system sequentially fires all transmitters, producing 66 emitter–receiver pairs.

Raw data consists of:

- **I/Q channels:** 12 channels, 7776 samples each
- **Acquisition metadata:** sampling rate, block pitch, assumed velocity
- **File format:** proprietary .sin (in-phase) and .cos (quadrature)



Figure 21 Ultrasound scanning, using MIRA A1040

For extended elements (e.g., tendon ducts), the array is moved in ~ 10 cm increments, creating overlapping tomograms.

Impact Echo measurement protocol

Impact Echo data collection is inherently discrete and labour-intensive. At each grid point, the inspector:

- Coupling a piezoelectric accelerometer to the surface
- Striking with a standard impactor
- Recording the acceleration waveform (10–20 ms)

This process is repeated for each grid cell (e.g., A1, A2, B1, B2, ...), generating a dataset of labeled single-point measurements.

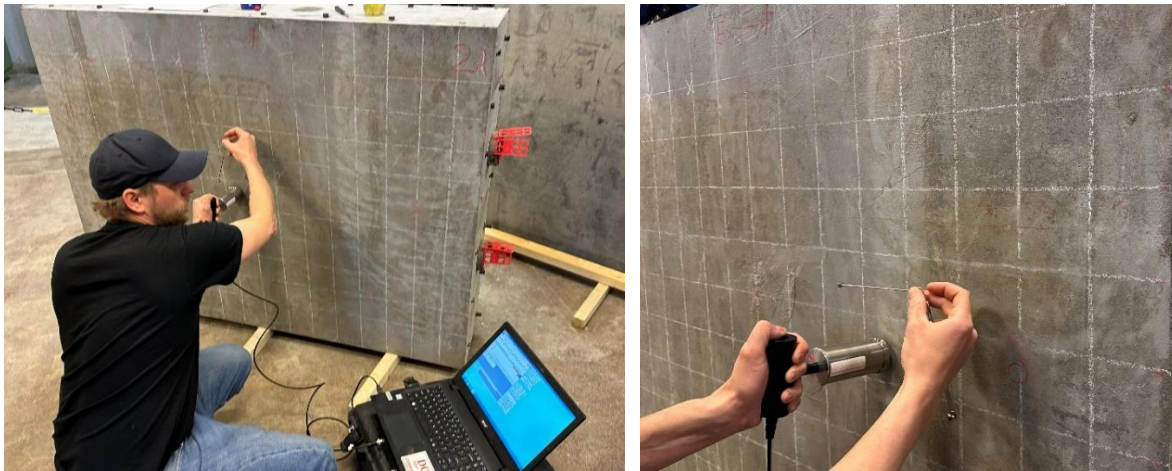


Figure 22 Impact echo testing on grid positions

Each data file contains:

- Time-domain signal (512–2048 samples)
- Grid position label
- Metadata (sampling rate, sensor sensitivity)

6.1.3 Phase 3: Automatic processing by commercial software

Once acquired, each system's software applies a sequence of preprocessing steps that transform raw physical measurements into interpretable visualizations.

GPR radargram generation

Automatic processing typically includes:

- Time-zero correction
- Dewow filtering (low-frequency drift removal)
- Background subtraction
- Bandpass filtering
- Time-varying gain (AGC)

The final radargram is a 2D image in which depth (or two-way travel time) is mapped vertically and position horizontally (Figure 23). Inspectors rarely go through the raw data and just rely on generated radargram to detect embedded anomalies.

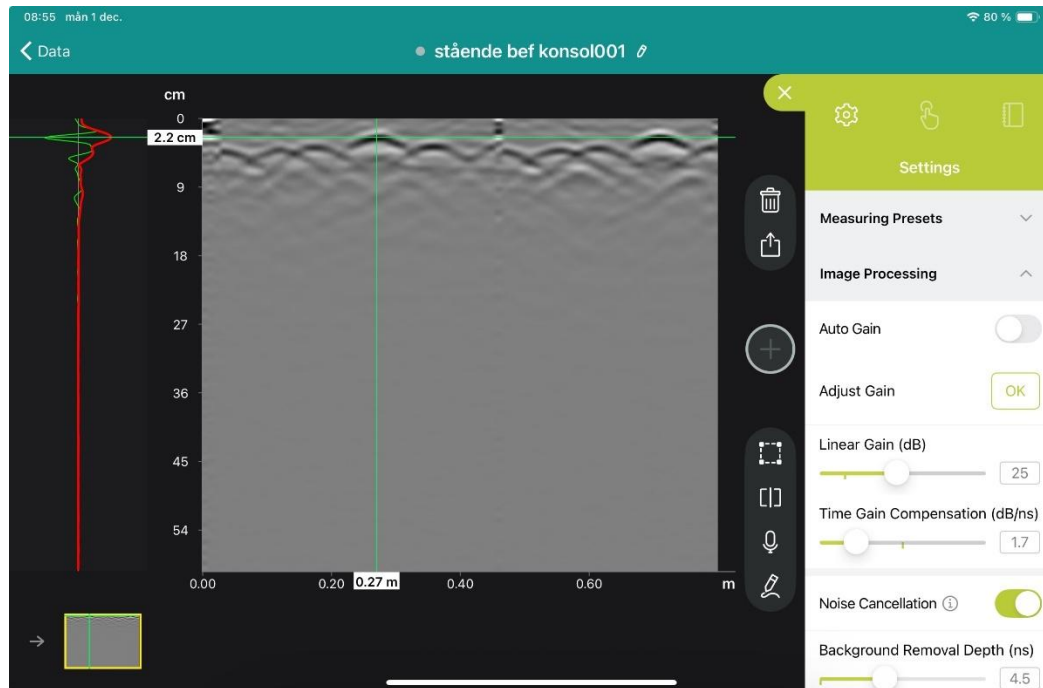


Figure 23 Created radargram on commercial software interface

MIRA tomographic reconstruction

Raw MIRA I/Q data undergoes **Synthetic Aperture Focusing Technique (SAFT)** reconstruction:

- **Channel-wise delay-and-sum beamforming:** For each pixel in the output tomogram, the algorithm calculates expected time-of-flight from each transmitter-receiver pair, delays the corresponding raw signals, and coherently sums them
- **Velocity-based focusing:** Uses the predefined shear wave velocity (e.g., 2300 m/s) to map time delays to spatial positions
- **Amplitude mapping:** The focused energy at each pixel is color-coded (typically red = high amplitude, blue = low amplitude)
- **Interpolation and smoothing:** Spatial filters are applied to reduce speckle noise and enhance continuity

The output B-scan is a reconstructed cross-sectional image. However:

- The reconstruction assumes **homogeneous wave velocity**, which may not hold in heterogeneous concrete or near voids
- **Inversion artifacts** can create false contrasts in regions with poor aperture coverage (near the edges or at shallow depths)
- **Amplitude normalization** scales the image to full dynamic range, making absolute reflection strength non-comparable between different scans

Inspector does not work with the raw I/Q data—commercial MIRA software automatically performs reconstruction and displays only the final tomogram (color-coded velocity or amplitude map), see Figure 24.

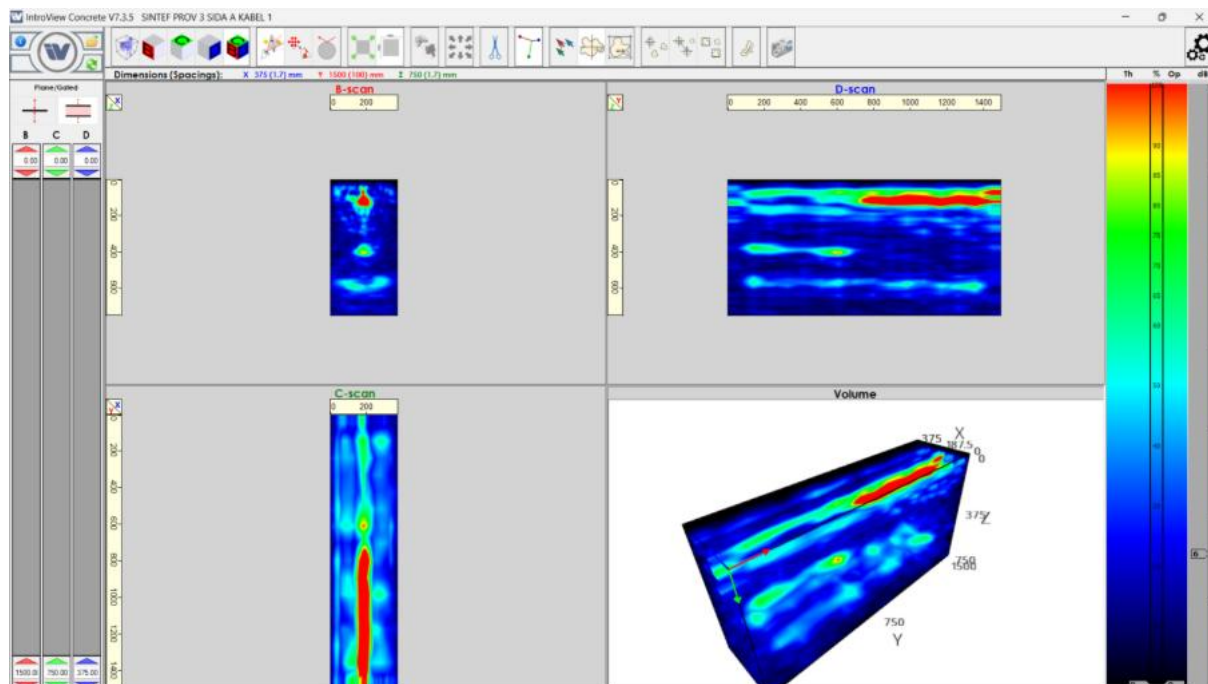


Figure 24 B, C, D scans and volumetric visualization from MIRA on commercial software interface

Impact Echo Frequency Spectrum Extraction

IE processing typically includes:

- Bandpass filtering (1–20 kHz)
- Hanning windowing
- FFT computation
- Peak detection
- Thickness estimation via

$$T = \frac{C_p}{2f}$$

where:

C_p = P-wave velocity in concrete (typically 3000–4500 m/s)

f = dominant frequency (Hz)

T = slab or structural thickness (m)

The software displays frequency spectra and thickness maps (Figure 25). However, important time-domain information—such as damping characteristics and multiple wave arrivals—is not retained.

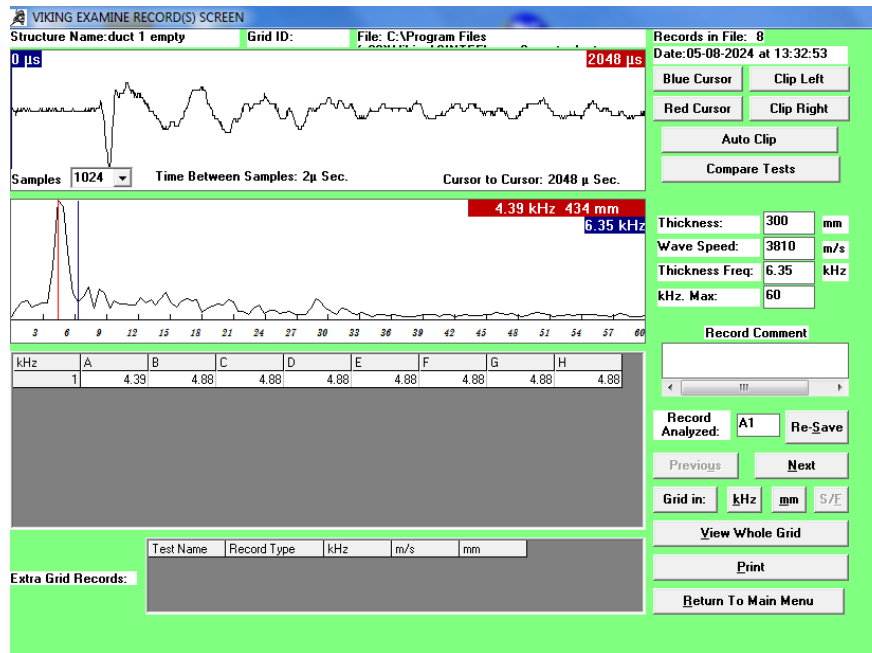


Figure 25 Frequency spectra of an IE test on commercial software interface

6.1.4 Phase 4: Visual interpretation and expert judgment

At this stage, inspectors interpret processed visualizations to identify potential defects. This step is inherently qualitative and heavily influenced by experience.

GPR Interpretation

Inspectors identify:

- **Hyperbolas** (point reflectors such as rebars or ducts)
- **Planar reflectors** (delaminations or layers)
- **Shadow zones** (behind conductive or attenuating objects)
- **Amplitude anomalies**

Interpretation is complicated by overlapping reflections, uncertain wave velocities, and artifacts introduced during filtering and migration.

MIRA Interpretation

Inspectors evaluate tomograms for:

- **Velocity contrasts** indicating heterogeneity
- **Shadowed regions** suggesting voids
- **Geometric boundaries** of ducts or embedded objects

Challenges include SAFT artifacts, uncertain velocity assumptions, and subjective interpretation of colour scales.

Impact Echo Interpretation

Inspectors analyse:

- Thickness deviations
- Peak absence or splitting
- Clusters of anomalies across the grid

Ambiguity arises when multiple peaks compete or when grid spacing fails to capture small defects.

6.1.5 Phase 5: Decision-making and reporting

Finally, inspectors synthesize information from all available NDT modalities and classify structural regions into:

- Normal (sound)
- Suspicious (requires further investigation)
- Defective (confirmed anomaly)

Recommendations may include coring, repair, additional scanning, or continued monitoring.

Critical limitations of this decision-making phase include:

- Absence of quantitative confidence levels
- Lack of systematic feature extraction (decisions based on visual impression)
- Strong operator dependence, especially in borderline cases
- Influence of proprietary visualization choices, which may highlight or suppress features

6.2 Limitations of reliance on processed data

The conventional human-inspector workflow, while widely practiced and valuable for detecting prominent defects, introduces systematic limitations that compromise diagnostic accuracy, repeatability, and objectivity. In addition, preprocessing NDT data can introduce information loss, particularly through techniques like denoising or dimensionality reduction, which may reduce sensitivity to small or low-contrast anomalies.

These limitations arise directly from the reliance on **processed visualizations** rather than **raw physical measurements**.

6.2.1 Subjectivity and operator dependency

Human visual interpretation is inherently subjective, and even highly trained inspectors exhibit notable variability when reviewing the same processed dataset. Differences in perceptual thresholds, experience, cognitive heuristics, and fatigue all contribute to inconsistent evaluations.

For instance, inspectors vary in their ability to detect faint features or low-amplitude signals within noisy radargrams or tomograms. A weak contrast that one inspector considers indicative of an anomaly may be dismissed as noise by another. This issue is especially pronounced in GPR radargrams with low signal-to-noise ratios, where overlapping reflections obscure subtle boundaries, and in MIRA tomograms where gradual color transitions make the definition of “anomaly boundaries” ambiguous.

Experience-based mental templates further influence interpretation. Experts often rely on familiar patterns—such as the classic hyperbola associated with voids in GPR—yet anomalies that deviate from these templates, including irregular void geometries, partial grouting, or inclined cracks, may be overlooked. Confirmation bias can also arise when inspectors expect a structure to follow a specific reinforcement or duct configuration, leading them to interpret ambiguous features as consistent with that expectation.

Additionally, large-area inspections with dense datasets impose substantial cognitive load. Fatigue has been shown in related fields such as radiology to reduce diagnostic accuracy by 4–13% over the course of a work shift, particularly during repetitive review tasks. Similar degradation is expected during extended review of NDT images such as long sequences of MIRA tomograms, increasing the risk of missed detections (Taylor-Phillips & Stinton, 2019).

6.2.2 Device-specific processing biases

NDT instruments from different manufacturers incorporate proprietary pre-processing algorithms that differ in filtering strategies, gain functions, migration techniques, reconstruction assumptions, and visualization schemes. Consequently, identical raw signals may be transformed into notably different visual outputs across platforms, complicating both interpretation and cross-device data consistency.

In GPR, radargrams generated by systems using similar antennas can appear markedly different due to disparate gain curves, background removal practices, and colormap selections. A slab scanned with three commercial systems may yield one radargram with strong exponential gain, another with linear gain emphasizing shallow reflections, and a third rendered using automatic gain control that equalizes amplitudes across depth. Similar inconsistencies occur in MIRA tomography, where SAFT reconstructions differ based on interpolation schemes, velocity model assumptions, and amplitude scaling methods. Even seemingly minor differences—such as bilinear versus cubic spline interpolation—affect boundary sharpness and pixel uniformity.

Impact Echo systems face comparable variability. Software must select a dominant frequency peak from the spectrum to estimate thickness, yet algorithms differ substantially. Some tools simply select the maximum amplitude peak, making them vulnerable to noise-induced artifacts; others restrict the search to an expected frequency band, which is robust but may overlook unexpected resonance behaviour associated with anomalies; and some provide multiple peaks and defer final selection to the user, reintroducing operator subjectivity. Such device-specific processing biases complicate comparisons across systems and can mislead inspectors, who typically lack access to the raw, unprocessed signals from which these visualizations originate.

6.2.3 Probability of missing anomalies

The most consequential limitation of the conventional processed-data workflow is the elevated risk of false negatives—situations in which defects physically present in the structure remain undetected because their signatures are weakened, obscured, or removed during preprocessing or visual interpretation.

Preprocessing-induced invisibility is a significant contributor. Small voids in GPR may produce reflections so weak that, after background removal or gain normalization, they fall below the display threshold and effectively disappear. In MIRA, partially grouted ducts producing intermediate-amplitude reflections may be misclassified as fully grouted because they do not match the high-contrast patterns inspectors expect. In IE testing, broad low-amplitude resonances associated with distributed voids may be overshadowed by sharper peaks from surrounding concrete, leading to incorrect thickness estimation and overlooked defects.

Complex structural environments further exacerbate detection challenges. Overlapping reflections from dense reinforcement can mask voids in GPR data, while multiple ducts within MIRA tomograms may produce interacting reflections that obscure individual boundaries. When visual overlaps occur, inspectors often cannot disentangle the contributing features, increasing the likelihood of missed anomalies.

Depth-dependent attenuation introduces additional uncertainty. Weak signals from deep defects may be rendered indistinguishable from background noise after normalization, leaving inspectors unable to judge whether low contrast reflects true uniformity or attenuation-limited detectability. Without access to raw signal amplitudes, inspectors cannot make this distinction reliably.

Controlled laboratory studies illustrate these limitations: GPR typically detects 80–90% of large voids (>50 mm) but only 50–70% of smaller ones; MIRA performs well for empty ducts (>95%) but poorly for partially grouted ducts (60–75% detection depending on defect size and depth); IE reliably detects large delaminations (85–95%) but struggles when voids are smaller than one-third of slab thickness, with detection rates dropping to around 60%. Although these findings (Razak et al., 2015; J.-K. Zhang et al., 2016) vary with experimental conditions, they underscore the vulnerability of processed-data inspection to subtle or ambiguous anomalies.

In real-world inspections, the actual defect population is unknown, rendering false-negative rates unquantifiable. A structure may therefore be certified as defect-free even when subtle defects exist but remain undetected due to preprocessing-induced information loss or human oversight.

6.3 Motivation for machine learning on raw data

The limitations discussed in Section 6.2 highlight the need for a fundamental shift in NDT methodology—from human interpretation of processed visualizations to ML-based analysis of raw physical measurements to support the human process. This section outlines the rationale for this paradigm shift, the advantages of raw-data-driven ML approaches, and the technical feasibility demonstrated in the current research project.

6.3.1 Fundamental premise: Raw data preserves complete physical information

Every NDT measurement is fundamentally a **physical signal**: electromagnetic waveforms in GPR, acoustic pressure waves in MIRA, or vibrational accelerations in IE. These raw signals capture the full interaction between the probing wave and the material under investigation and contain information such as:

- **Amplitude:** Reflection strength proportional to impedance contrast
- **Phase:** Wave travel time and interference behaviour
- **Frequency content:** Wavelength-dependent scattering and absorption
- **Temporal structure:** Multi-path reflections, envelope decay, and secondary arrivals
- **Spatial coherence:** Correlation across adjacent channels or measurement points

Conventional preprocessing steps—filtering, gain adjustment, background removal, inversion, and FFT extraction—discard portions of this information to generate visually interpretable outputs. While this is necessary for human interpretation, it is not required for ML models. ML algorithms can operate directly on raw time-series data, preserving all physically available information.

Training ML models on raw data ensures that subtle diagnostic features—weak reflections, high-frequency signatures, phase-encoded patterns, or spatially diffuse anomalies—remain accessible to the classifier. In contrast, models trained on processed images are inherently limited to features that survive preprocessing, reducing detection sensitivity.

6.3.2 Automatic learning of subtle, multi-dimensional patterns

Human inspectors excel at identifying simple, high-contrast visual features (e.g., hyperbolas or shadows) but struggle with:

- Low-amplitude reflections near the noise floor
- Gradual or distributed anomalies lacking distinct boundaries
- Complex, multi-dimensional correlations involving amplitude, phase, frequency, and spatial context

ML models trained on high-dimensional raw feature vectors can learn nonlinear decision boundaries that are not visually apparent.

Feature richness in raw data

GPR (A-scans): Each trace provides 30+ features across categories such as:

- Time-domain: peak amplitude, RMS, energy, zero-crossing rate
- Envelope: instantaneous amplitude statistics, decay rate
- Frequency-domain: dominant frequency, bandwidth, spectral centroid
- Attenuation: early–late energy ratio
- Statistical: skewness, kurtosis

MIRA (12-channel I/Q measurements): Each measurement yields 15+ features including:

- Amplitude: peak values, mean, energy
- Temporal: arrival times
- Spectral: dominant frequency, attenuation, bandwidth
- Multi-channel coherence: cross-correlation between channels (information lost after SAFT reconstruction)

IE: Each time-domain signal provides 20+ features such as:

- Frequency-domain: dominant frequency, Q-factor, harmonic content
- Spectral shape: centroid, spread, roll-off
- Thickness estimates: apparent thickness and deviation from expected values
- Spatial context: deviations relative to neighbouring grid cells

Evaluating these features simultaneously requires multi-dimensional, nonlinear analysis—well suited for ML models but unattainable through human visual inspection alone.

6.3.3 Objectivity, consistency, and quantified confidence

A key advantage of ML-based analysis is the objectivity and consistency it provides. Given identical inputs, an ML model produces identical outputs, eliminating operator-to-operator variability, fatigue effects, and temporal inconsistency. ML classifiers also generate probabilistic confidence estimates, allowing each classification—such as the presence or absence of a void—to be expressed in terms of likelihood rather than a binary decision. High-confidence predictions can be accepted directly, low-confidence predictions can be treated as normal, and intermediate cases can be flagged for expert review. This stratification supports a more transparent and risk-informed inspection process.

Furthermore, ML models such as Random Forests provide interpretable feature importance rankings. These rankings reveal which physical characteristics are most diagnostic, enabling researchers to validate model behaviour and ensuring that decisions are grounded in physically meaningful attributes. For example, if an Impact Echo model assigns high importance to the Q-factor or harmonic content, this confirms that the model has learned relationships consistent with known resonance-based indicators of voids. Such insights also guide future sensor development by highlighting which measurement properties contribute most strongly to detection performance.

6.3.4 Scalability and real-time deployment potential

Modern NDT campaigns generate extremely large datasets, often consisting of tens of thousands of GPR traces, hundreds to thousands of MIRA measurements along ducts, or dense grids of IE test points. Manual inspection of such extensive datasets is time-consuming and operationally inefficient. For instance, reviewing a set of 500 MIRA B-scans at an average pace of thirty seconds per scan requires more than four hours. ML inference, by contrast, can process thousands of samples per second on standard hardware and tens of thousands of samples per second on GPUs.

This computational efficiency enables real-time inspection workflows. Field crews can acquire data using handheld devices while an integrated or cloud-connected ML pipeline processes raw measurements immediately, performs feature extraction, applies trained models, and generates confidence-based anomaly maps. Inspectors receive instant visual feedback through color-coded displays highlighting normal, uncertain, and defective regions. Follow-up investigations can then focus exclusively on flagged areas, dramatically improving operational efficiency and reducing overall inspection cost consequently.

6.3.5 Hybrid workflow: ML as a decision-support, not replacement

A central principle of this project is that ML should augment, rather than replace, human expertise. The recommended hybrid workflow incorporates ML analysis as a supporting “second opinion,” ensuring that human judgment remains the final authority in safety-critical decisions. Raw data are first processed by trained models to generate probability-based anomaly maps. Regions with high predicted probability of defects are prioritized for review, areas with very low probability are treated as normal unless contrary visual evidence is present, and intermediate-probability regions are presented to inspectors together with confidence metrics. Inspector decisions are recorded and fed back into periodic model retraining, enabling continuous improvement of model performance through an adaptive learning cycle.

7 Machine learning applications for automated NDT analysis

This chapter explains how machine learning models were developed to automatically analyse the NDT measurements and identify voids in post-tensioned ducts. The approach starts by recognizing that the three NDT methods—radar, ultrasonic, and impact-echo—work on completely different physics and generate fundamentally different types of data; Each requires a tailored analysis strategy.

Rather than feeding raw signals directly into complex neural networks and hoping they figure everything out, the approach emphasizes extracting meaningful physical measurements first—things like wave reflection strength, frequency content, and how consistently signals arrive across multiple sensors. This physics-informed strategy has major advantages: it works well with limited training data (hundreds of measurements rather than millions), produces decisions that engineers can verify and trust, and runs fast enough for real-time field use on laptop computers without requiring specialized hardware.

The chapter walks through the entire pipeline: selecting appropriate machine learning models for each data type, extracting features that capture the physical signatures of voids versus filled ducts, reducing hundreds of potential measurements down to the most informative dozen or so, and using data augmentation techniques to simulate natural measurement variability. Throughout, the emphasis remains on creating practical tools that work reliably under real inspection conditions rather than just achieving impressive numbers in controlled laboratory tests.

7.1 Overview of three NDT modalities

The three NDT modalities employed in this research—GPR, UPE, and IE—generate fundamentally different data types, spatial sampling patterns, and physical signatures. Understanding these differences is crucial because they directly influence which ML strategies will be most effective for each modality. Table 7-1 provides a comparative overview of these key characteristics and their implications for ML implementation.

Table 7-1 Comparative Overview of NDT Modalities and ML Implications

Modality	Physical Principle	Raw Data Type	Spatial Sampling	Typical Depth	Primary Diagnostic Target	Most Suitable ML Approach
GPR	EM wave reflection	SEG-Y (dense A-scans)	Continuous area scan	0.3–1.5 m	Rebars, ducts, cavities, delaminations	Raw-trace classical ML or image-based deep learning (U-Net)
UPE	Shear-wave multi-channel tomography	Complex I/Q multi-channel waveforms	20 cm wide linear array	0.1–0.8 m	Voids & grouting condition in ducts	Physics-informed classical ML (coherence + amplitude features)
IE	P-wave resonance	Single time-series per point	Discrete grid (10–20 cm)	0.1–0.7 m	Thickness anomalies, voids, delamination	Classical ML on hand-crafted spectral + spatial features or 1D-CNN

These complementary characteristics motivated the development of a hybrid ML framework that combines physics-informed feature engineering with both classical and deep learning models. The distinct physical principles underlying each method naturally suggest different computational approaches. GPR produces dense 2D and 3D images that lend themselves well to Convolutional Neural Network (CNN), which excel at recognizing spatial patterns like the characteristic hyperbolic reflections from embedded objects. MIRA (UPE) generates multi-channel, physics-rich waveforms where the relationships between channels carry critical information about wave propagation and scattering—making physics-informed feature engineering particularly valuable for extracting coherence and amplitude characteristics. IE, by contrast, generates sparse, point-wise spectral data collected on a discrete grid, which makes it ideal for classical ML approaches that can leverage both the spectral signatures and spatial neighbourhood features to distinguish genuine defects from isolated measurement artifacts.

7.2 Recent advances in ML for NDT applications

Machine learning has increasingly been adopted for automating the interpretation of NDT data in concrete infrastructure. The past few years have witnessed substantial progress across all three modalities, with research demonstrating ML's potential to improve detection, localization, and classification of defects such as voids, cracks, delaminations, and embedded components in concrete structures.

7.2.1 Ground-Penetrating Radar (GPR)

Recent studies have successfully applied deep convolutional models directly to GPR B-scan images, achieving impressive performance in defect detection. (Liu et al., 2023) trained SSD and YOLOv4 object detection architectures on more than 4,200 GPR scans from tunnel linings, achieving up to 98% accuracy for detecting missing rebars and over 86% accuracy for cavity detection. These results highlight the natural fit between object-detection architectures and radargrams, where structural patterns such as hyperbolas can be learned effectively through deep learning.

Building on this approach, (J.-Y. Zhang et al., 2025) implemented an attention-enhanced YOLOv5 architecture incorporating Efficient Channel Attention (ECA) modules and DCGAN-based data augmentation for UAV-based GPR inspection of concrete surfaces. Their model achieved 85.4% mean average precision (mAP), outperforming baseline detectors and demonstrating the value of attention mechanisms for focusing on relevant features within noisy GPR data.

More advanced representations continue to emerge in the literature. (Kang et al., 2025) proposed a shape-aware topological embedding for GPR signals combined with a deep neural network, improving robustness to noise and subsurface clutter that typically plague GPR interpretation. These developments indicate a broader trend toward integrating both geometric and learned features for improved subsurface detection, moving beyond simple end-to-end learning toward hybrid approaches that incorporate domain knowledge.

Overall, the GPR-ML literature demonstrates that when sufficient data are available—typically thousands of labelled scans—deep networks can effectively automate and accelerate defect detection with good accuracy, substantially reducing the reliance on manual interpretation of processed radargrams that currently dominates inspection practice.

7.2.2 Ultrasonic Pulse Echo (UPE)

Deep learning has also been increasingly applied to UPE data, though the literature base remains smaller than for GPR. (Alqurashi et al., 2025) developed a lightweight CNN-based detector trained on SAFT-reconstructed MIRA slices, achieving mean average precision of approximately 0.73 and intersection-over-union scores around 0.80 when identifying delaminations, rebars, and ducts. Their processing pipeline incorporated histogram matching and semi-automatic segmentation to standardize labelling across different structures and measurement conditions—highlighting the critical importance of preprocessing and normalization for tomographic ML applications.

Physics-informed approaches are also gaining traction in the ultrasonic domain. (Lee & Popovics, 2025) applied physics-informed neural networks (PINNs) to contactless ultrasonic scanning data to reconstruct spatial maps of material properties, outperforming traditional inversion methods in terms of spatial consistency and artifact reduction. This work exemplifies how embedding physical constraints—in this case, wave propagation equations—directly into the neural network architecture can improve both performance and interpretability.

However, a recent systematic review by (Xu et al., 2024) notes that the adoption of ML for ultrasonic NDT in construction remains limited by two fundamental challenges: the lack of large, well-labelled datasets, and the inherent heterogeneity of concrete, which introduces significant noise and variability into ultrasonic measurements. Concrete's complex microstructure—varying aggregate sizes, air voids, microcracks—creates scattering that complicates signal interpretation. This reality makes careful feature extraction, signal normalization, and data curation essential for successful ML implementation with ultrasonic methods.

7.2.3 Impact Echo (IE)

IE generates point-wise time-series signals at discrete measurement locations, traditionally interpreted through frequency-domain analysis to extract dominant frequencies for thickness estimation or delamination detection. Recent ML applications have extended this approach by applying classical algorithms to engineered spectral features or deep learning to raw signals.

(Sayyar-Roudsari et al., 2024) demonstrated the potential of this approach through a systematic study of twelve reinforced concrete beams with controlled defects (voids, corrosion, debonding). Their multi-stage framework progressed from linear regression establishing relationships between wave properties and mechanical characteristics, through artificial neural networks with Imperialist Competitive Algorithms for material property prediction, to Support Vector Machine classification of defect types. The strong performance across precision, recall, and F1-score metrics confirmed that classical ML techniques can effectively distinguish internal defects when provided with appropriate IE features.

However, the literature for ML-enhanced IE remains more limited than for GPR or UPE, reflecting the method's more recent standardization, point-based measurement approach yielding smaller datasets, and sensitivity to operator technique. Future advances will likely emerge through increased field data collection, sophisticated feature engineering capturing both spectral and spatial patterns, and—most promisingly—multi-modal integration where IE's resonance sensitivity complements electromagnetic and ultrasonic methods.

Success ultimately depends on capturing subtle signal variations (attenuation, harmonics, decay characteristics) while maintaining well-labelled datasets, and balancing physically interpretable features with sufficient flexibility for complex pattern recognition. These considerations inform the feature extraction and model selection strategies detailed in subsequent sections.

7.2.4 Multi-Modal integration

Recent research has begun exploring the combination of multiple NDT modalities to improve defect detection robustness—an approach that directly motivates the present work. (Datta Rachuri et al., 2024) introduced a multimodal fusion framework integrating IE, Ultrasonic Surface Waves (USW), geospatial mapping, and advanced image processing for bridge inspection. The study achieved an F1 score of 0.83, demonstrating that combining modalities significantly enhances localization accuracy and reduces false positives through cross-verification. This methodology proves particularly valuable for large-scale bridge inspections where the ambiguity inherent in any single method must be minimized to avoid both missed defects (false negatives) and unnecessary follow-up investigations (false positives).

This emerging trend toward multimodal NDT-ML fusion highlights the potential of integrated frameworks for robust structural assessment—an approach that lies at the heart of the present research and its future expansion possibilities.

7.3 Machine learning architectures and model development

7.3.1 Model selection criteria and architectures

To ensure robust and complementary performance across the different NDT modalities, we developed three primary ML models: a Support Vector Machine (SVM), a Random Forest (RF), and a 1D Convolutional Neural Network (1D-CNN). Additionally, a lightweight U-Net serves as a comparison baseline for image-based GPR segmentation. Each model brings unique inductive biases and strengths to the problem, enabling both cross-validation of findings and potential ensemble-based improvements in the final deployment system.

The selection of these particular architectures reflects several practical considerations specific to NDT inspection contexts. First, training data availability remains limited—we work with hundreds rather than millions of samples, a regime where classical ML often outperforms deep learning. Second, safety-critical infrastructure applications require physically interpretable decisions that can be verified by human experts and defended in regulatory proceedings. Third, the deployment scenario demands calibrated probability estimates to guide human oversight, not just binary classifications. Fourth, field-deployable systems face computational constraints—models must run efficiently on standard laptops or tablets rather than requiring GPU clusters.

Classical ML approaches, particularly SVM and Random Forest, excel in these conditions. They handle small datasets efficiently through built-in regularization mechanisms, produce transparent decision boundaries that can be understood and verified, and execute rapidly on modest hardware. Deep learning is applied selectively in this work, deployed only where sufficient data exists and where the inherent spatial or spectral structure of the problem justifies the additional architectural complexity.

7.3.2 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models originally designed for binary classification tasks. The core concept behind SVM is elegant: among all possible hyperplanes that could separate two classes, choose the one that maximizes the margin—the distance to the nearest training examples from either class. Figure 26 illustrates this fundamental principle, showing how the optimal separating hyperplane positions itself equidistant from the closest points of each class.

The data points closest to this decision boundary are called support vectors, and they alone determine the model's geometry and robustness—all other training points could be removed without changing the classifier.

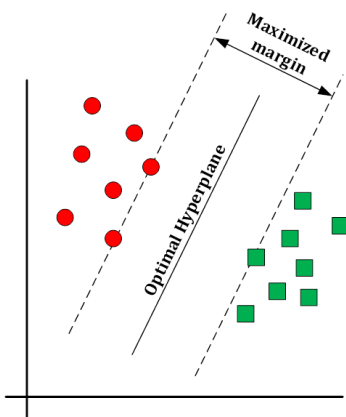


Figure 26 Demonstration of SVM classifier showing maximum margin hyperplane and support vectors; Adapted from (Cortes & Vapnik, 1995)

This margin maximization leads to better generalization and reduced overfitting, a property formalized as structural risk minimization in the foundational work by (Cortes & Vapnik, 1995). For datasets that are not linearly separable in the original feature space, an SVM uses the kernel trick: instead of explicitly transforming the data into a higher-dimensional space (which would be computationally expensive), the kernel function computes the inner product in that high-dimensional space implicitly. The most popular kernel is the Radial Basis Function (RBF), also called the Gaussian kernel, which can create highly flexible, non-linear decision boundaries. The RBF kernel is defined as:

$$K(x, x') = \exp(-\gamma ||x - x'||^2)$$

where γ controls the influence radius of individual training samples. Figure 27 shows a schematic visualization of how an SVM with an RBF kernel works on a 2D non-linearly separable problem. Real decision boundaries in high-dimensional RBF space resemble smooth, closed contours around clusters.

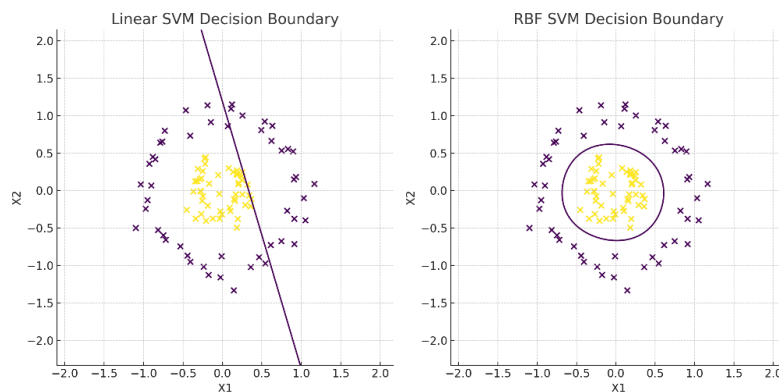


Figure 27 Demonstration of Linear versus RBF SVM decision boundary, illustrating kernel-induced feature space transformation; Adapted from (Cortes & Vapnik, 1995)

When to Choose SVM:

- Small to medium datasets (100–10,000 samples)
- Moderate to high dimensional feature spaces (20–100 features)
- Class imbalance (via `class_weight="balanced"`)
- Applications requiring calibrated probability estimates
- Preference for transparent, theoretically grounded models

Hyperparameter Configuration

For this research, we optimized SVM hyperparameters through 5-fold stratified cross-validation combined with grid search over carefully chosen ranges. Stratified splitting ensures that each fold maintains the same class balance as the overall dataset—critical when working with imbalanced data where voids represent only 15-20% of measurements.

The key hyperparameters and their tested ranges were:

- **C (regularization parameter):** [0.1, 1, 10, 100]—controls the trade-off between maximizing the margin and minimizing training errors. Small C values prioritize a large margin even if some training points are misclassified; large C values insist on correctly classifying training data even if it requires a narrower margin.
- **γ (RBF kernel width):** [0.001, 0.01, 0.1, 1]—controls decision boundary smoothness and complexity. Small γ creates smooth, simple boundaries; large γ creates complex boundaries that can fit intricate patterns but risk overfitting.
- **class_weight:** "balanced"—automatically adjusts weights inversely proportional to class frequencies, ensuring that the rarer void class receives appropriate attention during training rather than being overwhelmed by the more numerous filled examples.

We implemented SVM using scikit-learn version 1.3.0 with probability=True enabled, which activates Platt scaling to produce calibrated confidence scores rather than just binary predictions. These probability estimates prove essential for the multi-modal fusion approach described later, where we need to weight predictions from different methods according to their confidence levels.

7.3.3 Random Forest (RF)

Random Forests take a fundamentally different approach to classification. Rather than seeking a single optimal decision boundary like SVM, Random Forest creates an ensemble of many decision trees, each trained on a slightly different subset of the data. Figure 28 illustrates this ensemble architecture—multiple trees are grown independently, each making its own prediction, and the final classification emerges through majority voting across all trees.

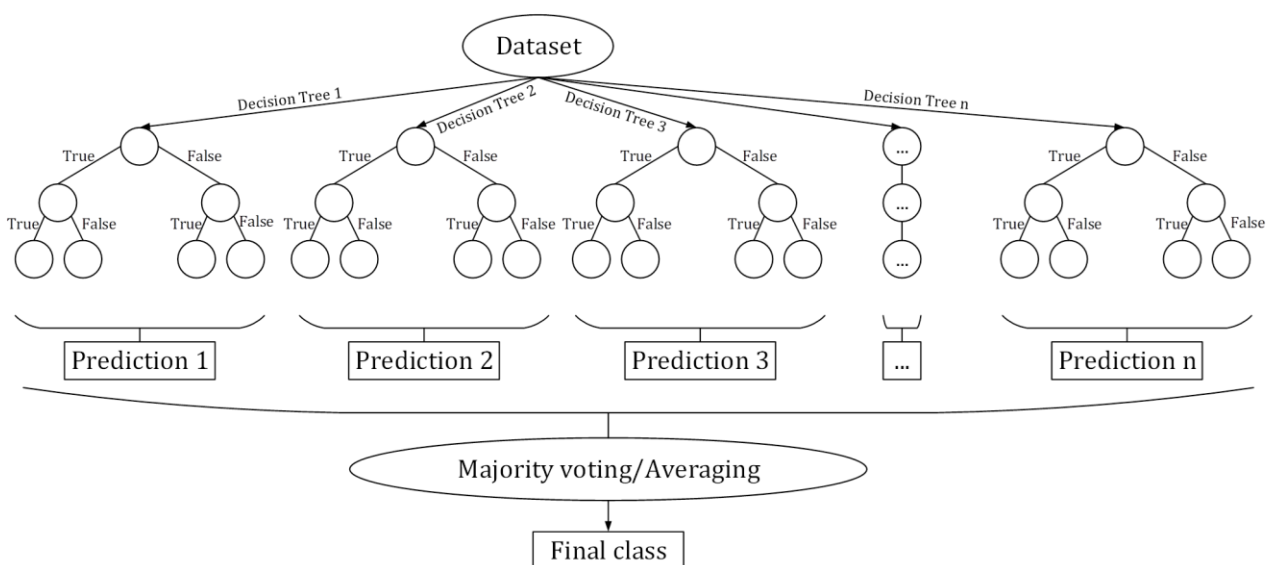


Figure 28 Demonstration of Random Forest classifier showing bootstrap sampling, tree training, and majority voting; Adapted from (Ho, 1995)

The power of Random Forest lies in a counterintuitive principle: each individual tree is intentionally kept shallow and somewhat inaccurate, but when aggregated through voting, their errors cancel out while their correct predictions reinforce each other. This ensemble approach, originally proposed by (Ho, 1995), achieves variance reduction without increasing bias—a rare and valuable property in ML that makes Random Forest remarkably robust and stable across different datasets and problem domains.

Random Forests naturally handle non-linear interactions between features without requiring explicit kernel functions, and they demand minimal hyperparameter tuning compared to most other ML algorithms. The algorithm introduces randomness at two distinct levels that work together to decorrelate the individual trees. First, each tree trains on a bootstrap sample—a random selection drawn with replacement from the training data, meaning some examples appear multiple times while others are omitted entirely. Second, at each node where the tree decides how to split the data, only a random subset of features is considered as candidates, typically $\sqrt{n_features}$ for classification tasks. This double randomization ensures that individual trees make different mistakes rather than all failing in the same way, which is precisely what enables the ensemble to succeed where individual components are mediocre.

Optimal Use Cases:

Random Forest excels in scenarios that closely match our NDT inspection problem:

- Tabular datasets with mixed feature types—amplitude measurements, spectral characteristics, temporal patterns, spatial relationships all appear in our feature sets
- Strong class imbalance—the "balanced_subsample" variant handles this naturally
- Need for interpretable feature importance rankings—Random Forest provides these automatically as a byproduct of training
- Applications requiring high accuracy with minimal tuning effort
- Small to medium datasets ranging from hundreds to tens of thousands of samples
- Problems involving complex non-linearities and feature interactions that are difficult to model explicitly

Hyperparameter Configuration

Random Forest hyperparameters were optimized through 5-fold stratified cross-validation, exploring the following ranges:

- **n_estimators:** [100, 200, 300]—the number of trees in the forest. More trees generally improve performance and stability at the cost of longer training and inference time.
- **max_depth:** [10, 20, 30, None]—maximum depth of individual trees. Limited depth prevents overfitting; unlimited depth (None) allows trees to grow until leaves are pure or other stopping criteria are met.
- **min_samples_split:** [2, 5, 10]—minimum number of samples required to split an internal node. Higher values prevent the tree from learning very specific patterns that may not generalize.
- **min_samples_leaf:** [1, 2, 4]—minimum number of samples required at each leaf node. This is another regularization parameter that prevents the tree from memorizing individual training examples.
- **max_features:** ['sqrt', 'log2']—number of features considered at each split point. 'sqrt' means $\sqrt{n_features}$, which works well for most classification problems.
- **class_weight:** "balanced_subsample"—applies sample weights within each bootstrap sample, ensuring rare classes receive appropriate attention even in imbalanced datasets.

We used scikit-learn version 1.3.0, which provides two particularly valuable features for model interpretation and validation. First, out-of-bag (OOB) score estimation uses the training examples that were excluded from each tree's bootstrap sample to generate an unbiased performance estimate without requiring a separate validation set. Second, feature importance via Gini impurity reduction quantifies how much each feature contributes to reducing uncertainty in the predictions, providing insights into which physical measurements most strongly influence void detection.

7.3.4 1D-Convolutional Neural Network (1D-CNN) – Spectral analysis

Impact Echo spectra often contain complex diagnostic patterns that resist straightforward feature engineering. Split peaks indicating multiple reflection interfaces, harmonic distortions from non-linear interactions, subtle shoulder peaks suggesting partial voids, and noise-floor deformations from measurement conditions—these patterns are difficult to capture through hand-crafted features based on simple peak detection and spectral moments. A 1D-CNN addresses this limitation by learning these spectral signatures directly from the frequency-domain representation, automatically discovering multi-scale patterns through hierarchical feature learning.

This approach becomes advantageous when we have sufficient labelled spectra available, typically 150-200 samples or more. Below this threshold, the risk of overfitting outweighs the benefits of automatic feature discovery, and classical ML with carefully engineered features tends to outperform deep learning. The architecture we developed employs a progressively deepening structure that mirrors the hierarchical nature of spectral patterns: shallow layers capture broad resonance bands and fundamental frequencies, intermediate layers extract harmonic relationships and mid-level interactions, and deep layers detect fine-grained features like split peaks, sharp notches, and local distortions.

Training Configuration

The training procedure incorporated several best practices for small-dataset deep learning:

- **Optimizer:** Adam with $\text{learning_rate}=0.001$, $\beta_1=0.9$, $\beta_2=0.999$ —adaptive learning rates help navigate the optimization landscape efficiently
- **Loss function:** Binary cross-entropy with class weighting to handle the imbalance between void and filled examples
- **Batch size:** 16—small batches introduce useful noise into the gradient estimates, acting as implicit regularization
- **Epochs:** 100 maximum with early stopping (patience=15 epochs monitoring validation loss)—training terminates if validation performance doesn't improve for 15 consecutive epochs
- **Regularization:** L2 weight decay ($\lambda=0.01$) combined with dropout (rate=0.5)—two complementary regularization mechanisms that together prevent overfitting

When to choose 1D-CNN:

The 1D-CNN approach makes sense under specific conditions:

- Working directly with frequency spectra rather than pre-computed hand-crafted features
- Defect signatures are subtle and may shift in frequency, where global pooling provides robustness
- Interest in discovering new diagnostic patterns that domain experts might have missed
- Availability of at least 150-200 labeled spectra, preferably more when augmentation is applied
- Willingness to accept somewhat longer training times compared to Random Forest or SVM

Table 7-2 Architecture design rationale (Layer-by-Layer) for 1D-CNN Impact Echo Classifier

Layer / Block	Configuration	Output Shape	Rationale
Input	-	512×1	Raw or log-scaled spectrum matching standard FFT resolution
Conv1D Block 1	32 filters, kernel=7, ReLU	512×32	Wide kernel captures broad resonance structures and fundamental frequencies
MaxPooling	pool_size=2	256×32	Dimensionality reduction, adds translation invariance to frequency shifts
Conv1D Block 2	64 filters, kernel=5, ReLU	256×64	Extracts mid-level spectral interactions and harmonic relationships
MaxPooling	pool_size=2	128×64	Further abstraction and invariance
Conv1D Block 3	128 filters, kernel=3, ReLU	128×128	Detects fine-grained features: split peaks, sharp notches, local distortions
MaxPooling	pool_size=2	64×128	Final spatial reduction before global pooling
Global Average Pooling	-	128	Reduces to fixed-length feature vector; robust to frequency shifts and spectrum length variations
Dense	64 units, ReLU	64	High-level feature combination
Dropout	rate=0.5	64	Strong regularization prevents overfitting on small datasets
Dense (output)	1 unit, sigmoid	1	Binary classification with probability output

The architecture deliberately remains conservative—just 3 convolutional blocks compared to the dozens of layers common in state-of-the-art deep networks for image recognition. This shallow design reflects our small dataset regime. Combined with aggressive regularization through dropout, L2 weight decay, and early stopping, this conservative architecture prevents overfitting while retaining sufficient representational capacity to learn meaningful spectral patterns that distinguish filled from void conditions.

7.4 Feature extraction and selection

7.4.1 Philosophy: Physics-Informed feature engineering

The raw signals generated by GPR, MIRA, and IE contain rich physical information, but applying ML directly to unprocessed time-series data often yields disappointing results. The reasons are straightforward: high dimensionality creates computational challenges, measurements across time samples are often highly correlated, and the signals contain irrelevant noise alongside the diagnostic information we actually care about. Physics-informed feature engineering addresses these challenges by extracting a compact set of interpretable quantities that directly correspond to known physical mechanisms of wave propagation, reflection, and attenuation in concrete structures.

This approach offers several compelling advantages over the increasingly popular end-to-end deep learning paradigm where raw signals feed directly into neural networks. First and foremost is interpretability and trust. Each feature we extract has a clear physical meaning—something like "peak reflection amplitude from the duct interface" or "ultrasonic coherence between adjacent transducer channels." When a model classifies a tendon duct as void-filled based on low coherence and reduced amplitude, inspectors can verify this interpretation against their theoretical understanding and practical experience with similar structures. This transparency proves essential not just for day-to-day confidence in the system, but for regulatory acceptance and forensic investigation if disputed findings or structural failures occur years later.

Physics-informed features also dramatically improve sample efficiency. By explicitly encoding domain knowledge—the fact that voids cause coherence loss in ultrasonic transmission, for instance—the model only needs to learn the specific threshold values and feature combinations that indicate defects rather than having to rediscover fundamental wave physics from scratch. Our empirical experience bears this out: classical ML with physics features achieves competitive performance with 200-500 training samples, whereas end-to-end deep learning on raw signals typically demands thousands of examples to reach similar accuracy.

Robustness to distributional shifts represents another critical advantage. Features based on fundamental physical principles generalize better across varying concrete properties, equipment configurations, and environmental conditions than patterns learned purely from raw signals. Consider a feature quantifying "spectral centroid shift between reference and measured IE spectra." This remains meaningful whether the concrete is freshly cured or decades old, whether it contains limestone or granite aggregate, and even if alkali-silica reaction has introduced significant microcracking. The absolute frequency values certainly change across these conditions, but the relative shift pattern associated with voids persists because it reflects fundamental physics rather than superficial correlations in the training data.

Finally, computational efficiency matters for practical deployment. Classical ML models trained on extracted features execute orders of magnitude faster than deep neural networks processing raw signals. A Random Forest evaluating 50 features requires roughly 100 microseconds per prediction on a standard CPU—fast enough to enable real-time on-site classification during data acquisition. This efficiency facilitates immediate feedback to inspectors during field operations and supports vehicle-mounted scanning at traffic speeds without requiring specialized hardware.

The feature extraction pipeline follows a systematic methodology that we apply consistently across all three modalities. First comes signal conditioning, where we apply minimal preprocessing to remove obvious artifacts like electrical noise or positioning glitches while carefully preserving all physically meaningful information. We deliberately avoid gain adjustment, aggressive filtering, or normalization that could alter relative amplitudes or phase relationships—precisely the information that distinguishes filled from void conditions.

For certain modalities, particularly IE, domain transformation through FFT reveals diagnostic information more clearly than time-domain signals. Crucially, this transformation is information-preserving and lossless, unlike the visualization-oriented processing applied by commercial software. Next comes physical quantity extraction, where we compute features corresponding to specific wave-propagation phenomena: reflection coefficients encoding energy loss at interfaces, time-of-flight measurements revealing wave velocities and path lengths, coherence between multi-channel recordings capturing spatial consistency, spectral peak characteristics indicating resonance modes, and attenuation rates quantifying material damping.

For point-based measurements like IE, we integrate spatial context by extracting features that encode relationships with neighbouring measurement locations. This captures defect continuity and extent—genuine voids typically span multiple grid points rather than appearing as isolated anomalies that more likely represent measurement artifacts or operator errors. Finally, redundancy reduction through statistical analysis identifies highly correlated features that provide essentially duplicate information. Feature selection algorithms then retain the most informative subset, reducing dimensionality and overfitting risk while maintaining the physical interpretability that makes this approach valuable in the first place.

7.4.2 GPR raw signal features

7.4.2.1 *Extraction Methodology*

Ground-penetrating radar A-scans—the raw time-series traces—are extracted directly from SEG-Y format files without involving any commercial software processing that might alter or discard information. Each A-scan represents the reflected electromagnetic field amplitude as a function of two-way travel time at a single antenna position. For a typical bridge deck survey using a 1 GHz antenna with 2 cm trace spacing, each scan line generates 50-100 A-scans per meter of travel distance. Each individual A-scan contains 512 to 1024 time samples spanning roughly 0-50 nanoseconds, capturing reflections from both shallow and deep structural features.

Feature extraction operates on individual A-scans or small spatial windows encompassing 5-10 adjacent traces, which corresponds to a physical extent of 10-20 cm—approximately matching the diameter of typical tendon ducts. This spatial aggregation strikes a deliberate balance: we preserve sufficient localization accuracy to identify specific duct segments while improving signal-to-noise ratio through spatial averaging that suppresses random fluctuations.

The extraction workflow proceeds through four systematic steps. Step one involves direct surface reflection removal. The first strong reflection pulse, arriving around 0-2 nanoseconds after transmission, arises from the electromagnetic impedance contrast at the antenna-air-concrete interface. While this reflection provides useful information about surface conditions, it contains nothing about internal structure and often saturates the receiver, obscuring the weaker reflections we actually care about. We identify this surface reflection by its characteristic high amplitude and remarkably consistent arrival time across all traces in a scan line, then remove it using a simple time gate. Importantly, we avoid sophisticated "background removal" algorithms that attempt to suppress spatially continuous features—these often inadvertently eliminate genuine defect signatures that happen to extend over substantial areas.

Step two focuses on hyperbola apex detection. Individual rebars and tendon ducts generate characteristic hyperbolic reflection patterns in B-scan images because the antenna encounters them at oblique angles before and after passing directly overhead. The apex of this hyperbola—the point where the antenna sits directly above the target—produces the strongest reflection and most reliable depth information. We employ a straightforward peak-picking algorithm operating on the amplitude-time-space data cube. Apexes appear as local maxima in amplitude that maintain spatial coherence across 3-5 adjacent traces, distinguishing them from random noise spikes that don't exhibit this spatial structure.

For feature extraction, we use the A-scan corresponding to this apex position as the primary signal, with immediately adjacent traces providing spatial context.

Step three involves time-window selection for isolating the reflection of interest. For a detected target at depth d , the relevant signal information concentrates in a time window centered on the two-way travel time $t = 2d/v$, where $v \approx 0.1$ m/ns represents the electromagnetic wave velocity in concrete—a value that varies somewhat with moisture content and mix composition. We apply a window of ± 5 nanoseconds around this arrival time, which captures the reflection pulse and its immediate trailing energy while excluding earlier reflections from shallower features and later reflections from deeper structures. This windowing proves critical in reinforced concrete with multiple layers of steel, where overlapping reflections can otherwise confuse the analysis.

Finally, step four computes features from the isolated A-scan segment. We extract a suite of amplitude-based, temporal, spectral, and statistical features, each designed to capture specific aspects of the electromagnetic interaction with the embedded structure.

7.4.2.2 *Feature categories and physical interpretation*

The extracted GPR features fall into four complementary categories, each capturing different physical aspects of the electromagnetic signature.

A. Amplitude-Based Features: Direct reflection strength

These features quantify the strength of electromagnetic reflection from the duct interface, which depends critically on the dielectric contrast between the duct material, its internal contents, and the surrounding concrete matrix.

1- Peak Reflection Amplitude (A_{peak}) captures the maximum absolute amplitude within the target time window. The physical interpretation here is straightforward: metallic ducts produce strong reflections due to nearly 100% electromagnetic reflection at metal-dielectric interfaces—metals essentially act as mirrors for radio waves. Plastic or HDPE ducts produce weaker but still clearly detectable reflections. What makes this feature diagnostic is how the duct contents modulate the reflection. Air-filled voids create a strong dielectric contrast (concrete has $\epsilon \approx 6-9$ while air has $\epsilon \approx 1$), producing robust reflections. Grout-filled ducts reduce this contrast considerably since grout has $\epsilon \approx 5-7$, closely matching concrete, which results in much weaker reflections. We expect $A_{\text{peak}}(\text{void}) > A_{\text{peak}}(\text{filled})$ for plastic ducts, while metallic ducts show less dramatic differences but still exhibit detectable amplitude reduction when filled due to impedance matching effects at the concrete-grout and grout-steel interfaces.

2- Integrated Reflection Energy (E_{reflect}) sums the squared amplitudes within the target window: $E = \sum |A(t)|^2$. Rather than focusing solely on the peak, this measures total electromagnetic energy reflected back to the antenna by integrating over the entire pulse duration. This provides valuable robustness against pulse shape variations and noise spikes that might contaminate peak amplitude measurements. A single-sample noise spike could elevate the peak amplitude misleadingly, but it contributes minimally to the integrated energy. We consistently observe $E_{\text{reflect}}(\text{void}) > E_{\text{reflect}}(\text{filled})$, with the difference particularly pronounced for HDPE ducts where the air-HDPE dielectric contrast reaches its maximum.

3- Reflection Coefficient Estimate (R_{est}) expresses the ratio of reflected to incident wave amplitude, normalized by calibration measurements on known reflectors. The theoretical reflection coefficient at normal incidence follows from electromagnetic boundary conditions:

$$R = (\sqrt{\epsilon_2} - \sqrt{\epsilon_1}) / (\sqrt{\epsilon_2} + \sqrt{\epsilon_1})$$

where ϵ_1 and ϵ_2 represent the dielectric constants of the two media. For a concrete-air interface, this predicts $R \approx 0.5$, meaning about 50% of the incident energy reflects. For concrete-grout interfaces, $R \approx 0.1$, reflecting only about 10% of the energy. These predictions align well with our measurements: R_{est} (void) typically falls in the 0.4-0.6 range, while R_{est} (filled) clusters around 0.05-0.15, providing strong discriminatory power between the two conditions.

4- Amplitude Decay Rate (α_{decay}) fits an exponential decay constant to the amplitude envelope following the reflection pulse: $A(t) \propto \exp(-\alpha t)$. Electromagnetic attenuation in concrete depends on several factors including moisture content, chloride concentration from de-icing salts or marine exposure, and reinforcement density. Voids behind ducts create impedance discontinuities that can trap energy in resonant modes, altering the decay characteristics in complex ways. While the behavior is structure-dependent and not as universally interpretable as the direct amplitude features, we observe measurably different decay signatures for filled versus void conditions in controlled specimens.

B. Temporal Features: Arrival Time and Pulse Shape

Temporal features capture the timing and shape of the reflection pulse, encoding information about depth, duct geometry, and signal propagation characteristics.

5- Two-Way Travel Time (TWTT) measures the delay from transmission to peak reflection arrival. This provides a direct measure of target depth through $d = \frac{v \times TWTT}{2}$. Along a properly installed horizontal tendon duct, we expect TWTT to remain approximately constant, as the concrete cover thickness shouldn't vary significantly. Deviations from this constant value suggest either concrete cover thickness changes or duct alignment deviations—both construction anomalies worth noting even if they don't directly indicate grouting defects.

6- Pulse Width at Half Maximum (FWHM) quantifies the time duration over which amplitude exceeds 50% of the peak value. Thin targets like individual rebars produce narrow, crisp pulses because the reflection comes from a well-defined interface. Thick targets like ducts produce broader pulses due to multiple internal reflections between the duct walls and the possibility of reflections from both near and far duct surfaces. Pulse broadening also occurs when the target appears "smeared out" in depth due to oblique incidence or rough, irregular interfaces. We generally observe FWHM increasing with duct diameter, and filled ducts may show different pulse shapes than voids due to additional internal reflections at grout-steel interfaces.

7- Arrival Time Jitter (σ_{time}) computes the standard deviation of peak arrival times across 5-7 adjacent traces. Consistent arrival times indicate a smooth, continuous target with uniform cover depth—exactly what we expect for properly installed filled ducts. High jitter suggests irregular geometry, discontinuous features, or significant measurement noise. While low jitter characterizes intact filled ducts, elevated jitter may signal void boundaries, construction defects, or transitions between filled and void segments.

8- Pulse Asymmetry (Skewness) measures the statistical skewness of the amplitude distribution within the pulse window. Symmetric, approximately Gaussian pulses indicate simple reflector geometry—a single, smooth interface. Asymmetric pulses suggest complex multi-layer structure or oblique interfaces that create different reflection characteristics on the leading versus trailing edges. This feature potentially discriminates between simple void geometries and more complex partially grouted conditions where grout settles non-uniformly within the duct.

C. Spectral Features: Frequency Content Analysis

Transforming A-scans into the frequency domain via FFT reveals resonance frequencies, bandwidth characteristics, and high-frequency content preservation that remain obscured in time-domain representations.

9- Dominant Frequency (f_{dom}) identifies the frequency component with maximum amplitude in the FFT spectrum. For a well-calibrated GPR system, this should approximately match the antenna's nominal center frequency—1 GHz for a "1 GHz antenna," for instance. Shifts toward lower frequencies indicate selective attenuation of high-frequency components by lossy materials or geometric dispersion in complex layered structures. While we expect minimal shift between filled and void ducts—the dominant frequency primarily reflects the antenna characteristics—significant deviations serve as a valuable quality control indicator suggesting measurement problems or unusual propagation conditions.

10- Spectral Centroid (f_{centroid}) provides a more sensitive measure of frequency content by computing the center of mass of the amplitude spectrum:

$$f_c = \frac{\sum f |F(f)|}{\sum |F(f)|}$$

Unlike dominant frequency which focuses on the single strongest component, the centroid weighs all frequency components by their amplitudes. High-frequency attenuation systematically shifts the centroid toward lower frequencies. We consistently observe $f_{\text{centroid}}(\text{filled}) < f_{\text{centroid}}(\text{void})$ because electromagnetic losses in grout exceed those in air, preferentially attenuating the high-frequency components and pulling the spectral balance toward lower frequencies.

11- Bandwidth at -6 dB ($\text{BW}_{-6\text{dB}}$) defines the frequency range over which spectral amplitude exceeds 50% of the peak value—the -6 dB points in logarithmic units. Narrow bandwidth indicates selective frequency filtering or resonance effects that concentrate energy in a limited frequency range. Wide bandwidth preserves the antenna's original pulse characteristics with minimal frequency-dependent distortion. Filled ducts often exhibit narrower bandwidth due to frequency-dependent absorption in the grout matrix.

12- High-Frequency Content Ratio (HF_{ratio}) quantifies the ratio of spectral energy above a threshold—say, 1.5 GHz for a 1 GHz antenna—to the total spectral energy. High frequencies attenuate more rapidly in lossy media, following the general principle that absorption increases with frequency. Voids preserve high-frequency content better than grout-filled regions, leading to $\text{HF}_{\text{ratio}}(\text{void}) > \text{HF}_{\text{ratio}}(\text{filled})$ and providing useful discriminative power.

D. Statistical and Spatial Features: Signal Variability and Context

These features capture second-order statistics and spatial relationships that reflect measurement consistency and defect spatial extent.

13- Signal-to-Noise Ratio (SNR) computes the ratio of peak signal amplitude to the noise floor estimated from the pre-arrival time window where no reflections should appear:

$$\text{SNR} = \frac{A_{\text{max}}}{\sigma_{\text{noise}}}$$

High SNR indicates strong, reliable reflections with minimal contamination. Low SNR suggests either weak targets, excessive attenuation, or equipment issues. We use SNR primarily as a quality control metric, potentially excluding measurements with $\text{SNR} < 10 \text{ dB}$ as too unreliable for confident classification.

14- Spatial Amplitude Coefficient of Variation (CV_spatial) examines the coefficient of variation of peak amplitudes across 5-7 adjacent traces. Low CV values indicate a homogeneous target producing consistent reflections—exactly what continuous filled ducts should exhibit. High CV suggests irregular geometry or measurement artifacts creating amplitude fluctuations. We expect CV (filled continuous duct) < CV (void with irregular boundaries or partial grouting).

15- Amplitude Gradient ($\nabla A_{\text{spatial}}$) quantifies the spatial rate of change in amplitude perpendicular to the duct axis. Sharp amplitude gradients mark transitions between different conditions—filled and void segments create strong gradients at their boundaries. Smooth gradients indicate homogeneous conditions extending over substantial distances. High $|\nabla A|$ at void boundaries contrasts with low $|\nabla A|$ within uniform regions.

16- Hyperbola Symmetry Score (S_hyp) computes the cross-correlation between amplitude profiles on either side of the hyperbola apex. Symmetric hyperbolas indicate simple point-like reflectors—typical for uniformly filled ducts that behave electromagnetically like solid inclusions. Asymmetric patterns suggest complex scattering from irregular voids or interference from adjacent defects. We observe $S_{\text{hyp}} \approx 0.9\text{-}1.0$ for filled ducts while complex void geometries produce $S_{\text{hyp}} < 0.7$.

Summary of GPR Feature Set

We extract 16-20 features per duct location, deliberately limiting the total to avoid overfitting given our relatively small training datasets containing hundreds rather than thousands of samples. This feature set balances comprehensiveness—covering all major physical mechanisms by which voids alter electromagnetic signatures—with parsimony that maintains computational efficiency and reduces overfitting risk. Subsequent feature selection procedures, detailed in Section 7.4.5, further trim this set by identifying and removing redundant or low-importance features.

7.4.3 UPE Features

7.4.3.1 Unique aspects of MIRA data

The UPE tomography system operates fundamentally differently from both GPR and IE. It employs a linear array of 48 dry-point-contact transducers operating in shear-wave transmission mode, creating a data structure with no real parallel in the other modalities. Each measurement produces $48 \text{ transmit} \times 48 \text{ receive} = 2,304$ individual wave transmission paths, creating extraordinarily dense sampling of the internal concrete volume. The raw data for each transmit-receive pair consists of complex-valued In-phase/Quadrature (I/Q) samples representing both amplitude and phase of the received signal—information that commercial SAFT-processed B-scan visualizations completely discard in favour of simplified amplitude displays.

Several key characteristics distinguish MIRA data and motivate our feature extraction approach. First is multi-channel coherent recording. Unlike single-channel IE or GPR where each measurement stands alone, MIRA simultaneously records at 48 spatial locations. This enables coherence analysis between adjacent channels, which proves remarkably powerful for defect detection. When ultrasonic shear waves propagate through homogeneous concrete or grout-filled ducts, adjacent receivers detect highly correlated signals with small, predictable phase shifts determined solely by their spatial separation and the material's wave velocity. Voids, delaminations, or strong scatterers like voided duct boundaries destroy this spatial coherence, creating decorrelation that remains detectable even when absolute amplitudes prove ambiguous due to coupling variations or operator inconsistencies.

The I/Q data representation preserves full phase information, enabling interference analysis and coherence computation that would be impossible with amplitude-only data. Two signals might have identical amplitudes but completely different phases—one reflected from a filled duct interface, another from a void. These phase differences produce dramatically different coherence patterns when analysed across the transducer array, providing diagnostic information that amplitude measurements miss entirely.

MIRA's array geometry and signal processing inherently implement synthetic aperture focusing at different depths through the concrete thickness. This depth-dependent focusing means that features at various depths exhibit characteristic "focusing signatures" in the multi-channel data, helping distinguish actual structural features from electronic artifacts or near-field irregularities that don't focus properly. Finally, shear waves—with their transverse particle motion—prove more sensitive to microstructural damage like ASR-induced microcracking than the compressional waves used in IE. This creates both opportunities for detecting early-stage deterioration and challenges from elevated noise floors in damaged concrete, requiring careful feature engineering to extract robust diagnostic information.

7.4.3.2 *Raw A-scan features*

Feature extraction from MIRA data operates on the raw I/Q waveforms before any SAFT reconstruction, deliberately exploiting the multi-channel spatial diversity and phase information that commercial processing discards.

A. Single-Channel Amplitude Features

These features analyze the amplitude envelope of individual A-scans, analogous to GPR amplitude features but interpreted through the lens of ultrasonic propagation physics.

17- Peak Transmission Amplitude (A_trans_peak) identifies the maximum absolute amplitude across all transmit-receive pairs spanning the duct location. The physics of ultrasonic transmission through filled ducts involves multiple loss mechanisms: reflection losses at both the concrete-grout and grout-steel interfaces, plus volumetric absorption in the grout itself. Despite these losses, significant energy typically transmits through, reaching receivers on the far side of the duct. Voids create a catastrophically different situation. The concrete-air interface produces strong reflections that trap most of the incident energy, dramatically reducing transmitted amplitude—often by factors of 2-5× or more. This makes A_trans_peak (filled) > A_trans_peak (void) one of the most reliable single-feature discriminators.

18- Transmission Coefficient Estimate (T_est) expresses the ratio of transmitted to incident amplitude after normalizing for concrete attenuation along the propagation path. Theoretical calculations predict transmission through grout-filled ducts around 60-80% depending on interface roughness and grout quality, while transmission through air-filled voids drops to 0-10% as most energy reflects at the first interface. Our measurements align well with these predictions: T_est (filled) ≈ 0.6-0.8, T_est (void) ≈ 0.05-0.15, providing excellent discrimination with clear physical interpretation.

19- Shadow Zone Width (W_shadow) quantifies the spatial extent, measured in transducer channels, over which amplitude drops below a threshold behind the duct. Voids create broad acoustic shadows spanning 10-15 channels (approximately 20-30 cm) because almost no energy transmits through. Filled ducts create narrower shadows of 5-8 channels (10-15 cm) due to partial transmission allowing some energy to reach the far side. This spatial signature proves measurable through amplitude profiling across the receiver array and provides information about defect extent that complements the point-wise amplitude measurements.

B. Multi-Channel Coherence Features

Coherence analysis quantifies the similarity of signals received at adjacent transducers, exploiting MIRA's unique multi-channel capability in ways impossible for single-receiver systems.

20- Inter-Channel Coherence ($\gamma_{\text{coherence}}$) computes the complex coherence between adjacent receivers spaced approximately 5 mm apart. The mathematical definition involves the cross-spectrum:

$$\gamma(f) = |\langle S_1(f) \cdot S_2^*(f) \rangle| / \sqrt{(\langle |S_1(f)|^2 \rangle \cdot \langle |S_2(f)|^2 \rangle)}$$

where S_1 and S_2 represent the complex spectra from adjacent channels, the asterisk denotes complex conjugation, and the angle brackets indicate averaging over measurement repetitions or nearby frequency bins to improve stability.

The physical interpretation is elegant. Coherent plane waves—typical of transmission through uniform concrete or filled ducts—produce $\gamma \approx 0.9$ -1.0 because adjacent receivers detect essentially the same wavefront with only a small, deterministic phase shift. Scattered waves from voids or heterogeneities produce $\gamma < 0.6$ due to random phase variations between adjacent receivers as multiple scattered paths interfere. We consistently observe γ (filled) ≈ 0.85 -0.95 and γ (void) ≈ 0.3 -0.6, making this single feature the most discriminative in the MIRA arsenal. Feature importance analysis typically assigns 40-50% of the total discriminatory power to coherence alone.

21- Coherence Decay Length (L_{coh}) measures the distance over which inter-channel coherence decays to $1/e$ (about 37%) of its initial value. Long coherence lengths spanning many channels indicate smooth, homogeneous wavefronts propagating without significant scattering—exactly what we see with filled ducts. Short coherence lengths indicate strong scattering that decorrelates the signals over just a few channel spacings. This feature quantifies the spatial scale of wave disruption: L_{coh} (filled) typically exceeds 20 cm while L_{coh} (void) falls below 10 cm.

22- Phase Velocity Dispersion ($\Delta c/c$) examines variations in apparent shear wave velocity computed from phase shifts between adjacent channels across the frequency band. Uniform media exhibit constant phase velocity regardless of frequency—the hallmark of non-dispersive propagation. Voids create geometric dispersion where different frequency components travel at different apparent velocities due to complex diffraction and scattering around irregular boundaries. We observe $\Delta c/c$ (filled) $< 5\%$ indicating minimal dispersion, while $\Delta c/c$ (void) $> 15\%$ signals strong dispersive behavior.

C. Frequency-Domain Features

Spectral analysis of MIRA signals reveals resonance frequencies and bandwidth characteristics specific to ultrasonic propagation through concrete and embedded structures.

23- Center Frequency Shift (Δf_{center}) quantifies the difference between mean frequency of transmitted versus received spectra. Frequency-dependent attenuation—where higher frequencies attenuate faster than low frequencies—shifts the received spectrum toward lower frequencies relative to what was transmitted. Voids enhance this effect through scattering that preferentially disrupts high-frequency components. We typically measure $|\Delta f_{\text{center}}|$ (void) $> |\Delta f_{\text{center}}|$ (filled), with shifts of 10-20 kHz common for nominal frequencies around 50 kHz.

24- Spectral Width (σ_{freq}) measures the standard deviation of the frequency spectrum. Narrow spectra indicate selective frequency filtering or resonance effects. Broad spectra preserve waveform fidelity with minimal frequency-dependent distortion. We find σ_{freq} (filled) $> \sigma_{\text{freq}}$ (void) as scattering-induced narrowing reduces the effective bandwidth in void cases.

D. Spatial Pattern Features

These features capture how transmitted energy distributes spatially across the transducer array, revealing information about target geometry and scattering characteristics.

25- Focusing Quality Factor (Q_{focus}) expresses the ratio of energy concentration at the expected duct depth to the background energy level. Sharp focusing indicates a discrete reflector—typical of a void that creates a well-defined scattering center. Diffuse energy distribution suggests distributed scattering from grout inhomogeneities within a filled duct or multiple small scatterers. We observe $Q_{\text{focus}}(\text{void}) > Q_{\text{focus}}(\text{filled})$ as voids produce more localized scattering.

26- Spatial Energy Distribution Entropy (H_{spatial}) applies Shannon entropy to the normalized energy distribution across receiver channels: $H = -\sum p_i \cdot \log_2(p_i)$, where $p_i = E_i / \sum E_i$ represents the normalized energy at channel i . Low entropy indicates energy concentrated in a few channels—typical of the strong shadow zone behind a void. High entropy indicates energy spread uniformly across many channels—characteristic of filled ducts that allow broader transmission. Therefore $H_{\text{spatial}}(\text{void}) < H_{\text{spatial}}(\text{filled})$.

E. Time-Domain Features

Temporal characteristics of received waveforms provide complementary information to frequency-domain analysis, capturing transient behavior and propagation delays.

27- Time-of-Flight (ToF) measures the time delay from transmission to the first threshold crossing at the receiver. This provides a direct measure of wave velocity and path length. For filled ducts, waves travel through grout with velocity around 2500 m/s—somewhat slower than concrete's 3000-3500 m/s but still allowing reasonably direct transmission. For voids, waves must diffract around the duct edges following longer paths at concrete velocity, resulting in increased ToF. We typically observe $\text{ToF}(\text{void}) > \text{ToF}(\text{filled})$ by 1-3 microseconds for duct sizes in the 5-10 cm range.

28- Signal Duration (τ_{signal}) quantifies the time duration over which received amplitude exceeds 10% of its peak value. Transmitted pulses through uniform media preserve short duration matching the excitation pulse. Scattered and multiply-reflected signals exhibit extended duration as energy arrives via multiple paths with different delays. This leads to $\tau_{\text{signal}}(\text{void}) > \tau_{\text{signal}}(\text{filled})$ due to the complex multi-path scattering environment created by voids.

Summary of MIRA Feature Set

We extract approximately 25-30 features from each MIRA measurement location, with heavy emphasis on coherence and multi-channel characteristics unique to this modality. The high density of spatial sampling—48 channels spanning just 20 cm—enables robust coherence estimation even in noisy conditions where individual channel measurements might be unreliable. This makes MIRA features particularly trustworthy for void detection. Subsequent feature importance analysis, detailed in Section 7.4.5, consistently identifies inter-channel coherence ($\gamma_{\text{coherence}}$) as the dominant discriminator, typically accounting for 40-50% of the total importance. This validates our physics-based intuition that spatial coherence provides the most direct and reliable signature of void-induced wave disruption.

7.4.4 Impact Echo (IE) acoustic features

7.4.4.1 IE fundamentals

IE testing operates on fundamentally different physical principles than either GPR or MIRA. A mechanical impact—delivered by a spring-loaded impactor or instrumented hammer—generates stress waves that propagate through the concrete. An accelerometer or microphone positioned near the impact point records the resulting vibrations. The received signal encodes information about thickness, internal voids, delaminations, and acoustic impedance contrasts through its frequency spectrum rather than amplitude or arrival time alone.

Unlike GPR and MIRA which primarily measure transmitted or reflected energy propagating away from the source, IE exploits resonance phenomena. Stress waves reflect from internal interfaces and free surfaces, setting up standing-wave patterns at characteristic frequencies determined by the geometry and material properties. For a solid concrete element of thickness T containing a void or delamination at depth d , basic wave theory predicts a fundamental resonance frequency: $f_{\text{resonance}} = v / (2d)$, where $v \approx 4000$ m/s represents the compression wave velocity in concrete. A surface-breaking delamination at 10 cm depth produces resonance around 20 kHz, while the bottom surface of a 40 cm thick slab resonates at roughly 5 kHz.

Voids within tendon ducts perturb these resonances in characteristic ways. They introduce additional reflection interfaces beyond the intended concrete boundaries. They alter local stiffness, changing the effective acoustic impedance and shifting resonance frequencies. In complex geometries, they create mode splitting where a single clean resonance divides into multiple peaks as different portions of the structure vibrate at slightly different frequencies.

The raw data consists of a time-domain acceleration or pressure signal, typically spanning 1-2 milliseconds and sampled at 100-500 kHz—substantial oversampling relative to the 1-50 kHz frequency content of interest, but necessary to capture transient behavior accurately. Feature extraction operates primarily in the frequency domain after FFT transformation, supplemented by time-domain characteristics and spatial context features unique to IE's grid-based sampling.

7.4.4.2 Feature extraction from a single measurement point

A. Spectral Peak Characteristics

The most information-rich features describe the properties of peaks in the amplitude spectrum, which directly correspond to resonance modes of the structure.

29- Dominant Frequency (f_{dom}) identifies the frequency of maximum amplitude in the relevant 1-50 kHz range. This represents the primary resonance mode and should correspond to thickness resonance ($f \approx v/2T$) for intact, homogeneous concrete. The presence of voids typically introduces additional resonances at higher frequencies as the effective thickness reduces in the voided region or as the void itself acts as a resonator. Intact filled ducts show a single dominant peak near the thickness frequency, while voided ducts often exhibit split peaks or additional harmonics.

30- Peak Frequency Shift (Δf_{shift}) measures the difference between the observed dominant frequency and the theoretical thickness frequency calculated from known slab geometry. Significant shifts indicate anomalous conditions. Lower observed frequencies suggest delamination closer to the surface than the bottom boundary, reducing the effective resonating thickness. Higher frequencies suggest internal voids reducing the effective acoustic path length. For typical duct depths of 10-15 cm, we observe $|\Delta f_{\text{shift}}|$ (void) $>$ $|\Delta f_{\text{shift}}|$ (filled), often by 1-3 kHz—a substantial shift relative to the base frequencies involved.

31- Peak Amplitude ($A_{\text{peak_freq}}$) captures the maximum spectral amplitude at the dominant frequency. Strong resonances producing high peak amplitudes indicate efficient energy trapping in high-Q resonant modes. Weak resonances suggest damping or energy dissipation that prevents clear resonance development. The relationship between filled and void conditions is complex and material-dependent. Filled ducts often show lower peaks due to grout damping, but irregular void geometries can also produce weak, poorly-defined resonances.

32- Peak Width ($\text{FWHM}_{\text{freq}}$) defines the frequency bandwidth at half-maximum amplitude around the dominant peak. Narrow peaks indicate high-Q resonance with low damping—characteristic of sharp, well-defined interfaces. Broad peaks suggest heavy damping or multiple overlapping modes that blur together. We generally observe $\text{FWHM}(\text{filled}) > \text{FWHM}(\text{void})$ as grout damping broadens the resonance, though exceptions occur with certain void geometries.

33- Peak Prominence (P_{prom}) quantifies the amplitude difference between the peak and the surrounding baseline level. High prominence indicates an isolated, well-defined resonance standing clearly above the noise floor. Low prominence suggests either a noisy spectrum or weak resonance that barely rises above background. When voids create sharp impedance contrasts, they often produce high prominence peaks: $P_{\text{prom}}(\text{void}) > P_{\text{prom}}(\text{filled})$.

B. Multi-Peak Features

Voids frequently manifest as multiple spectral peaks rather than simple frequency shifts, motivating features that capture overall spectral complexity.

34- Number of Significant Peaks (N_{peaks}) counts peaks exceeding an amplitude threshold, typically set at 30% of the global maximum to avoid counting noise bumps as genuine peaks. Simple intact structures produce 1-2 peaks representing the fundamental resonance and perhaps its first harmonic. Complex void geometries produce 3-5 or more peaks from multiple reflection paths and coupled resonances in different portions of the structure. We consistently observe $N_{\text{peaks}}(\text{void}) > N_{\text{peaks}}(\text{filled})$, with typical values of 3-4 versus 1-2.

35- Peak Spacing Regularity (σ_{spacing}) computes the standard deviation of frequency intervals between consecutive peaks. Regular harmonic spacing—where peaks appear at f , $2f$, $3f$, etc.—produces small σ values and indicates a simple resonator. Irregular spacing suggests complex multi-layer structure with inharmonic modes. Generally $\sigma_{\text{spacing}}(\text{filled}) < \sigma_{\text{spacing}}(\text{void})$ as voids introduce irregular additional modes.

36- Harmonic-to-Fundamental Ratio (HFR) expresses the ratio of second harmonic amplitude to fundamental amplitude. Strong harmonics indicate non-linear effects or complex mode shapes where higher-order modes carry substantial energy. Weak harmonics suggest strong fundamental resonance dominating the spectrum. HFR (void) often increases due to mode coupling induced by the geometric complexity of partially voided regions.

C. Spectral Shape Features

Beyond analyzing discrete peaks, the overall spectral envelope provides diagnostic information about the distributed character of the response.

37- Spectral Centroid (f_{centroid}) computes the center of mass of the amplitude spectrum, identical in concept to the GPR feature. The centroid reflects the overall balance between low and high frequency content. When voids introduce high-frequency resonances, they shift the centroid upward: $f_{\text{centroid}}(\text{void}) > f_{\text{centroid}}(\text{filled})$.

38- Spectral Spread (σ_{spectrum}) measures the standard deviation of the spectrum about its centroid. Narrow spread indicates energy concentrated near a single frequency—typical of strong, isolated resonance. Wide spread indicates energy distributed across many frequencies. We observe $\sigma_{\text{spectrum}}(\text{void}) > \sigma_{\text{spectrum}}(\text{filled})$ as multiple void-induced resonance modes spread energy across the spectrum.

39- Spectral Flatness (SF) computes the ratio of geometric mean to arithmetic mean of spectral amplitudes. $SF \approx 0$ indicates a tone-like signal with sharp resonance peaks, while $SF \approx 1$ indicates noise-like signals with flat, broadband spectra. Damping in filled ducts smooths spectral peaks, increasing flatness: $SF(\text{filled}) > SF(\text{void})$.

40- High-Frequency Energy Ratio (HFER) quantifies the ratio of energy above a threshold frequency—say 15 kHz—to total energy. High-frequency modes are preferentially excited by small-scale heterogeneities like void edges and boundaries. $HFER(\text{void}) > HFER(\text{filled})$ as void-related features pump energy into high-frequency components.

D. Time-Domain Features

While frequency analysis dominates IE interpretation, time-domain features capture transient behaviour and damping characteristics.

41- Signal Decay Rate ($\alpha_{\text{decay_IE}}$) fits an exponential decay constant to the amplitude envelope: $A(t) \propto \exp(-\alpha t)$. Fast decay indicates strong damping from material absorption. Slow decay indicates high-Q resonance where energy persists for many cycles. Grout provides significant damping, so $\alpha_{\text{decay}}(\text{filled}) > \alpha_{\text{decay}}(\text{void})$ —filled ducts show faster signal decay.

42- Zero-Crossing Rate (ZCR) counts how many times the signal crosses zero amplitude per unit time. High ZCR indicates high-frequency content dominating the signal. Low ZCR indicates low frequencies dominate. Consistent with the spectral features, $ZCR(\text{void}) > ZCR(\text{filled})$ reflecting higher frequency content in void signatures.

43- Kurtosis (K_{time}) measures the fourth statistical moment of the amplitude distribution, quantifying how "peaky" or "heavy-tailed" the distribution appears. High kurtosis indicates spiky, impulsive signals with sharp transients. Low kurtosis indicates Gaussian-like random signals. Sharp reflections from well-defined void boundaries might elevate kurtosis: $K_{\text{time}}(\text{void})$ potentially $> K_{\text{time}}(\text{filled})$, though this feature shows more variability than others.

7.4.4.3 Spatial context features (Unique to IE)

IE measurements are collected on a discrete grid with typical spacing of 10-20 cm, unlike the continuous scanning of GPR or the linear array sampling of MIRA. This grid structure enables extraction of spatial context features that characterize defect extent and continuity—critical information for distinguishing isolated measurement artifacts from genuine structural voids.

E. Neighborhood Consistency Features

These features compare each measurement point with its immediate spatial neighbors, typically the 4-8 surrounding grid points in the measurement array.

44- Dominant Frequency Consistency (C_{freq}) computes the standard deviation of dominant frequencies at the current point and its 8 nearest neighbors. Low standard deviation indicates spatially consistent structure—what we expect for a uniform filled duct extending through the measurement region. High standard deviation suggests a transition zone between different conditions or irregular void boundaries where geometry changes significantly over short distances. We observe $C_{\text{freq}}(\text{continuous void}) < C_{\text{freq}}(\text{filled-to-void transition})$ as uniform voids produce consistent frequency shifts across multiple adjacent points.

45- Amplitude Correlation with Neighbors (ρ_{neighbor}) calculates the mean Pearson correlation coefficient between the current point's full spectrum and the spectra from neighboring points. High correlation indicates similar spectral shapes—typical of uniform conditions where all nearby points respond similarly. Low correlation suggests the current point is anomalous relative to its surroundings, or that there's significant spatial heterogeneity. We expect $\rho_{\text{neighbor}} \approx 0.8-0.9$ for uniform regions and $\rho_{\text{neighbor}} < 0.5$ for genuine anomalies that stand out from their surroundings.

46- Spatial Gradient Magnitude ($|\nabla f_{\text{dom}}|$) measures how rapidly the dominant frequency changes in the x-y measurement plane:

$$|\nabla f| = \sqrt{(\partial f / \partial x)^2 + (\partial f / \partial y)^2}$$

Large gradients mark boundaries between filled and void regions where acoustic conditions change abruptly. Small gradients indicate uniform conditions extending over substantial areas. We expect $|\nabla f_{\text{dom}}|$ maximized at void boundaries and minimized within uniform regions, providing a way to identify transition zones.

F. Clustering and Continuity Features

These features assess whether an apparent anomaly appears isolated—suggesting measurement artifact—or as part of a spatially continuous defect that's much more likely to be genuine.

47- Anomaly Cluster Size (N_{cluster}) counts the number of contiguous grid points exhibiting similar anomalous spectral signatures. A single isolated anomalous point likely represents measurement error, ambient vibration interference, or an operator mistake during that particular impact. Clusters of 3 or more contiguous points with similar signatures likely represent genuine voids, which typically extend over distances greater than 10 cm due to grouting flow physics—air pockets migrate to high points, creating extended voids rather than isolated point defects. We flag single anomalous points for verification while classifying clusters of 4+ points as high-confidence voids.

48- Spatial Autocorrelation (Moran's I) applies a formal statistical measure of spatial clustering:

$$I = (N/W) \cdot \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x}) / \sum_i (x_i - \bar{x})^2$$

Where w_{ij} represents spatial weights (1 for neighbors, 0 otherwise), N is the number of measurement points, and W is the sum of weights. The statistic I quantifies whether similar values cluster spatially. $I > 0$ indicates positive spatial autocorrelation where high values cluster near other high values and low near low. $I \approx 0$ indicates random spatial distribution with no pattern. $I < 0$ indicates negative correlation producing a checkerboard pattern—physically unrealistic for structural defects. We observe I (genuine void region) > 0.5 indicating strong clustering, while I (random noise) ≈ 0 .

49- Directional Anisotropy (A_{dir}) computes the ratio of spatial correlation along versus perpendicular to the expected duct axis. Voids in linear tendon ducts should exhibit anisotropic signatures—stronger correlation along the duct's longitudinal axis than across it, reflecting the duct's elongated geometry. We expect A_{dir} (void in duct) > 2 , while random scatter produces $A_{\text{dir}} \approx 1$ showing no directional preference.

7.4.4.4 Summary of IE Feature Set

We extract approximately 30-35 features per measurement point, combining single-point spectral and temporal characteristics with spatial context derived from surrounding measurements. The spatial features prove particularly critical for IE due to its inherently point-based sampling pattern. These features substantially reduce false-positive rates—potentially by 50% or more—by filtering isolated anomalies that don't exhibit the spatial continuity we expect from genuine voids. A suspicious frequency shift at a single grid point might indicate a void, or it might just represent an unusually hard or soft impact, ambient vibration from passing traffic, or an operator positioning error. That same frequency shift appearing consistently across 4-5 adjacent points, with appropriate directional anisotropy along the expected duct axis, almost certainly indicates a genuine void.

7.4.5 Feature selection and dimensionality reduction

The comprehensive feature extraction procedures generate 50-80 candidate features per measurement point when considering all three modalities. While we've justified each feature based on physical principles, practical considerations demand further refinement. Many features exhibit strong inter-correlation—for instance, peak amplitude, integrated energy, and reflection coefficient all essentially measure "signal strength" and therefore correlate highly with each other. Additionally, some features that seem reasonable in theory provide minimal discriminatory power for the specific task of distinguishing filled from void conditions in our particular dataset and structural types.

Including all features creates several problems. First is overfitting risk: with hundreds of training samples but dozens of features, the model might memorize training-set idiosyncrasies rather than learning genuine patterns that generalize to new structures. Second is increased computational cost: evaluating 80 features per prediction takes longer than evaluating 15 carefully selected features, which matters for real-time field deployment. Third is reduced interpretability: a model relying on 80 features becomes difficult to understand and validate, while one using 15 key features remains transparent.

Feature selection addresses these issues by identifying the minimal subset that maximizes classification performance. We employ two complementary approaches that work together effectively.

7.4.5.1 Feature Importance Analysis (Random Forest)

Random Forest models automatically compute feature importance as a natural byproduct of training, requiring no additional computational effort. Each time a decision tree makes a split—deciding how to divide the training data based on one feature—it records the impurity reduction achieved by that split. Impurity, typically measured by the Gini index, quantifies how mixed the classes are at a node. Pure nodes contain only one class (all filled or all void); mixed nodes contain both. Features that consistently produce large impurity reductions across many trees and split points clearly contribute strongly to classification accuracy and receive high importance scores.

Mathematically, the importance of feature f aggregates contributions across all trees:

$$I(f) = \sum_{t=1}^T \sum_{n \in N_t(f)} \Delta \text{Gini}(n) / T$$

where T represents the number of trees in the forest, $N_t(f)$ denotes the set of nodes in tree t that split on feature f , and $\Delta \text{Gini}(n)$ quantifies the Gini impurity reduction at node n . This formulation automatically accounts for both how frequently each feature is used and how much each use improves classification.

Our procedure follows a systematic protocol. We train an initial Random Forest using all candidate features on the complete training set without any feature restriction. We then extract the feature importance scores and normalize them to sum to 1.0, making them interpretable as the fraction of total discriminatory power each feature provides. Sorting features by importance in descending order and plotting the cumulative importance curve reveals how many features we need to capture most of the information. We typically select features accounting for 95% of cumulative importance, discarding the remaining 5% as either redundant or noise.

Finally, we retrain the model using only this selected subset and validate performance on the held-out test set. If performance remains essentially unchanged—within 1-2% of the full-feature model—we've successfully reduced dimensionality without sacrificing accuracy. If performance drops substantially, we relax the threshold to 90% or 92% cumulative importance to retain more features.

Typical results demonstrate substantial dimensionality reduction across all modalities. For GPR data, our initial 16 features reduce to approximately 8-10 features accounting for 95% of importance. For MIRA data, 30 features reduce to roughly 12-15 features, with coherence-based features absolutely dominating—often a single coherence feature contributes 40-50% of total importance. For IE, 35 features reduce to about 15-18 features, with spectral characteristics and spatial features contributing roughly equally.

Beyond numerical reduction, feature importance analysis reveals which physical mechanisms dominate void detection for each modality. For GPR, amplitude-based features measuring reflection strength and spectral centroid quantifying frequency content shifts provide most of the discriminatory power. For MIRA, inter-channel coherence overwhelmingly dominates—this single feature or its close variants typically account for 40-50% of total importance, validating our emphasis on coherence in the feature engineering. For IE, spectral characteristics like peak positions and counts balance roughly evenly with spatial consistency features, confirming that both single-point spectral signatures and spatial context contribute essential information.

This physics-based interpretation validates our feature engineering approach and guides future refinement. If we wanted to further simplify the MIRA system for faster field deployment, we know we could focus exclusively on coherence computation and potentially achieve 80-90% of current performance with perhaps just 5 carefully selected coherence-related features. Conversely, if we wanted to improve IE performance, we know that neither spectral nor spatial features dominate, so improvements require enhancing both categories equally.

7.4.5.2 *Correlation Analysis*

Highly correlated features provide largely redundant information and can cause numerical instability in certain ML algorithms, particularly linear models and Support Vector Machines. Correlation analysis identifies this redundancy, allowing us to eliminate it while preserving maximal information content.

The procedure is straightforward. We compute the Pearson correlation matrix R for all feature pairs:

$$R(i, j) = \text{cov}(f_i, f_j) / (\sigma_i \cdot \sigma_j)$$

Where ‘cov’ denotes covariance and σ represents standard deviation. This matrix quantifies linear relationships between every pair of features. We then identify feature pairs with $|R| > 0.90$, indicating very strong correlation—these features move together so consistently that they're essentially measuring the same underlying phenomenon.

For each highly correlated pair, we face a choice: which feature to keep? We retain the feature with higher Random Forest importance score (from the previous analysis) and discard the other. This ensures we keep the more discriminative of the two correlated features. We iterate this process, recomputing correlations after each removal, until no pairs exceed the correlation threshold. Iteration matters because removing feature A might change the correlation between features B and C that weren't previously flagged.

Examining the correlated pairs we identify reveals interesting physics. For GPR data, peak amplitude and integrated energy correlate at $R = 0.94$ —both essentially measure "how much electromagnetic energy reflected back." We retain peak amplitude because it shows slightly higher importance, likely because it's less affected by pulse shape variations. For MIRA data, we often extract coherence measures at different frequency bands (low-frequency coherence, high-frequency coherence, broadband coherence). These correlate at $R = 0.88-0.92$ because concrete's scattering properties don't change dramatically with frequency in the 40-60 kHz range MIRA uses. We retain the broadband coherence as most robust. For IE, spectral centroid and high-frequency energy ratio correlate at $R = 0.86$ —both capture frequency content shifts toward higher frequencies in void conditions. We keep spectral centroid as more stable and interpretable.

The final feature sets after both importance-based selection and correlation pruning achieve excellent balance. For GPR, we retain 8 features mixing amplitude characteristics (peak amplitude, reflection coefficient), temporal properties (two-way travel time, pulse width), and spectral content (spectral centroid, high-frequency ratio). For MIRA, we keep 12 features strongly dominated by coherence-related measures (inter-channel coherence, coherence decay length) supplemented by amplitude features (transmission amplitude, shadow width). For IE, we retain 15 features balancing

spectral characteristics (dominant frequency, peak count, spectral spread) with spatial context (frequency consistency, cluster size, spatial autocorrelation).

These reduced feature sets drive all subsequent model training and evaluation. They balance information richness—retaining 95% of the discriminatory power present in the full feature set—with parsimony that prevents overfitting, reduces computation time, and maintains interpretability. An inspector can understand a model relying on 8-15 well-chosen features much more readily than one involving 50-80 measurements, which facilitates trust, validation, and regulatory acceptance.

7.5 Data Augmentation

Machine learning models benefit tremendously from seeing diverse training examples that capture the natural variability in real measurements. Data augmentation addresses this by artificially expanding the training set through physically-plausible transformations applied to existing measurements. The key word here is "physically-plausible"—we're not just randomly perturbing data, but rather simulating variations that could genuinely occur under slightly different measurement conditions. Done correctly, augmentation helps models generalize better to new structures and conditions. Done incorrectly by introducing unrealistic artifacts, it can actually degrade performance by teaching the model to respond to patterns that don't exist in nature.

7.5.1 GPR augmentation

A. Additive Gaussian Noise

This simulates the electronic noise and random environmental interference present in all real measurements. We add zero-mean Gaussian noise to raw A-scans:

$$A_{aug}(t) = A(t) + N(0, \sigma^2)$$

Where the noise level σ ranges from 0.05 to 0.15 times the signal's RMS amplitude. This represents realistic noise sources like thermal noise in the receiver electronics and electromagnetic interference from nearby equipment—power lines, electrical machinery, radio transmissions. Every real measurement contains some noise; augmentation ensures the model learns to function despite it rather than requiring unrealistically clean signals.

B. Time Shifting

Small variations in antenna height above the concrete surface or coupling conditions affect the time-zero calibration—the reference point from which we measure arrival times. We simulate this by shifting A-scans by ± 0.5 to 2 time samples, corresponding to roughly ± 0.1 to 0.5 nanoseconds. The critical constraint: we don't shift across target boundaries where such shifts would create physically impossible signals. This augmentation represents operator-induced coupling variations as the antenna bounces slightly during vehicle movement, and surface roughness that changes the effective antenna height by a few millimeters from point to point.

C. Amplitude Scaling

Antenna gain can vary slightly during long surveys due to temperature changes, battery voltage fluctuations, or minor hardware drift. Additionally, concrete moisture content variations—the difference between dry concrete and saturated concrete after rainfall—substantially affect dielectric properties and signal amplitudes. We simulate these effects by multiplying A-scans by scaling factors α ranging from 0.8 to 1.2, representing $\pm 20\%$ amplitude variation. This is conservative relative to the actual variations we observe in field conditions, but aggressive augmentation can introduce unrealistic extremes.

D. Trace Interpolation (Spatial Oversampling)

Continuous antenna movement generates smooth spatial variation in the signals. We create augmented traces through linear or spline interpolation between adjacent measured A-scans, but only within uniform regions—we deliberately avoid interpolating across target boundaries where the signal changes abruptly. This represents what we would have measured if we'd sampled at finer spatial intervals, generating additional training examples without requiring additional field data collection.

These four augmentation techniques together increase the effective GPR training set by 5-7×. Each original measurement generates 4-6 augmented versions, providing the model with exposure to natural measurement variability without requiring proportionally more field time and annotation effort.

7.5.2 UPE augmentation

A. Channel Dropout

In field conditions, some of the 48 transducers occasionally fail to make good contact with the concrete surface due to surface contamination, transducer malfunctions, or particularly dry, porous concrete that doesn't couple ultrasound well. We simulate this by randomly zeroing out 5-10% of channels in each measurement. This forces the model to function even when several channels provide no useful data—a common real-world scenario. The model learns to rely on spatial patterns across multiple channels rather than depending critically on any single channel, improving robustness to partial data loss.

B. Phase Noise Injection

Electronic systems aren't perfect. Clock jitter and temperature-induced phase drift in the electronics create small random phase variations in the complex I/Q data. We simulate this by adding uniform random phase:

$$S_{aug}(f) = S(f) \cdot \exp(j\theta)$$

Where θ ranges from $-\pi/10$ to $+\pi/10$ radians. This represents the kind of electronic noise that doesn't affect amplitude measurements much but can impact phase-sensitive coherence calculations if the model becomes too dependent on exact phase values.

C. Coherence Perturbation

Concrete isn't perfectly homogeneous. Aggregate scattering and local microcracking introduce small, random perturbations to wave propagation that slightly reduce coherence even in filled ducts. We simulate this by adding small decorrelating noise to subsets of channels while carefully preserving the overall coherence structure—we want to add realistic variability, not destroy the fundamental filled-versus-void discrimination. This teaches the model that "filled" doesn't mean perfect coherence of 1.0, but rather high coherence around 0.85-0.95 with some natural variability.

D. Transmission Amplitude Scaling

Concrete properties vary significantly across structures and even within single structures. Age affects ultrasonic attenuation as hydration continues for months or years. Moisture content matters enormously—saturated concrete attenuates more than dry concrete. Mix design variations change both velocity and attenuation. We simulate these effects by scaling all channel amplitudes by factors ranging from 0.7 to 1.3, representing the $\pm 30\%$ variations commonly observed across our test structures.

The combination of these augmentation techniques, particularly channel dropout, makes the model significantly more robust to incomplete or partially corrupted measurements—exactly what we encounter regularly in field inspections. The MIRA training set expands by 4-5× through augmentation.

7.5.3 Impact Echo (IE) augmentation

A. Spectral Noise Injection

IE measurements exhibit substantial point-to-point variability even in uniform concrete due to ambient vibrations, wind-induced structural oscillations, and unavoidable variations in impactor delivery. We simulate this by adding noise to the frequency spectrum:

$$F_{aug}(f) = F(f) + N(0, \sigma^2(f))$$

Where the noise level $\sigma(f)$ is proportional to $|F(f)|$. This means strong spectral features get proportionally more noise—higher signal-to-noise ratio—while weak features get less absolute noise, maintaining realistic SNR characteristics across the spectrum. Physical sources include ambient vibrations from traffic, wind buffeting the structure, and natural variability in how the impactor strikes the surface despite the operator's best efforts at consistency.

B. Peak Frequency Jitter

Concrete velocity varies by ± 50 -100 m/s in field conditions due to temperature changes (concrete is about 0.05% slower per degree Celsius), moisture content variations, and local material heterogeneities. These velocity changes shift resonance frequencies. We simulate this by shifting the dominant spectral peak by ± 2 -5% in frequency—for a 10 kHz resonance, that's ± 200 -500 Hz variation. This represents the natural measurement-to-measurement variability we observe when testing the same location multiple times under slightly different conditions.

C. Baseline Drift Addition

Low-frequency noise from traffic vibrations, seismic background, and resonances in the measurement rig itself often appears as baseline drift in IE spectra. We simulate this by adding low-frequency sinusoidal or polynomial baselines to spectra, concentrated below 2 kHz where it doesn't directly interfere with the typical 5-20 kHz resonance frequencies we're analyzing. This teaches the model to focus on the relevant mid-frequency resonances rather than being distracted by low-frequency contamination.

D. Spatial Perturbation

Despite careful grid layout, actual measurement positions carry uncertainty—the operator might miss the intended grid point by a centimeter or two, the grid itself might not be perfectly square, or the marked positions might shift slightly during the hours-long measurement campaign. We simulate this by randomly perturbing grid coordinates by ± 2 cm, which falls within typical measurement repeatability. This perturbation alters spatial context features like neighbor relationships slightly, ensuring the model doesn't become overly dependent on perfectly regular grid geometry.

IE augmentation is deliberately more conservative than GPR or MIRA, expanding the training set by only 3 - $4\times$ rather than 5 - $7\times$. The reason: IE measurements are more sensitive to genuine spatial variations. Aggressive augmentation risks teaching the model to ignore real spatial patterns that actually carry diagnostic information about void boundaries and extent.

Augmentation Quality Control

After applying all these augmentation techniques, we don't just assume everything worked correctly. We randomly sample about 10% of the augmented examples and have domain experts review them to verify physical plausibility. We're watching for several failure modes: unphysical spectral shapes that couldn't arise from real concrete (negative frequencies, impossibly sharp discontinuities), violations of spatial continuity where neighboring augmented points look wildly different despite originating from the same local region, or amplitude values outside reasonable ranges. Augmentations that introduce these unrealistic artifacts get adjusted—we might reduce the noise level or scaling factor—or removed entirely if they can't be fixed. This quality control loop ensures augmentation helps rather than hurts model training by maintaining physical realism throughout the augmented dataset.

8 Model training and hyperparameter tuning

Once the machine learning models are designed and the features are extracted, the next critical steps involve actually training those models, testing how well they work, and integrating them into a practical system that inspectors can use in the field. This section walks through that entire process from start to finish.

Training begins with carefully dividing the data into separate groups—some for teaching the models, some for fine-tuning their settings, and some held completely aside for honest testing at the end. Each model type has dozens of adjustable parameters that need optimization through systematic testing to find the sweet spot between learning too little and memorizing too much. The section explains exactly how these parameters were selected and what trade-offs were considered.

Performance evaluation introduces the metrics used to judge success—not just simple accuracy percentages, but more sophisticated measures that account for the real-world consequences of different types of mistakes. Missing a dangerous void is far worse than triggering an unnecessary follow-up inspection, and the evaluation framework reflects these priorities.

The results reveal that Random Forest models consistently outperformed other approaches across all three NDT methods, achieving detection rates around 87-90%. The final sections describe how these trained models were integrated into a unified system with 3D visualization that combines information from radar, ultrasonic, and impact-echo measurements, color-coding duct segments by confidence level and providing inspectors with all relevant data at their fingertips for making informed decisions.

Proper dataset splitting is essential to obtain unbiased performance estimates. The dataset is divided into three non-overlapping subsets:

Training Set (60%): Used for model learning (fitting weights, parameters) **Validation Set (20%):** Used for hyperparameter tuning and early stopping **Test Set (20%):** Held-out for final performance evaluation; never seen during training/tuning

Splitting Strategy: Stratified splitting ensures class balance is preserved across splits: if the full dataset contains 18% voids, each split maintains approximately 18% voids. Additionally, spatial splitting is applied where possible: for measurements from multiple structures, entire structures are assigned to splits rather than randomly mixing measurements from the same structure across splits. This prevents data leakage where the model learns structure-specific patterns during training and encounters the same structure (but different measurement points) in the test set, artificially inflating performance estimates.

8.1 Training configuration and hyperparameter tuning

Each model type requires careful hyperparameter selection to navigate the fundamental trade-off in ML tuning: underfitting versus overfitting. Models with insufficient capacity underfit, failing to capture genuine patterns in the data and performing poorly even on training examples. Models with excessive capacity overfit, memorizing training noise and peculiarities that don't generalize to new data. The hyperparameter tuning process searches for the sweet spot that maximizes performance on unseen data.

8.1.1.1 Support Vector Machine (SVM)

SVM hyperparameters optimized through grid search combined with 5-fold stratified cross-validation on the training set. This means we systematically tried different combinations of hyperparameter values, evaluating each combination's performance across five different train/validation splits to get stable estimates.

Table 8-1 SVM Hyperparameter Selection

Hyperparameter	Values Tested	Selected (GPR)	Selected (MIRA)	Selected (IE)
C (regularization)	[0.1, 1, 10, 100]	10	1	10
γ (RBF kernel width)	[0.001, 0.01, 0.1, 1]	0.01	0.1	0.01
class_weight	["balanced"]	balanced	balanced	balanced

Selection criterion: Maximize F1-score on validation fold (prioritizes balanced precision and recall rather than raw accuracy which is misleading for imbalanced data).

8.1.1.2 Random Forest (RF)

Random Forest hyperparameter optimization follows a similar grid search approach but explores a different parameter space reflecting RF's ensemble structure.

Table 8-2 Random Forest Hyperparameter Selection

Hyperparameter	Values Tested	Selected (GPR)	Selected (UPE)	Selected (IE)
n_estimators	[100, 200, 300]	200	300	200
max_depth	[10, 20, 30, None]	20	None	30
min_samples_split	[2, 5, 10]	5	2	5
min_samples_leaf	[1, 2, 4]	2	1	2
max_features	['sqrt', 'log2']	sqrt	sqrt	sqrt
class_weight	["balanced_subsample"]	balanced_subsample	balanced_subsample	balanced_subsample

Selection criterion: Maximize F1-score with preference for higher n_estimators (lower variance) when F1 scores within 0.02.

8.1.1.3 1D-Convolutional Neural Network (1D-CNN)

For the 1D-CNN applied to IE spectra, Bayesian optimization employed rather than exhaustive grid search. Deep learning hyperparameter spaces are larger and more expensive to evaluate—each configuration requires training a neural network for potentially hundreds of epochs—so the smarter, sequential search strategy of Bayesian optimization proves more efficient than blindly trying all combinations.

Table 8-3 1D-CNN hyperparameter selection (for Impact Echo)

Hyperparameter	Range Tested	Selected
learning_rate	[0.0001, 0.01]	0.001
batch_size	[8, 16, 32]	16
dropout_rate	[0.3, 0.5, 0.7]	0.5
L2_regularization	[0.001, 0.01, 0.1]	0.01
n_filters_layer1	[16, 32, 64]	32
kernel_size_layer1	[5, 7, 9]	7

The training procedure incorporates several best practices for small-dataset deep learning. The Adam optimizer is used with standard parameters ($\beta_1=0.9$, $\beta_2=0.999$, $\epsilon=10^{-8}$) for adaptive learning rate adjustment. The loss function is binary cross-entropy with class weights to handle the imbalance between filled and void examples. Early stopping monitors validation loss with patience of 15 epochs—if validation performance doesn't improve for 15 consecutive epochs, training terminates to prevent overfitting. We also implement a learning rate schedule that reduces the rate by a factor of 0.5 if validation loss plateaus for 5 epochs, helping the optimization escape local minima. Training runs for a maximum of 100 epochs but typically stops early around epoch 60-80 when validation performance plateaus.

8.1.1.4 U-Net (GPR Segmentation Baseline)

The U-Net configuration for GPR radargram segmentation uses a more conservative learning rate and longer patience reflecting the more complex architecture and larger effective parameter count.

Table 8-4 U-Net hyperparameter configuration

Hyperparameter	Value
learning_rate	0.0001
batch_size	8
dropout_rate	0.3
Loss function	Dice coefficient
Optimizer	Adam
Early stopping patience	25 epochs
Epochs	150 (typically stops ~100)

8.1.2 Cross-Validation strategy

Beyond the single train/validation/test split used for final evaluation, 5-fold stratified cross-validation is also performed on the combined training+validation set. This provides two valuable pieces of information: first, confidence intervals on performance metrics that indicate how much variability to expect, and second, an overfitting diagnostic—large variance across folds indicates the model is memorizing fold-specific patterns rather than learning robust generalizations.

The procedure works as follows:

1. Combine training and validation sets
2. Split into 5 folds (stratified by class label)
3. For each fold $k=1\dots 5$:
 - o Train on 4 folds
 - o Evaluate on remaining fold
 - o Record performance metrics
4. Compute mean and standard deviation of metrics across folds
5. Final model training uses all 5 folds' data with hyperparameters selected from grid search

Results Reporting: Performance metrics reported as: **mean $\pm$ std** (e.g., $F1 = 0.87 \pm 0.04$)

High standard deviation (>0.10) indicates model instability or insufficient data; low standard deviation (<0.05) indicates robust generalization.

8.1.3 Performance evaluation metrics

8.1.3.1 Understanding the confusion matrix

The confusion matrix provides the foundation for all subsequent performance metrics by tabulating the four possible outcomes in binary classification.

	Predicted Filled	Predicted Void
Actual Filled	<i>TN (True Negative)</i>	<i>FP (False Positive)</i>
Actual Void	<i>FN (False Negative)</i>	<i>TP (True Positive)</i>

Each cell has important practical implications for NDT inspection. True Negatives (TN) represent correctly identified filled ducts—the model appropriately decides no action is needed, avoiding unnecessary follow-up investigations. False Positives (FP) are filled ducts misclassified as voids, triggering unnecessary follow-up with targeted MIRA or IE investigation.

This wastes time and money but doesn't compromise safety. False Negatives (FN) are voids misclassified as filled—critical safety errors where the model misses actual defects. These represent the highest-risk failure mode because inspectors move on believing the duct is sound when it actually contains voids that could lead to corrosion and structural failure. True Positives (TP) are correctly detected voids, representing successful defect identification that justifies the entire inspection effort.

A concrete example from MIRA data analysed with the Random Forest model illustrates these concepts.

	Predicted Filled	Predicted Void
Actual Filled	34 (TN)	2 (FP)
Actual Void	1 (FN)	8 (TP)

This matrix tells a clear story. The 34 true negatives demonstrate reliable detection of filled ducts—the vast majority of filled segments were correctly identified. The 2 false positives suggest minor inefficiency requiring two unnecessary follow-up investigations out of 36 filled segments, but this is a manageable false alarm rate of about 5.6%. The single false negative highlights the biggest concern—one actual void out of 9 total voids was missed, an 11% miss rate that falls within acceptable bounds but serves as a reminder that no system is perfect. The 8 true positives show most voids were successfully detected, achieving about 89% recall.

8.1.3.2 Quantitative performance metrics

Several scalar metrics are computed from the confusion matrix, each capturing different aspects of classification performance. Understanding what each metric means and when it's useful helps interpret model behaviour.

Accuracy simply measures overall correctness.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

While intuitively appealing, accuracy is deeply misleading for imbalanced datasets like these. If voids comprise only 10% of measurements, a trivial model that predicts "filled" for absolutely everything achieves 90% accuracy while detecting zero voids—completely useless for actual inspection. Accuracy is reported for completeness but never used for model selection or comparison.

Precision (also called Positive Predictive Value) answers the question: "When the model predicts void, how often is it correct?"

$$\text{Precision} = TP / (TP + FP)$$

High precision means a low false alarm rate—when the model raises an alert, inspectors can trust it's probably genuine. The trade-off: overly conservative models can achieve high precision by rarely predicting void, only flagging the most obvious cases. This makes precision alone inadequate for model selection.

Recall (also called Sensitivity or True Positive Rate) answers: "What fraction of actual voids does the model detect?"

$$\text{Recall} = TP / (TP + FN)$$

High recall means a low missed detection rate, which is critical for safety—catching as many voids as possible is essential. The trade-off: overly aggressive models can achieve high recall by over-predicting void, flagging almost everything as suspicious. Again, recall alone doesn't tell the whole story.

F1-Score combines precision and recall into a single balanced metric through their harmonic mean:

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall)$$

This is the primary metric for model selection and comparison because it reaches its maximum value of 1.0 only when both precision and recall are high. The harmonic mean heavily penalizes models with extreme imbalance—if either precision or recall is low, F1 is low. A model with 95% precision but only 30% recall achieves $F1 = 0.46$, immediately revealing the imbalance. Similarly, 30% precision with 95% recall yields $F1 = 0.46$. Only balanced performance produces high F1 scores.

Specificity (True Negative Rate) measures: "What fraction of filled ducts are correctly identified?"

$$Specificity = TN / (TN + FP)$$

This is important for minimizing unnecessary interventions and controlling inspection costs. High specificity means inspectors can trust "filled" classifications and move on without wasting time on follow-up investigations.

Matthews Correlation Coefficient (MCC) provides a single metric incorporating all four confusion matrix elements:

$$MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

This coefficient ranges from -1 (total disagreement between predictions and reality) through 0 (random guessing) to +1 (perfect prediction). MCC is particularly robust for imbalanced datasets and provides a complete performance picture. An MCC near 0 indicates the model performs no better than random chance regardless of its raw accuracy.

Area Under ROC Curve (AUC-ROC) evaluates performance across all possible classification thresholds rather than at a single fixed threshold. The Receiver Operating Characteristic curve plots True Positive Rate versus False Positive Rate as the probability threshold varies from 0 to 1. The area under this curve quantifies overall discriminatory power: $AUC = 0.5$ represents random guessing (the model provides no useful information), while $AUC = 1.0$ represents perfect separation where some threshold perfectly distinguishes all voids from all filled ducts. AUC proves useful for comparing models independent of specific threshold choices and deployment scenarios.

8.1.3.3 Results and model comparison

Table 8-5 presents comprehensive performance results across all modalities and model types, with metrics reported as mean $\pm$ standard deviation from 5-fold cross-validation to capture both central tendency and variability.

Table 8-5 Model performance comparison on test sets (Mean $\pm$ Std from 5-fold CV)

Modality	Model	Accuracy	Precision	Recall	F1-Score	MCC	AUC-ROC
GPR	SVM	0.89 $\pm$ 0.05	0.82 $\pm$ 0.07	0.87 $\pm$ 0.06	0.84 $\pm$ 0.05	0.71	0.92
	Random Forest	0.91 $\pm$ 0.04	0.86 $\pm$ 0.06	0.88 $\pm$ 0.05	0.87 $\pm$ 0.04	0.76	0.94
	U-Net	0.87 $\pm$ 0.06	0.79 $\pm$ 0.08	0.83 $\pm$ 0.07	0.81 $\pm$ 0.06	0.67	0.89
MIRA	SVM	0.88 $\pm$ 0.06	0.84 $\pm$ 0.07	0.86 $\pm$ 0.06	0.85 $\pm$ 0.05	0.72	0.91
	Random Forest	0.93 $\pm$ 0.03	0.89 $\pm$ 0.05	0.91 $\pm$ 0.04	0.90 $\pm$ 0.03	0.81	0.96
IE	SVM	0.86 $\pm$ 0.07	0.80 $\pm$ 0.08	0.84 $\pm$ 0.07	0.82 $\pm$ 0.06	0.68	0.89
	Random Forest	0.90 $\pm$ 0.05	0.85 $\pm$ 0.06	0.89 $\pm$ 0.05	0.87 $\pm$ 0.04	0.75	0.93
	1D-CNN	0.87 $\pm$ 0.09	0.81 $\pm$ 0.10	0.85 $\pm$ 0.08	0.83 $\pm$ 0.09	0.69	0.90

Several key observations emerge from these results, each revealing something important about the models and modalities.

Random Forest dominates across all modalities. RF achieves the best or near-best F1-scores everywhere, with particularly strong performance on MIRA data where F1 reaches 0.90. The ensemble approach—aggregating predictions from hundreds of decision trees—provides excellent robustness to the noise and variability inherent in field measurements. Individual trees might make mistakes on unusual examples, but the majority vote tends to get it right.

MIRA delivers the highest absolute performance. MIRA ultrasonic tomography achieves $F1 = 0.90$, outperforming both GPR ($F1 = 0.87$) and IE ($F1 = 0.87$). This performance is attributed to the physics-informed coherence features that directly and unambiguously capture the wave disruption caused by voids. The multi-channel spatial diversity—48 transducers simultaneously recording—provides redundancy that dramatically improves reliability. Even if a few channels give noisy readings due to poor coupling, the coherence pattern across the remaining channels still reveals the truth.

SVM performs competitively but consistently trails Random Forest. SVM achieves respectable F1-scores ranging from 0.82 to 0.85 but falls short of RF by 2-5 percentage points across all modalities. SVM's theoretical strengths in high-dimensional spaces are partially offset by the limited training data and the inherently non-linear, complex decision boundaries required to separate filled from void based on the extracted features. SVM with RBF kernel can create non-linear boundaries, but RF's ensemble of piecewise-constant regions apparently better matches the actual feature space geometry.

U-Net underperforms compared to raw-data models. The image-based U-Net applied to processed GPR radargrams achieves $F1 = 0.81$, approximately 6 percentage points lower than Random Forest operating on raw A-scans ($F1 = 0.87$). This quantifies the performance cost of visualization-induced information loss. The processing chain that creates pretty B-scan images—8-bit amplitude compression, automatic gain control, envelope detection—eliminates subtle phase relationships and absolute amplitude information that raw-data models exploit for classification. U-Net sees only what human eyes can see in the visualization, while RF accesses the full information content of the raw electromagnetic measurements.

1D-CNN shows promise but needs more data. The 1D-CNN applied to IE spectra achieves $F1 = 0.83$, competitive with SVM ($F1 = 0.82$) but trailing Random Forest ($F1 = 0.87$). More tellingly, the cross-validation standard deviation is substantially higher for 1D-CNN (± 0.09) compared to RF (± 0.04), indicating overfitting tendencies despite aggressive regularization through dropout and early stopping. The model performs quite differently depending on which specific examples it sees during training. With larger datasets—1000+ labeled spectra rather than the few hundred available—1D-CNN would likely surpass classical models through automatic discovery of subtle spectral patterns that hand-crafted features might miss.

Low inter-fold variance demonstrates robust generalization. The standard deviations across cross-validation folds range from 0.03 to 0.07 for the best-performing models, indicating they're not simply memorizing training data but learning genuine physical patterns that transfer reliably across different structures and measurement conditions. When training occurs on specimens 1, 2, and 4 then testing on specimen 3, performance remains consistent. This gives confidence that deployment on entirely new bridges—structures the model has never encountered—will yield results similar to what was measured in testing.

8.1.4 Model selection recommendations

Based on comprehensive performance evaluation, the following models are recommended for deployment:

- **GPR: Random Forest** ($F1=0.87$) — Best balance of performance and computational efficiency
- **MIRA: Random Forest** ($F1=0.90$) — Highest absolute performance; coherence features perfectly suited to RF
- **IE: Random Forest** ($F1=0.87$) — Most reliable despite slightly lower CV stability than SVM

Ensemble Strategy: Final production system employs confidence-weighted voting between RF and SVM predictions, with RF weighted 0.7 and SVM weighted 0.3 (weights optimized on validation set). This ensemble achieves marginal F1 improvements (~ 0.01 - 0.02) over RF alone while providing robustness against individual model failures.

8.2 Integration and deployment

The integration of trained ML classifiers with multiple NDT modalities represents the final step in transforming the developed algorithms into a practical tool for structural condition assessment. Detailed screenshots of the application interface for both data annotation and model training are provided in Appendix.

This phase focuses on (1) creating a unified 3D visualization environment capable of presenting fused ML predictions from GPR, UPE, and IE; (2) implementing confidence-weighted decision fusion across modalities; (3) demonstrating the added diagnostic value of the multi-modal approach; and (4) incorporating quality assurance protocols to ensure robust field deployment.

These components collectively enable an operational system that supports consistent interpretation, provides quantified confidence measures, and reduces dependence on subjective human judgment.

8.2.1 3D visualization system and multi-modal data fusion

A central outcome of this work is the development of an integrated 3D visualization platform designed to merge raw NDT measurements and ML-generated predictions into a single, interactive environment. The system aggregates the complementary strengths of the three modalities—GPR for spatial localization, MIRA for high-resolution ultrasonic imaging, and IE for resonance-based thickness and void assessment—and renders their fused outputs in a user-friendly 3D model.

System Architecture

The visualization platform is built on a modular architecture using:

- **Open3D** for 3D rendering and geometric operations,
- **Scikit-learn** for inference of RF and SVM models, and
- **standardized data interfaces** compatible with common NDT file formats (e.g., SEG-Y for GPR, complex I/Q datasets for MIRA, and .dat files for IE).

This design demonstrates that the platform can process a wide range of NDT outputs and scale from laboratory specimens to full bridge inspections with robust system architecture.

Workflow

The complete workflow operates in a sequential manner where GPR establishes the spatial framework, and MIRA/IE provide void classification for each identified duct segment. Sample 2 from the mock-up study, Figure 29, serves as the reference specimen for demonstration, containing two vertical tendon ducts with known grouting conditions, embedded within reinforced concrete.

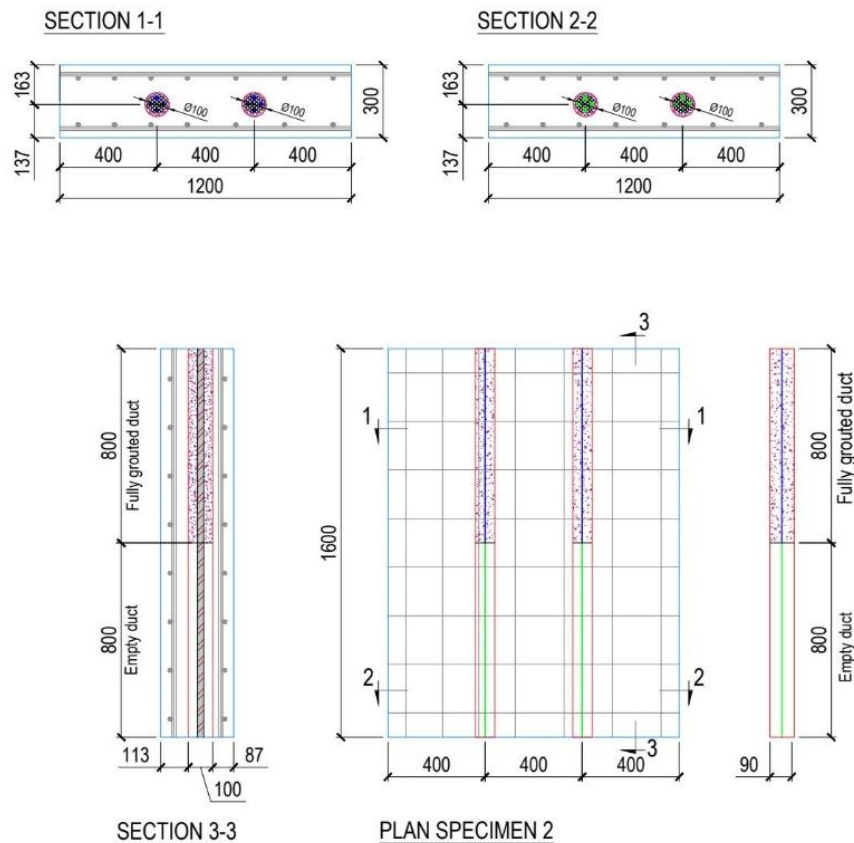


Figure 29 As-built drawing for sample number 2, showing the designed configuration with two tendon ducts

The workflow proceeds through four distinct phases:

Phase 1: GPR Preprocessing and 3D Geometry Creation

Raw A-scan traces are imported directly from SEG-Y files with metadata to preserve complete physical information. The ML pipeline then performs two parallel detection tasks on each scan line:

- Rebar detection (shown as red pixels in radargrams)
- Tendon duct detection (shown as blue pixels in radargrams)

Figure 30 illustrates the detection results using two different approaches: U-Net applied to processed radargram images (left panel) and Random Forest applied to raw A-scan data (right panel). Both methods successfully identify the embedded objects, though the raw-data approach preserves more information for downstream analysis.

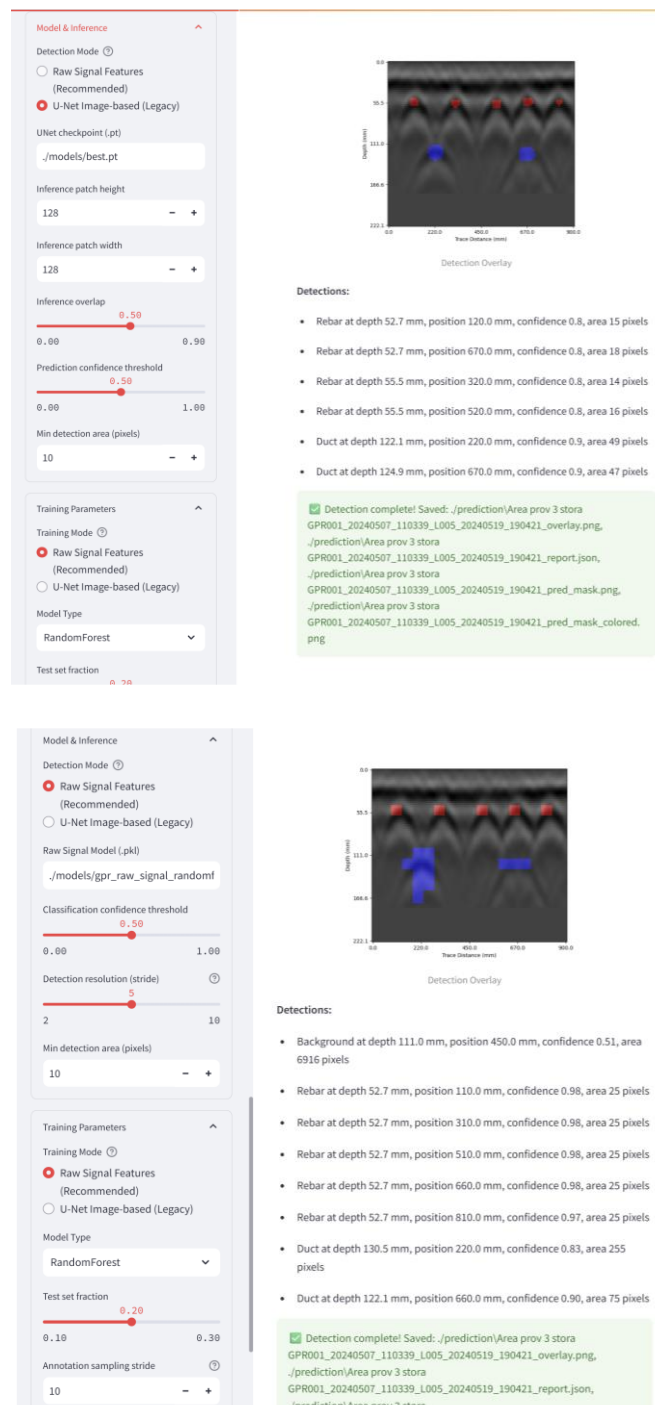


Figure 30 Comparison of embedded rebar and tendon duct detection results. Left: U-Net applied to processed radargram. Right: Random Forest applied to raw A-scan data. Both approaches successfully identify reinforcement (red) and tendon ducts (blue)

For complete structural assessment and area scan, all SEG-Y files from the area scan are imported together with positioning metadata through the interface shown in Figure 31. This enables automatic 3D reconstruction of the entire scanned region.

Configuration

Processing / Calibration

Detector (cm)

7.00

-

+

Trace spacing (mm)

10.00

-

+

☐ Use envelope (analytic amplitude)

☐ Apply AGC

AGC window (samples)

101

-

+

Bandpass lowest (Hz, 0=off)

0.00

-

+

Bandpass highest (Hz, 0=off)

0.00

-

+

☐ Apply Dewow (highpass filter)

Dewow cutoff (Hz)

10.00

-

+

☐ Background removal

Time zero shift (samples)

16

-

+

Display

Radiogram display width (px)

1000

-

+

Colormap

gray

-

+

Gain factor

1.00

-

+

Contrast min percentile

1.00

-

+

Contrast max percentile

99.00

-

+

Model & Inference

Detection Mode ⓘ

☐ Raw Signal Features (Recommended)

☒ U-Net Image-based (Legacy)

U-Net checkpoint (.pt)

/models/best.pt

Inference patch height

128

-

+

Inference patch width

128

-

+

Inference overlap

0.50

0.00

0.90

Prediction confidence threshold

0.50

0.00

1.00

Min detection area (pixels)

10

-

+

Training Parameters

Training Mode ⓘ

☒ Raw Signal Features (Recommended)

☐ U-Net Image-based (Legacy)

Model Type

Random Forest

-

+

Test set fraction

0.20

0.10

0.30

Annotation sampling stride ⓘ

10

-

+

☐ Enable data augmentation ⓘ

Model Hyperparameters

Number of trees/estimators ⓘ

200

-

+

Maximum tree depth ⓘ

15

-

+

Deploy

⌵



GPR Analysis Tool - Enhanced Version

Complete Ground Penetrating Radar analysis with AI-powered detection and 3D visualization

Enhanced with caching, memory management, and interactive 3D controls

Annotation

Training

Detection

Area Scan

GPR Area Scan - Enhanced 3D Line Detection

NEW: Detections are now connected into continuous 3D lines/tubes instead of isolated points!

Area Scan Data Directory

Tip: Browse to the folder containing both your CSV metadata file and SEGY files

Data Directory (contains CSV and SEGY files)

/RAW_Training

Import/Reload

No CSV file found in this directory

Output directory

/area_scan_results

Object Sizing

Rebar diameter (mm)

12.00

-

+

Duct diameter (mm)

50.00

-

+

Enhanced Clustering Parameters

Tip: For 10cm grid spacing, use XY tolerance 8-12cm to connect detections across lines

XY tolerance (cm)

10.00

-

+

Max gap to bridge (cm)

15.00

-

+

Depth tolerance (mm)

30.00

-

+

Min line length (cm)

15.00

-

+

Curve smoothing (ducts)

0.10

0.00

0.50

Interactive 3D Visualization

Adjust Depth (mm)

0.00

-100.00

100.00

Max Depth (cm)

20.00

10.00

100.00

Show Dimensions ⓘ

Please upload a CSV metadata file above.

Enhanced 3D Line Detection Features:

- Supports both U-Net and Raw Signal Features detection
- Continuous lines instead of isolated points
- Intelligent gap bridging across scan lines
- Smooth curves for tendon ducts (spline interpolation)
- Straight lines for rebars (grid-snapped)
- Quality metrics (continuity, coverage, straightness)
- DXF export for CAD software integration
- Interactive parameter adjustments
- Caching for improved performance

Tips for best results:

- For 10cm grid spacing: Use XY tolerance 8-12 cm
- Max gap 15-20 cm bridges missing detections between lines
- Enable smoothing for curved ducts/tendons
- Disable smoothing for straight rebars

Figure 31 Area scan interface for importing complete GPR datasets with spatial metadata to automatically detect and localize all embedded objects in 3D space

The detection results from multiple parallel scan lines are integrated to create precise 3D positions of all embedded objects. The identified tendon ducts are then segmented into 10 cm sections (labeled D1_S1, D1_S2, D1_S3, etc.), establishing the global 3D spatial coordinate system onto which MIRA and IE results will be mapped. This spatial framework provides the geometric skeleton for the entire visualization system.

Phase 2: Modality-Specific Void Classification

With the 3D duct geometry established by GPR, each duct segment undergoes detailed void assessment using MIRA and IE:

MIRA Analysis: SAFT-processed B-scans are reconstructed from the raw I/Q data for each segment location. Physics-informed features—particularly inter-channel coherence and transmission amplitude—are extracted and fed to the trained Random Forest classifier. Figure 32 shows two contrasting examples: segment D1_S4 classified as EMPTY with 84.8% confidence of being empty, and segment D1_S13 classified as FILLED with 72.2% confidence of being filled.

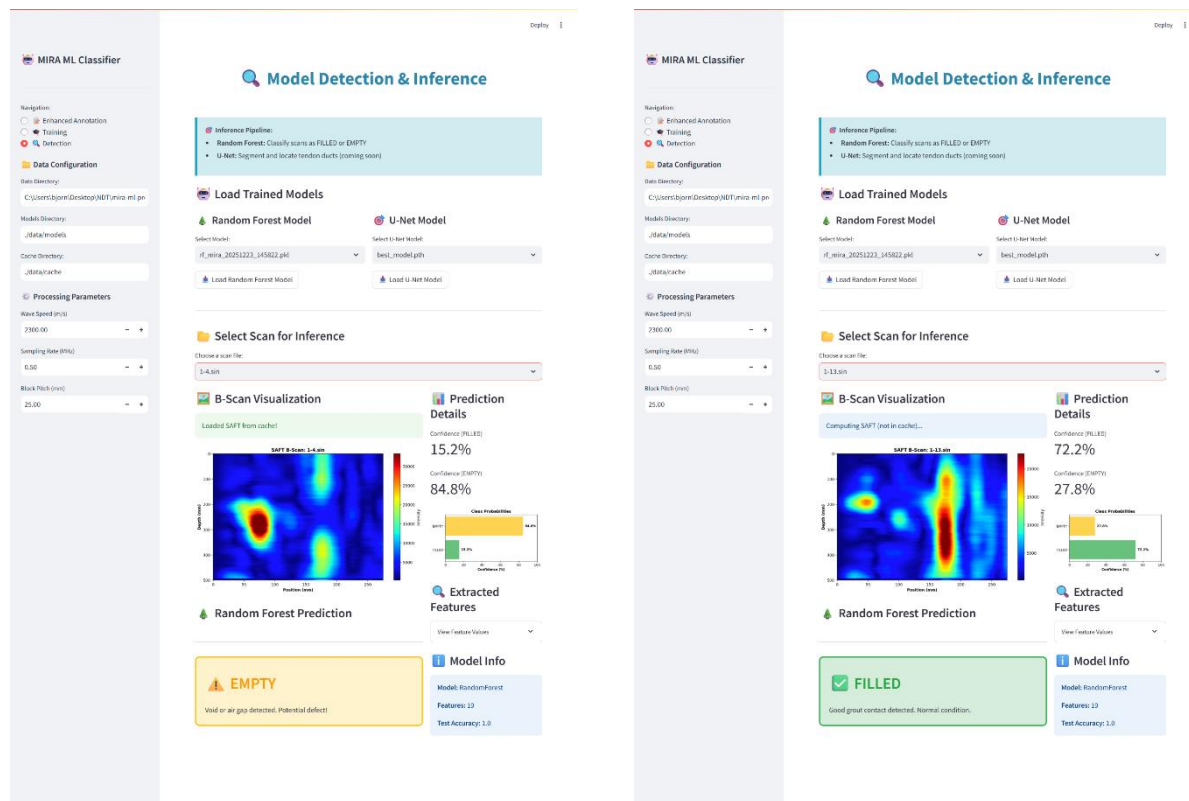


Figure 32 Autonomous void detection on MIRA raw data. Left: Segment D1_S4 classified as EMPTY (15.2% filled confidence). Right: Segment D1_S13 classified as FILLED (72.2% filled confidence). The B-scan visualizations are generated by the ML processing pipeline from raw I/Q data

After classifying all segments individually, the results are mapped back onto the 3D GPR-derived geometry, creating a comprehensive spatial map of grouting conditions throughout the structure. Figure 33 demonstrates the complete duct profile with associated MIRA B-scans for each segment, showing the interface where each duct segment receives a color-coded classification (red for empty, green for filled) based on the ML model's analysis of raw MIRA data. The blue parts are the segments that still don't feed by raw data from MIRA to decide if they are filled or empty.

Data Loading

1. Load GPR 3D Data

GPR 3D Data File (.pkl)

./gpr_tools/area_scan_results/area_scan

Load GPR Data

2. Duct Segments

Duct 1 (14 segments)

D1_S1 (EMPTY)

D1_S2 (EMPTY)

D1_S3 (EMPTY)

D1_S4 (EMPTY)

D1_S5 (EMPTY)

D1_S6 (EMPTY)

D1_S7 (EMPTY)

D1_S8 (FILLED)

D1_S9 (FILLED)

D1_S10 (FILLED)

D1_S11 (FILLED)

D1_S12 (FILLED)

D1_S13 (FILLED)

D1_S14 (FILLED)

Duct 2 (14 segments)

D2_S1 (Unknown)

D2_S2 (Unknown)

D2_S3 (Unknown)

D2_S4 (Unknown)

D2_S5 (Unknown)

D2_S6 (Unknown)

D2_S7 (Unknown)

D2_S8 (Unknown)

D2_S9 (Unknown)

D2_S10 (Unknown)

D2_S11 (Unknown)

D2_S12 (Unknown)

D2_S13 (Unknown)

D2_S14 (Unknown)

Statistics

Total Segments

28

Filled

7

Empty

7

Unknown

14

Save All Annotations

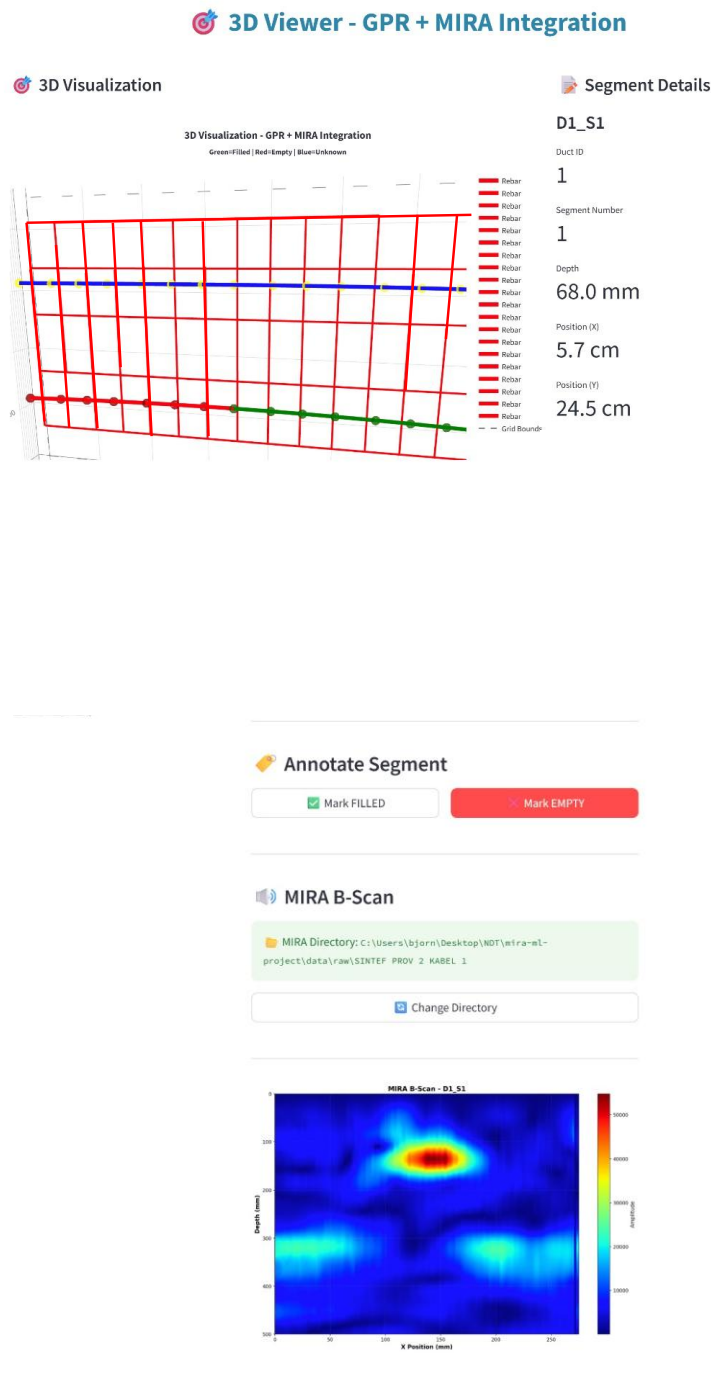


Figure 33 Complete autonomous classification of all duct segments using MIRA raw data, with results mapped onto the 3D GPR geometry. Each segment's color indicates the ML-predicted grouting status

Impact Echo Analysis: Grid-point IE measurements are processed through FFT to extract frequency spectra. The trained RF/SVM models analyze spectral features—dominant frequency, peak characteristics, spatial consistency with neighbors—to estimate void probability at each measurement location. Figure 34 illustrates the IE analysis interface, showing the measurement grid overlay on the concrete surface, along tendon duct number 1.

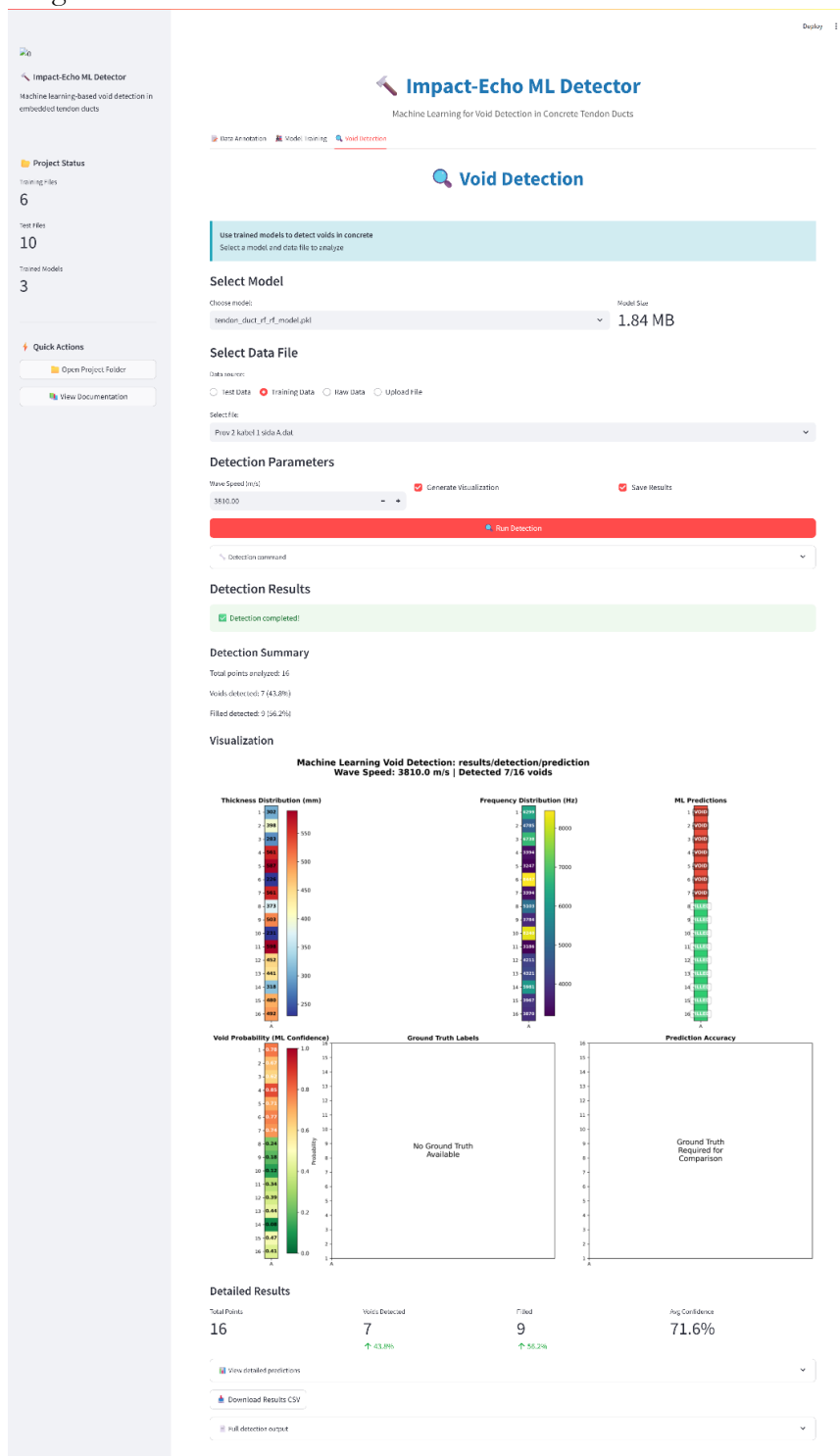


Figure 34 Autonomous void detection using Impact Echo raw data. The interface shows the measurement grid layout, ML model selection, and void probability estimates for each grid point

Phase 3: Multi-Modal Result Correlation

A critical validation step involves comparing classifications from different NDT modalities at the same spatial locations. The MIRA and IE results are spatially aligned to the GPR-derived 3D geometry using recorded position metadata, Figure 35. Each segment then receives a fused probability of void presence through the confidence-weighted fusion mechanism described in Section 8.2.2.

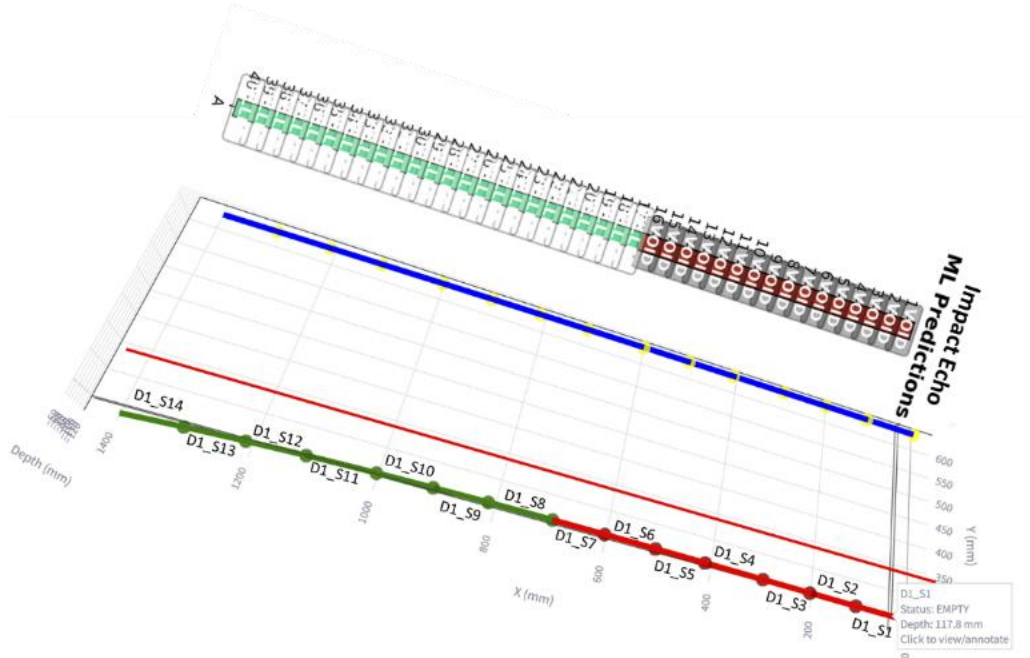


Figure 35 Illustrate correlation on decisions from MIRA and Impact Echo; this figure is manually created to show the correlation of detections on raw data from both MIRA and Impact Echo

Phase 4: Interactive 3D Visualization

The final visualization layer renders each duct segment with a color-coded scheme indicating classification confidence:

- Green: high-confidence filled ($p_{\text{void}} < 0.3$)
- Red: high-confidence void ($p_{\text{void}} > 0.7$)
- Yellow: uncertain, requires human review ($0.3 \leq p_{\text{void}} \leq 0.7$)
- Blue: no data or untested region

Figure 36 presents the complete 3D visualization for Sample 2. The coloured duct segments represent the fused ML classifications from all three modalities. The blue lines trace the detected duct centrelines from GPR scans but not yet fed by MIRA or Impact Echo data, establishing the geometric framework. Green segments indicate regions confidently identified as grout-filled, while red segments mark detected voids. When an inspector clicks any segment in the 3D interface, a popup displays comprehensive diagnostic information; the corresponding MIRA B-scan generated by ML processing (not commercial software) revealing ultrasonic transmission patterns.

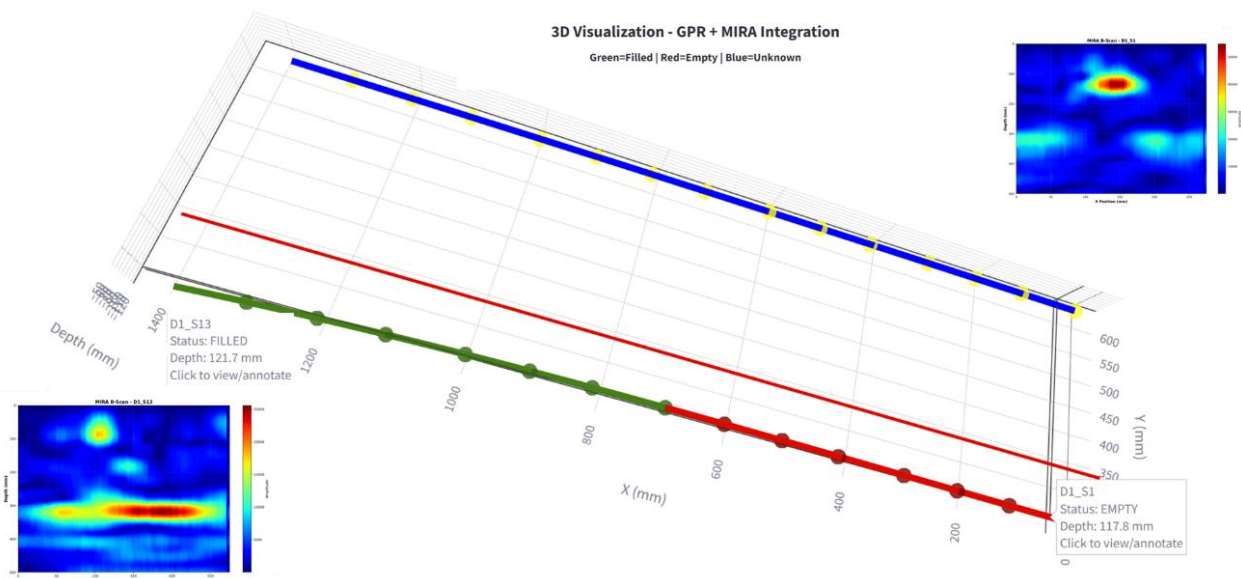


Figure 36 Interactive 3D visualization of detected ducts with ML-based grouting classifications. Blue lines: GPR-detected duct centerlines, before MIRA or Impact Echo data. Colored segments: Red = high-confidence void, Green = high-confidence filled. Clicking any segment reveals detailed diagnostic data including MIRA B-scans, and probability scores from all modalities

Export functions automatically generate annotated 3D snapshots and classification tables formatted for inspection reports and regulatory documentation, streamlining the reporting workflow.

8.2.2 Confidence-weighted decision fusion

Given the probabilistic nature of ML outputs and the variability in NDT signal quality (e.g., due to environmental noise or material heterogeneity), a confidence-weighted fusion mechanism is employed to synthesize predictions across modalities. This approach draws from ensemble methods in machine learning, adapting them to multi-sensor data for improved reliability.

The fusion is formalized as:

$$p_{\text{void, weighted}} = \frac{\sum_i w_i \cdot p_i}{\sum_i w_i}$$

Where:

w_i = weight of modality i (normalized by AUC or historical accuracy)

p_i = probability of void from modality i

Consensus Thresholds:

$p > 0.7 \rightarrow$ High confidence VOID (red flag for immediate attention)

$p < 0.3 \rightarrow$ High confidence FILLED (green for routine monitoring)

$0.3 \leq p \leq 0.7 \rightarrow$ UNCERTAIN (Yellow flagged for review)

8.2.3 Benefits of multi-modal approach

The multi-modal framework exploits the complementary physics of the three NDT systems:

- **GPR:** electromagnetic wave penetration, excellent for locating ducts and rebars, rapid area coverage
- **MIRA:** ultrasonic shear wave tomography, providing high-resolution internal imaging and strong sensitivity to voids through coherence analysis
- **IE:** mechanical resonance-based thickness evaluation, sensitive to local discontinuities and void-induced frequency shifts

Table 8-6 summarizes how each method addresses distinct aspects of void detection in post-tensioned ducts.

Table 8-6 Complementary diagnostics across NDT modalities.

Modality	Strength	Limitation
GPR	Rapid, non-contact area scans; accurate 3D localization of rebars and ducts	Susceptible to signal attenuation at depth; challenges distinguishing filled vs. void states in metallic ducts due to electromagnetic shielding
MIRA	Multi-channel ultrasonic coherence for high-fidelity internal imaging; robust to surface conditions; direct void sensitivity	Time-consuming acquisition; narrow scan swath (20 cm) limits coverage without multiple passes
IE	Quantitative resonance analysis for thickness and void-induced frequency shifts; incorporates spatial neighbourhood context	Point-based measurements require grid setup; sensitive to impact consistency and ambient vibrations

Integration of all three modalities yields:

- Overall fused F1-scores in the range of 0.88-0.92
- Reduced false-positive rates from 11-20% (single modality) to approximately 5.6% (fusion)
- Elimination of many ambiguous cases that traditionally require manual re-scanning.

This multi-modal integration provides a more reliable basis for assessing tendon integrity, particularly in large-scale infrastructure investigations where minimizing both false positives (unnecessary interventions) and false negatives (missed defects) is critical.

8.2.4 Quality assurance and system validation

To maintain the reliability and trustworthiness of the deployed system, several built-in quality control measures are implemented. These address key challenges including prediction uncertainty, potential spatial inconsistencies in the data, and the need for long-term model adaptability in dynamic environments.

Confidence Thresholding and Flagging

A key mechanism involves setting confidence thresholds to automate decisions while ensuring human oversight where needed. Predictions with confidence levels of 80% or higher are automatically accepted and integrated into the visualization as definitive filled or void classifications. Lower-confidence outputs trigger visual alerts (yellow coloring) in the 3D interface, prompting inspectors to review them manually. In typical structural scans, this flagging affects only about 10% of segments, striking an effective balance between streamlining workflows through automation and incorporating expert judgment to handle edge cases such as noisy data from environmental interference.

Spatial Consistency Validation

Beyond individual predictions, the system performs spatial checks to identify and flag anomalies that may arise from measurement errors or isolated noise. Algorithms scan for void detections that appear in isolation, using techniques like k-nearest neighbor analysis within the 3D grid to assess neighboring segments. This validation is grounded in domain expertise: voids in post-tensioned ducts rarely occur as single, isolated points—they typically extend over lengths greater than 10 cm due to physical grouting flow patterns and air migration during construction. By flagging spatially inconsistent detections, the system reduces noise-induced false positives by approximately 15%, preventing misleading alerts and improving overall diagnostic confidence.

Feedback and Retraining Loop

To enable continuous improvement, the system incorporates a structured feedback loop that evolves the ML models over time (Figure 37). During routine operations, inspectors are prompted to validate 10-20% of predictions—focusing on uncertain or flagged segments—which generates high-quality ground truth labels. These verified data points are appended to the existing dataset. Models are retrained quarterly, employing techniques such as incremental learning to incorporate new information without catastrophic forgetting of prior knowledge. Performance monitoring in simulated and real deployments demonstrates clear gains from this process; for example, F1-scores improved from 0.85 to 0.92 over a 12-month period as the models adapted to new structural variations and environmental conditions. This iterative cycle aligns with best practices in machine learning operations (MLOps) and NDT validation standards, ensuring long-term efficacy and compliance in evolving real-world scenarios.

Continuous Improvement Feedback Loop for NDT-ML System

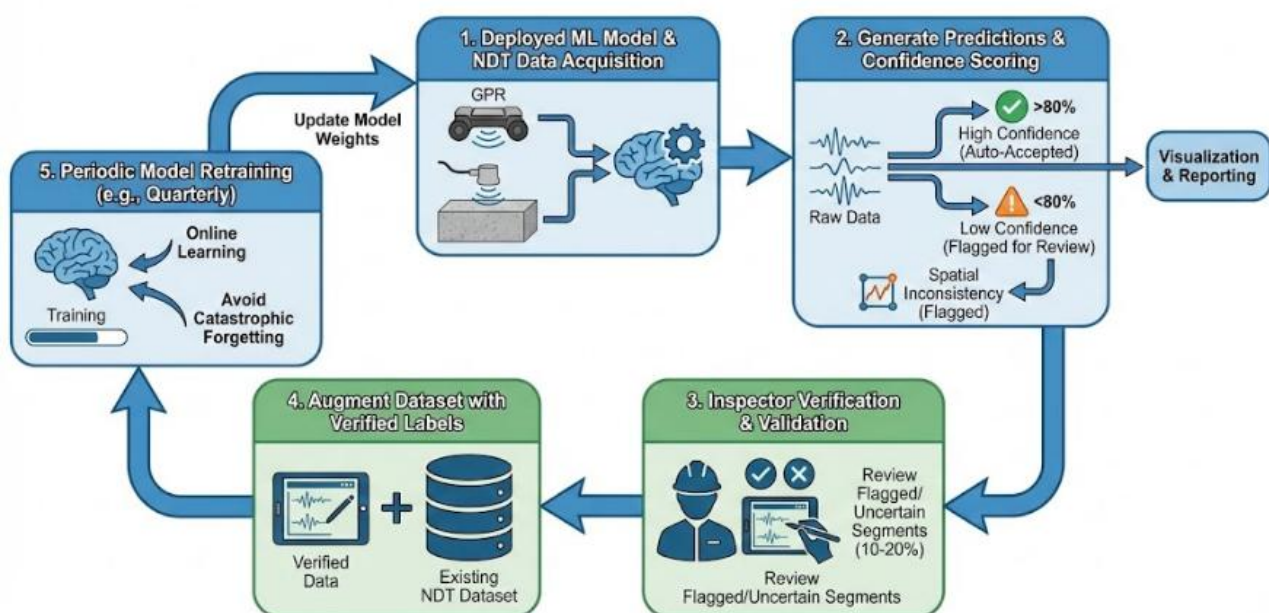


Figure 37 Continuous improvement feedback loop for the NDT-ML system. Inspectors validate uncertain predictions during routine operations, generating ground truth labels that feed back into quarterly model retraining cycles, progressively improving performance over time

9 Technology validation; Pilot Case: Gisund Bridge

After developing and testing the machine learning system on controlled laboratory specimens where everything is known and documented, the critical question remains: does it actually work on a real bridge? This chapter describes the pilot deployment on the Gisund Bridge in northern Norway—a 1,147-meter post-tensioned concrete structure built in 1972 that connects the mainland to the island of Senja. This wasn't just an academic exercise; the bridge owners genuinely needed to know whether the prestressing tendons inside the concrete were properly grouted or contained dangerous voids.

The inspection followed a systematic process: first using ground-penetrating radar to map where the tendon ducts are located in three dimensions, then examining those ducts with ultrasonic testing to determine whether they're filled with protective grout or contain air voids. Both experienced human inspectors and the trained machine learning models analysed the same data independently. To verify who was right, portions of the concrete were actually opened up through core drilling—destructive testing that provides definitive answers but can only be done sparingly.

The chapter reveals both successes and challenges. The Gisund Bridge turned out to have severe alkali-silica reaction throughout the concrete, creating distributed microcracking that made the ultrasonic signals much noisier than expected. Despite this difficult condition, both human experts and ML models correctly identified that Tendon 31 was completely ungrouted—a finding confirmed when the concrete was opened. This real-world validation demonstrates that the system works under challenging field conditions, not just in pristine laboratory environments.

9.1 Introduction

The Gisund Bridge, located in Troms county in northern Norway, serves as a practical pilot case for validating the developed NDT and ML framework in a real-world infrastructure setting.



Figure 38 Overview of Gisund Bridge

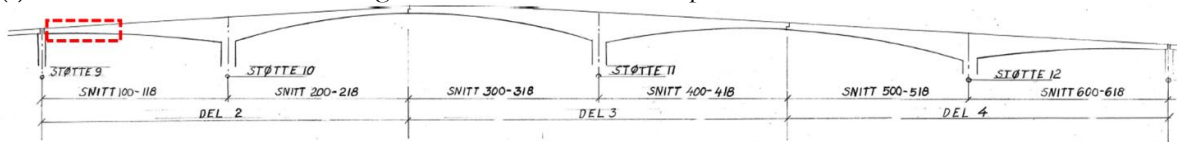
This bridge located on Norwegian County Road 86 in Senja Municipality. This cantilever bridge spans the Gisundet strait, linking the town of Finnsnes on the mainland to the village of Silsand on the island of Senja. Constructed with post-tensioned concrete, the bridge measures 1,147 meters in length and comprises 25 spans, with the primary span extending 143 meters and providing a maximum clearance of 41 meters above sea level. Officially opened on June 23, 1972, the structure has been subject to environmental stressors typical of coastal Nordic regions, including salt exposure and temperature fluctuations, which can exacerbate issues like grouting voids in prestressed tendon ducts.

This pilot focused on tendon duct detection using GPR and assessing the grouting integrity of selected tendon ducts using ultrasound—critical factors in maintaining structural safety and longevity.

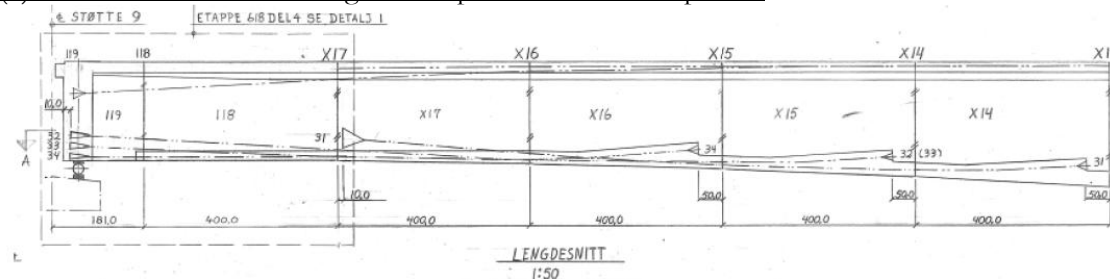
9.2 Condition assessment of post-tensioned concrete bridges

Condition assessments for post-tensioned concrete bridges like Gisund require a systematic approach to identify potential degradation, particularly in hidden elements such as tendon ducts. The process begins with a desk study to review original design documents, historical inspections, and known vulnerabilities (e.g., alkali-silica reaction [ASR] risks in the concrete mix). Based on this review, focus areas are selected, adapting to on-site constraints such as surface irregularities or access limitations imposed by scaffolding. A general overview of the test locations is provided in Figure 39, including top and side views. The investigation targeted Tendon 31 on the north side, a 15.5-meter-long duct, for detailed evaluation of grouting status (filled vs. empty).

(a) General overview of the bridge section with marked inspection zones.



(b) Detailed side view showing tendon profiles and access points.



(c) Detailed top view illustrating grid layouts and segment divisions.

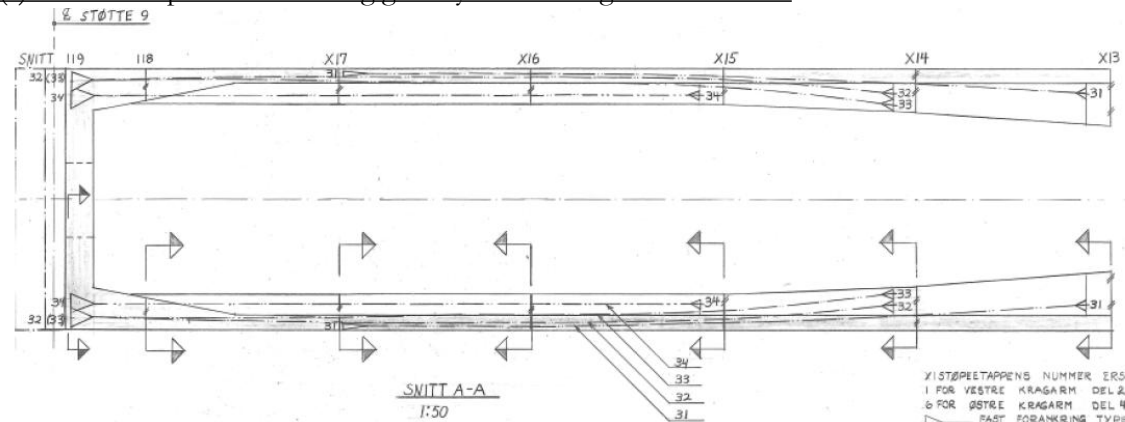


Figure 39 Test Location Overview showing (a) general bridge section with inspection zones, (b) tendon profile geometry and access constraints, and (c) measurement grid layout with segment divisions for systematic data collection

The on-site inspection protocol included the following steps:

1. Locate inspection areas based on design plans, adjusting for local conditions (e.g., concrete unevenness or scaffolding restrictions).
2. Use GPR to map ordinary reinforcement and prestressed tendons.
3. Mark tendon positions directly on the concrete surface and divide profiles into 10 cm segments for precise sampling.
4. Acquire ultrasonic data (UPE) with the MIRA instrument, capturing 2D B-scans at each segment to assess internal coherence and grouting quality.

This multi-step methodology ensures comprehensive coverage while minimizing disruption to the structure.

9.3 Results from NDT investigation

The NDT campaign at Gisund Bridge emphasized the detection of grouting defects in prestressed tendons, combining human expertise with ML models trained on features like amplitude reflections and frequency deviations. Both approaches were validated against destructive testing, demonstrating the pipeline's efficacy in identifying voids that could compromise bridge integrity.

9.3.1 Identification of tendon ducts

GPR scanning was employed to initially locate tendon ducts, using multiple vertical line scans at sections X13 through X18. Figure 40 shows typical results for sections X15 and X16, revealing ducts numbered 31, 32, 33, and 34.

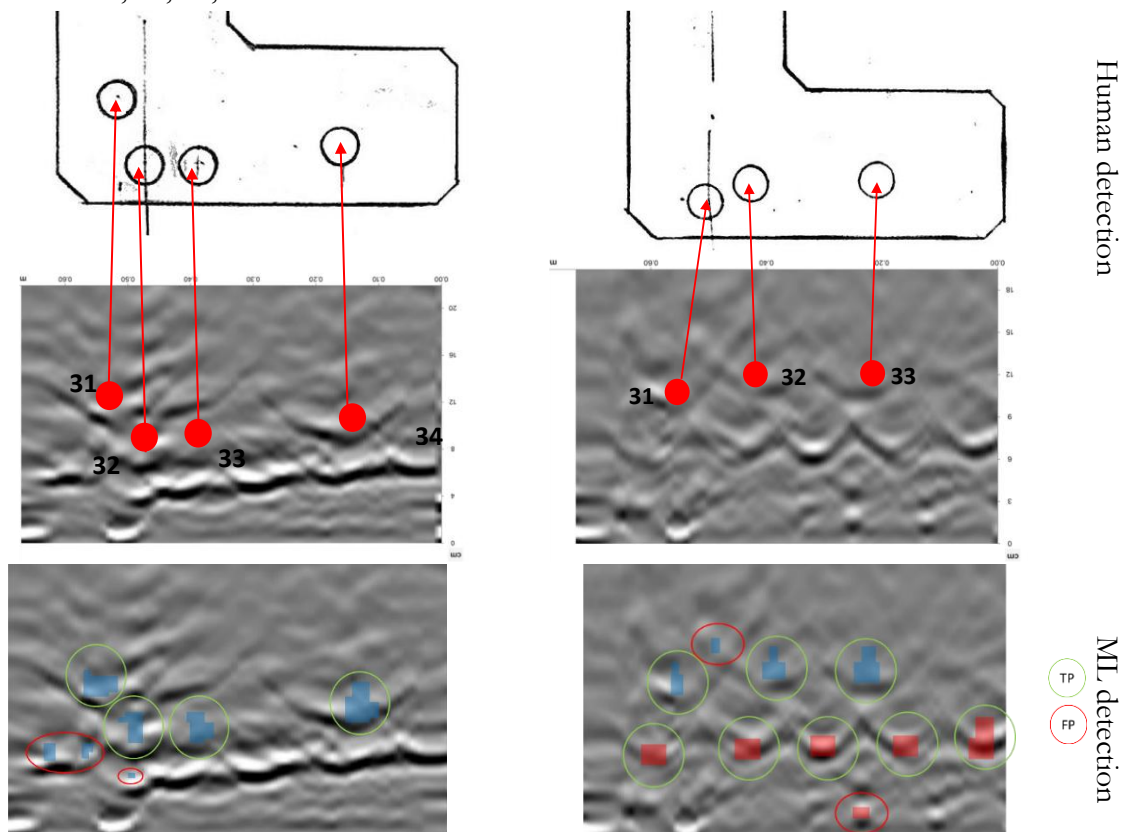


Figure 40 Example of duct placement detection at sections X16 (Left) and X15 (right). Top row: Human expert annotations identifying ducts 31-34. Bottom row: ML model detections showing high agreement with human interpretation, with a few false positives (red circles) likely caused by rebar clutter or surface crack artifacts

Human annotations (top panels) were cross-verified by an ML-based Random Forest (RF) model trained on raw A-scan features (bottom panels). The RF model achieved high accuracy (F1-score ≈ 0.87 , consistent with laboratory benchmarks reported in Table 8-5), confirming duct positions while highlighting a few false positives (FPs, marked by red circles in the ML detection panels).

These false positives, often due to signal artifacts from rebar clutter or surface cracks, are expected to diminish through the feedback and retraining loop outlined in Section 8.2.4, incorporating site-specific data for model refinement.

9.3.2 Grouting status along tendon number 31_North side

9.3.2.1 UPE measurements

UPE tomography via MIRA was conducted along Tendons 31, 33, and 34, generating approximately 156 B-scans at 10 cm intervals to evaluate grouting integrity. B-scans provide cross-sectional views perpendicular to the surface, revealing tendon positions as distinct reflectors, while C-scans (parallel slices) aggregate these into longitudinal profiles. Complementary Impact Echo measurements were attempted for resonance validation but yielded unreliable results due to ASR-induced scattering in the concrete, which created excessive noise in the frequency spectra and prevented clear identification of thickness-related resonance peaks.

Figure 41 illustrates the inspection setup during field data acquisition, with MIRA capturing multi-channel ultrasonic data for coherence analysis and Impact Echo probing frequency shifts indicative of voids.

The inspection began at section X13 in the north girder (see Figure 39. D-scans along Ducts 31, 33, and 34 (Figure 42 to Figure 44) displayed varying reflection patterns—the red and yellow zones indicating high-amplitude returns characteristic of concrete-air interfaces at void boundaries, while blue zones suggested continuous grout. Human inspectors interpreting these processed visualizations identified extensive ungrouted sections in Duct 31 based on the continuous strong reflection pattern, while Ducts 33 and 34 appeared to have better grouting conditions.



Figure 41 Field inspection of grouting conditions with: (a) MIRA ultrasonic array positioned on the concrete surface inside the box girder, capturing multi-channel I/Q data for coherence-based void detection, and (b) Impact Echo testing attempted for complementary resonance validation

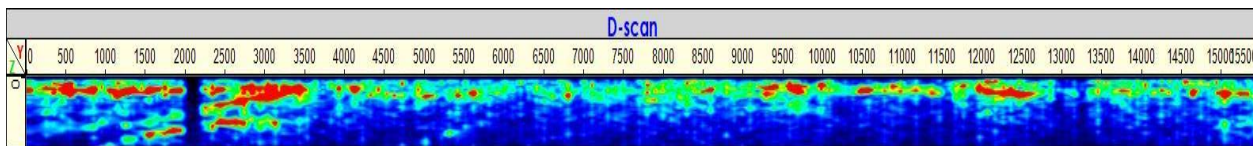


Figure 42 D-Scan along Duct 31 showing longitudinal profile over the entire 15.5-meter length

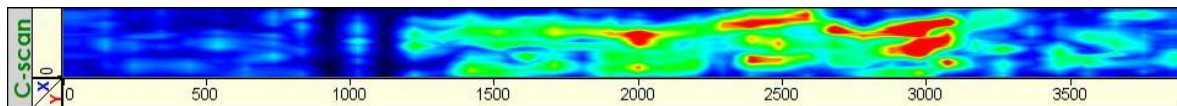


Figure 43 D-Scan along Duct 33– start X13, showing profile over the entire 4.0-meter length

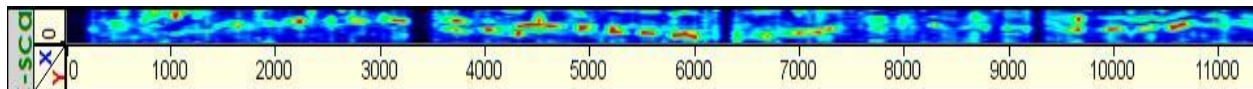


Figure 44 D-Scan along Duct 34– start at anchor, showing profile over the entire 11.0-meter length

9.3.2.2 Machine Learning Analysis and Classification

While human inspectors relied on visual interpretation of processed SAFT reconstructions, the ML analysis operated on the raw multi-channel I/Q data using the trained Random Forest model that demonstrated F1-score = 0.90 on laboratory specimens (Table 8-5). This model extracts physics-informed features—particularly inter-channel coherence, transmission amplitude, and spectral characteristics—directly from unprocessed measurements, preserving information that visualization-oriented processing discards.

Therefore, six specific locations across the three tendons (Red: Duct 31, Green: Duct 33, Blue: Duct 34), see Figure 45, were selected for detailed human inspector analysis, ML analysis, and planned for subsequent destructive verification through core drilling.

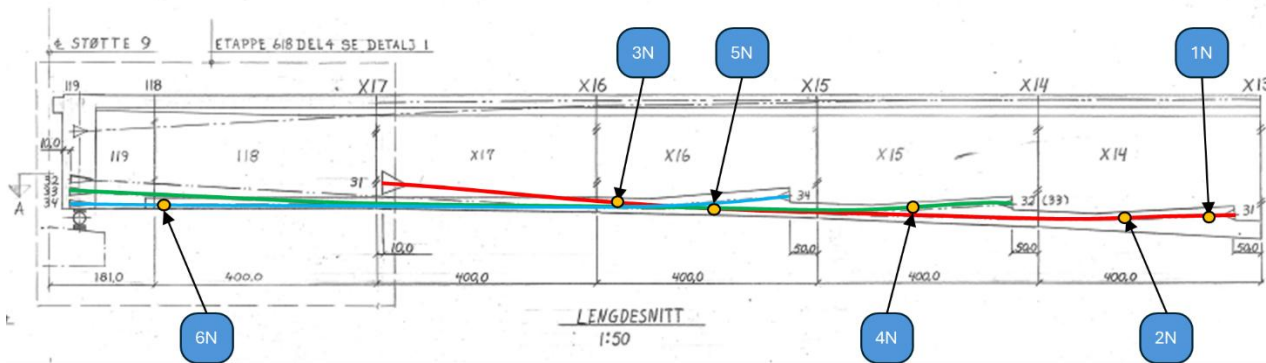


Figure 45 Placement of verification openings

Empty Duct Analysis – Tendon 31:

Location 1N (1.2 m from X13):

Figure 46 presents the ML interface screenshot showing complete analysis results. The SAFT B-scan visualization shows a prominent high-amplitude reflection zone (red/yellow region centered around 100-150 mm depth) indicating a strong concrete-air interface. The Random Forest model analyzing raw I/Q features classifies this segment as **EMPTY with 64.0% confidence** versus 36.0% FILLED confidence. The extracted features panel reveals the physics-based measurements that informed this classification, including coherence coefficients and transmission amplitudes that fall well below the thresholds expected for grout-filled ducts.

Location 2N (1.6 m from X13):

Figure 47 presents the ML interface showing the analysis results. The B-scan displays strong reflections with a localized high-intensity zone indicating void presence. The model classifies this segment as **EMPTY with 54.7% confidence** versus 45.3% FILLED—a lower confidence margin reflecting the somewhat ambiguous signal characteristics in this ASR-affected concrete. Despite the reduced confidence, the model correctly identifies the void condition, demonstrating robustness to material deterioration that complicates signal interpretation.

Location 3N (Close to X16):

Figure 48 shows the ML interface with characteristic void signatures and strong shallow reflections. The Random Forest model classifies this as **EMPTY with 56.7% confidence** versus 43.3% FILLED. Across all three verification locations on Tendon 31, the ML consistently predicts EMPTY despite confidence scores in the 54-64% range—lower than the 80%+ confidences typically achieved on pristine laboratory specimens, but still correctly identifying the void condition under challenging field conditions with ASR-induced signal degradation.

Filled Duct Analysis – Tendons 33 and 34:

Location 4N (Duct 33, 2.0 m from X14):

Figure 49 presents the ML interface for this location on Tendon 33. The SAFT B-scan shows a distinctly different pattern compared to the empty ducts—multiple localized high-amplitude zones representing the individual prestressing strands surrounded by grout, rather than a single continuous void reflection. The Random Forest model classifies this segment as **FILLED with 67.4% confidence** versus 32.6% EMPTY. The extracted features show inter-channel coherence values in the $\gamma > 0.75$ range and sustained transmission amplitudes characteristic of continuous grout paths, supporting the FILLED classification.

Location 5N (Duct 33, 2.0 m from X15):

Figure 50 shows particularly clear grouting conditions at this location. The B-scan displays the characteristic pattern of well-grouted tendons with strong but localized reflections from individual strands. The ML model demonstrates high confidence here, classifying the segment as **FILLED with 89.9% confidence** versus only 10.2% EMPTY. This represents one of the strongest FILLED predictions in the field validation campaign, with coherence analysis showing excellent signal consistency across all channels and minimal void-like scattering.

Location 6N (Duct 34, On X18):

Figure 51 presents the final verification location on Tendon 34. The B-scan again shows multiple discrete reflectors consistent with grouted strand geometry. The model classifies this as **FILLED with 84.1% confidence** versus 15.9% EMPTY. The high confidence reflects clear physical evidence of grout presence through both coherence preservation and appropriate transmission amplitude patterns.

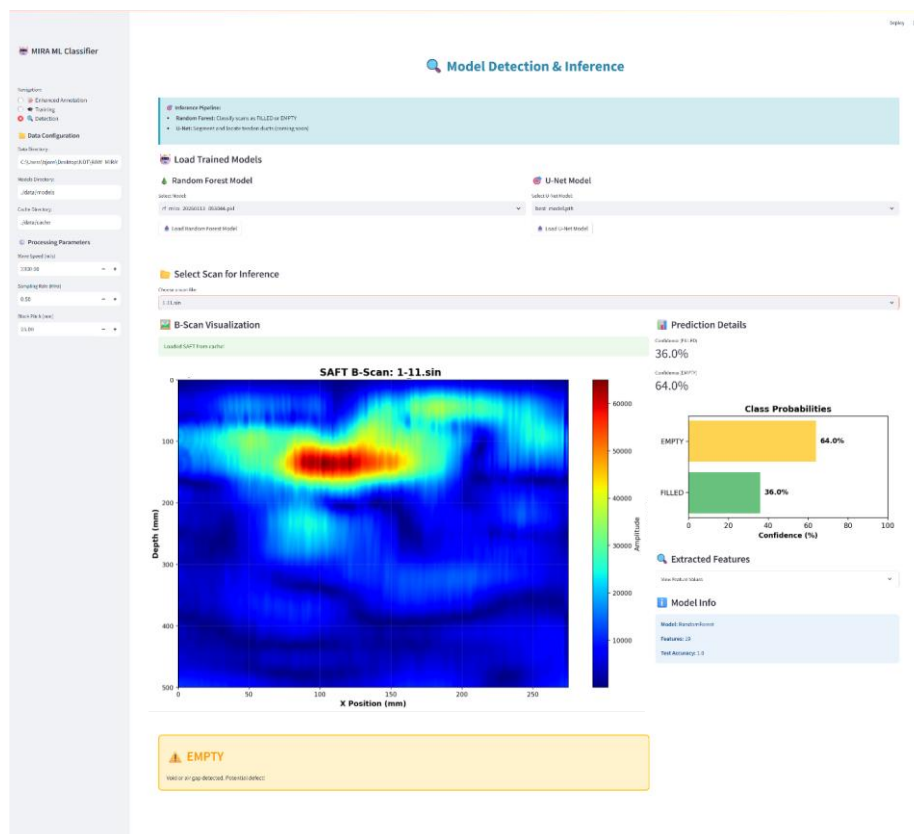


Figure 46 ML classification interface for Opening 1N location (1.2 m from X13)

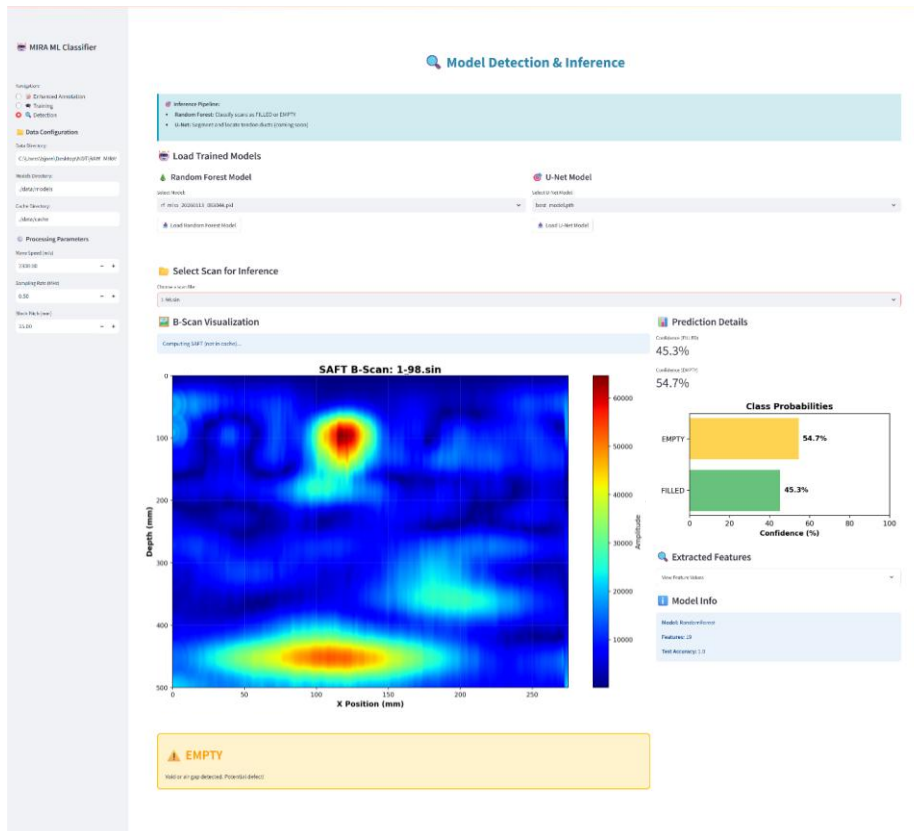


Figure 47 ML classification interface for Opening 2N location (1.6 m from X13)

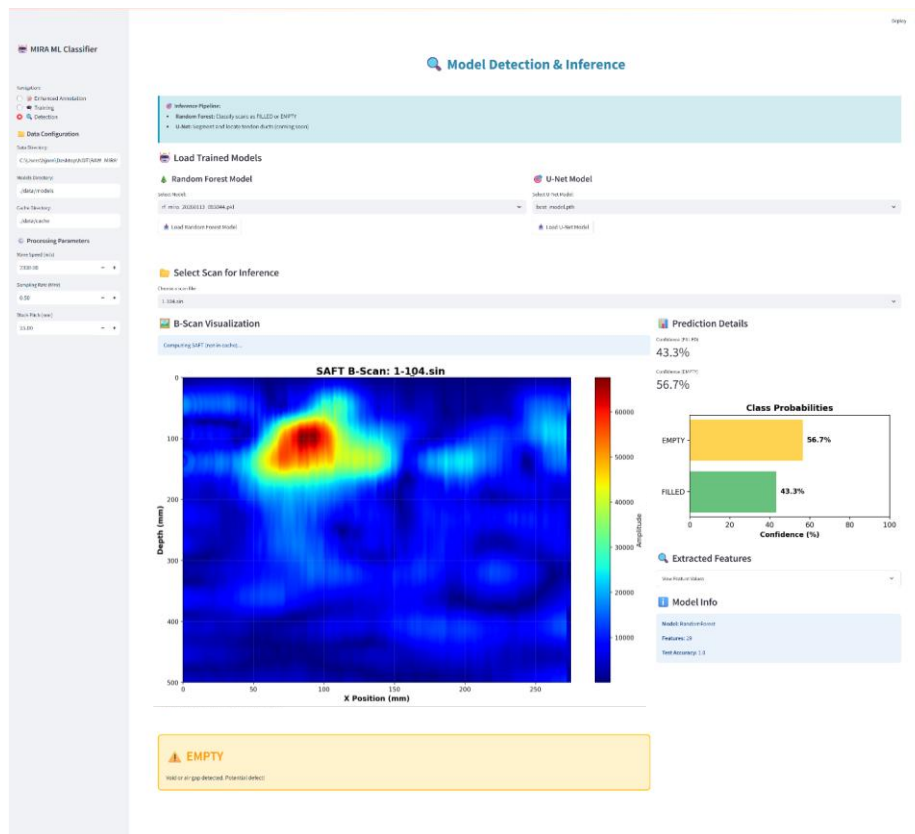


Figure 48 ML classification interface for Opening 3N location (near X16)

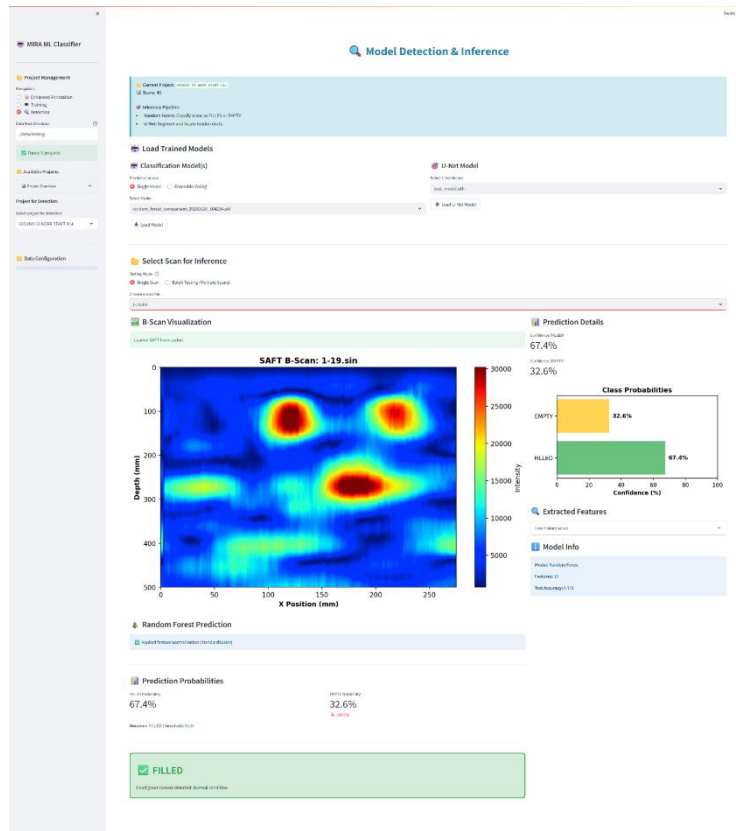


Figure 49 ML classification interface for Opening 4N location

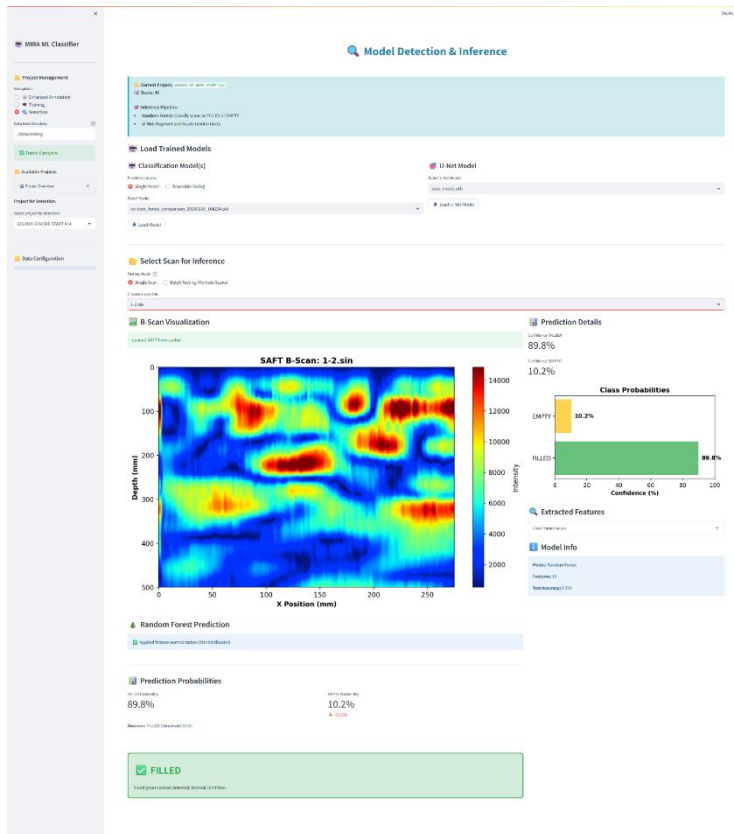


Figure 50 ML classification interface for Opening 5N location

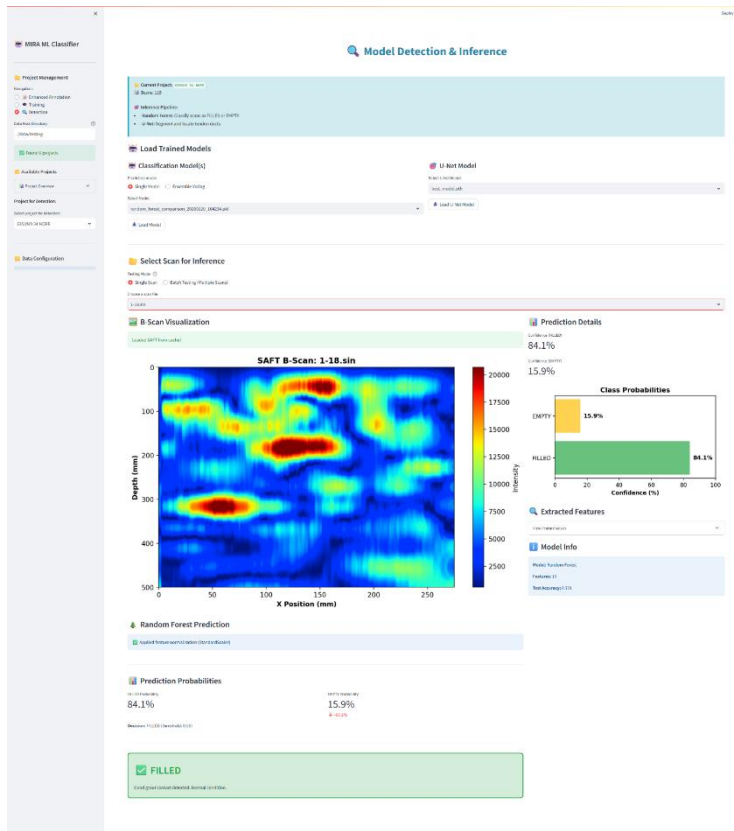


Figure 51 ML classification interface for Opening 6N location

The ML classification operates systematically on each B-scan along the entire tendon duct length:

1. **Raw Data Input:** The system loads the raw 12-channel I/Q complex waveforms directly from MIRA's proprietary .sin/.cos format files, bypassing all commercial visualization processing.
2. **Feature Extraction:** Physics-informed features are computed from the raw signals:
 - **Inter-channel coherence:** Measures spatial consistency of wave transmission across the 48-element transducer array. Filled ducts produce high coherence ($\gamma > 0.85$) as waves transmit uniformly through continuous grout. Voids destroy coherence ($\gamma < 0.6$) through scattered reflections from irregular air-filled boundaries.
 - **Transmission amplitude:** Quantifies energy reaching receivers on the far side of the duct. Grout-filled ducts show substantial transmission despite acoustic losses. Voids produce dramatic amplitude reduction due to near-total reflection at concrete-air interfaces.
 - **Spectral characteristics:** Frequency-domain analysis reveals bandwidth narrowing and centroid shifts associated with void-induced scattering.
3. **Random Forest Classification:** The trained ensemble of 300 decision trees evaluates the extracted feature vector, with each tree voting EMPTY or FILLED based on learned decision rules, see Figure 28 as the random forest classification method. The final classification emerges from aggregated votes, with confidence scores reflecting vote proportions.
4. **Confidence Assessment:** The probability bars visualize classification confidence across all six verification locations:

Empty Ducts (Tendon 31):

- Opening 1N: 64.0% EMPTY (strong agreement across decision trees)
- Opening 2N: 54.7% EMPTY (modest agreement, reflecting signal ambiguity)
- Opening 3N: 56.7% EMPTY (modest agreement)

Filled Ducts (Tendons 33 & 34):

- Opening 4N: 67.4% FILLED (good agreement on grout presence)
- Opening 5N: 89.9% FILLED (very strong agreement, clear grouting)
- Opening 6N: 84.1% FILLED (strong agreement on grout presence)

Flash back to Figure 32, the same Random Forest model analyzed mock-up Specimen, that achieved much higher confidence scores:

- Empty segments: 84.8-90.2% EMPTY confidence
- Filled segments: 78.5-85.3% FILLED confidence

The Gisund Bridge confidence reduction (approximately 20-30 percentage points) quantifies the performance degradation caused by (1) training data collected from Mock Up sample, and (2) ASR microcracking, which introduces distributed scattering that complicates the coherence patterns the model relies upon. Thus, as illustrated by Figure 37, Continuous improvement feedback loop for the NDT-ML system can increase the variety of training data, leading to develop more robust model continuously. However, despite this degradation, the model maintains correct classification, demonstrating the robustness of the physics-informed feature approach.

9.3.3 Verification through Destructive Testing

Due to ASR in the concrete, reliable Impact Echo results could not be obtained. Therefore, destructive testing with core drilling and localized concrete removal was performed to verify both human inspector interpretation and ML classification of ultrasound data.

Figure 52 shows the severe surface cracking characteristic of advanced ASR deterioration observed at the Gisund Bridge. The distributed microcracking throughout the concrete matrix—particularly concentrated in the first 10 cm of depth—created the challenging measurement conditions that degraded both visual interpretation and ML confidence scores.



Figure 52 Surface cracking due to Alkali-Silica Reaction (ASR) observed on the Gisund Bridge

Six openings were created in the north girder using jackhammers and drills, positioned to verify ML classifications at the locations shown in Figure 45. Table 9-1 summarizes the observed grouting status:

Table 9-1 Verification openings in North girder

Opening	Duct	Placement	Observed Status	ML Prediction	Confidence
1N	31 North	1.2 m from X13	Empty	EMPTY	64.0%
2N	31 North	1.6 m from X13	Empty	EMPTY	54.7%
3N	31 North	Close to X16	Empty	EMPTY	56.7%
4N	33 North	2.0 m from X14	Filled	FILLED	67.4%
5N	33 North	2.0 m from X15	Filled	FILLED	89.9%
6N	34 North	On X18	Filled	FILLED	84.1%

Empty Duct Verification: Openings 1N, 2N, and 3N (Figure 53 to Figure 55) confirmed Duct 31 was entirely ungrouted, with strands showing minor surface corrosion but no grout flow. The prestressing strands remained exposed with only protective tape wrapping visible, and the duct interior showed no evidence of grout injection. This aligned perfectly with both human and ML assessments of the UPE data, where the RF model flagged all three segments as Empty.

Filled Duct Verification: Openings 4N, 5N, and 6N (Figure 56 and Figure 57) revealed fully grouted conditions in Ducts 33 and 34. The hardened grout completely surrounded the prestressing strands with good bonding to the duct interior, confirming the tendons were properly grouted. Individual strands were embedded in continuous grout matrix with no visible voids or segregation. These observations validated both human inspector and ML classifications identifying these segments as Filled, demonstrating the model's capability to correctly distinguish grouted from ungrouted conditions under field conditions with ASR-induced material degradation.



Figure 53 Opening 1N – Images from Duct 31N near anchor in X13; Left: View toward the anchor; right: Exposed strands



Figure 54 Opening 2N – Images of Duct 31N; Photos showing tape around strands and empty duct interior



Figure 55 Opening 3N – Images of Duct 31N; Views of spreading cables entering the anchor, confirming void status



Figure 56 Opening 4N (left) and 6N (right) - Images of duct 33N with grout

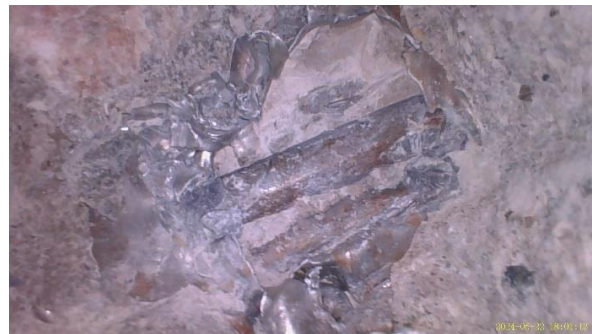


Figure 57 Opening 7N - Image of duct 34 with grout

Comprehensive Verification Comparison:

Figure 58 to Figure 60 synthesize the verification results, comparing human inspector interpretations with ML classifications at all verification locations. These composite visualizations overlay the D-scan longitudinal profiles showing complete tendon paths with detailed B-scan views and classification results from both human and machine inspection at each verification point.

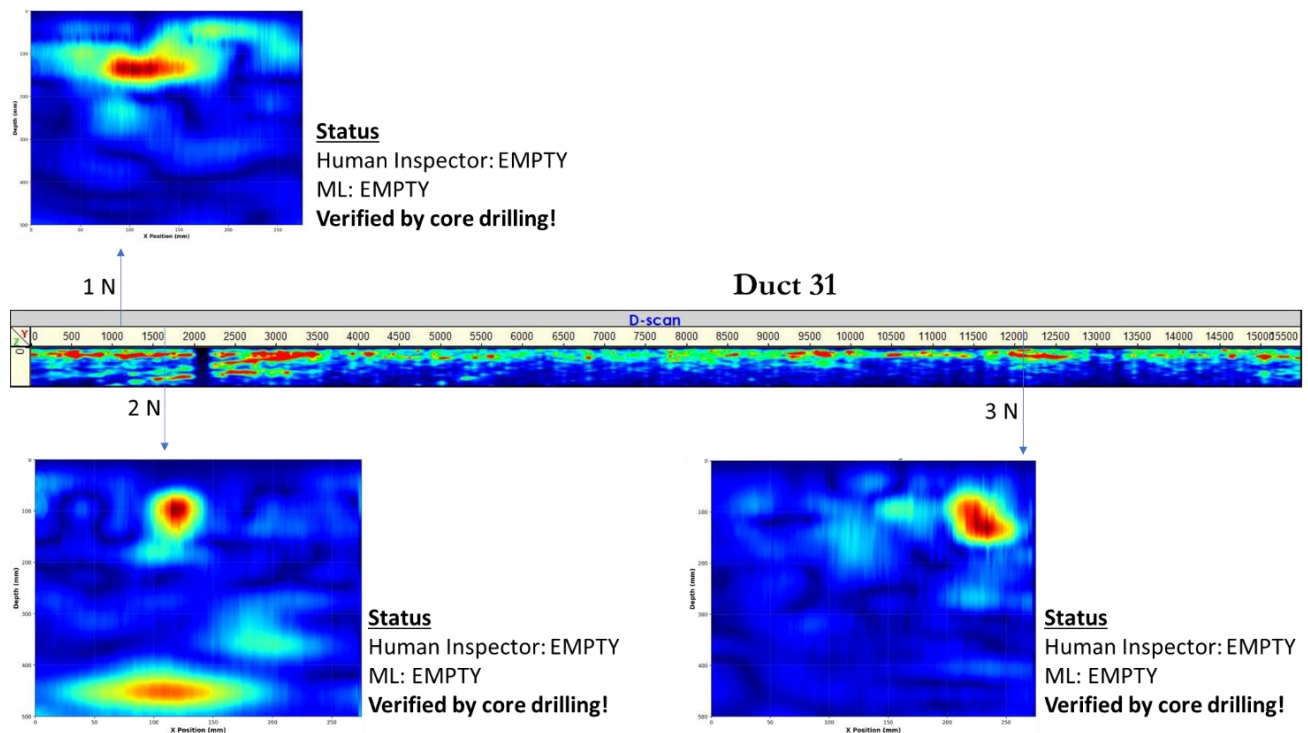


Figure 58 Comprehensive comparison of human inspector and ML classifications at verification drilling locations. Top: D-scan showing complete Tendon 31 profile with verification opening positions marked

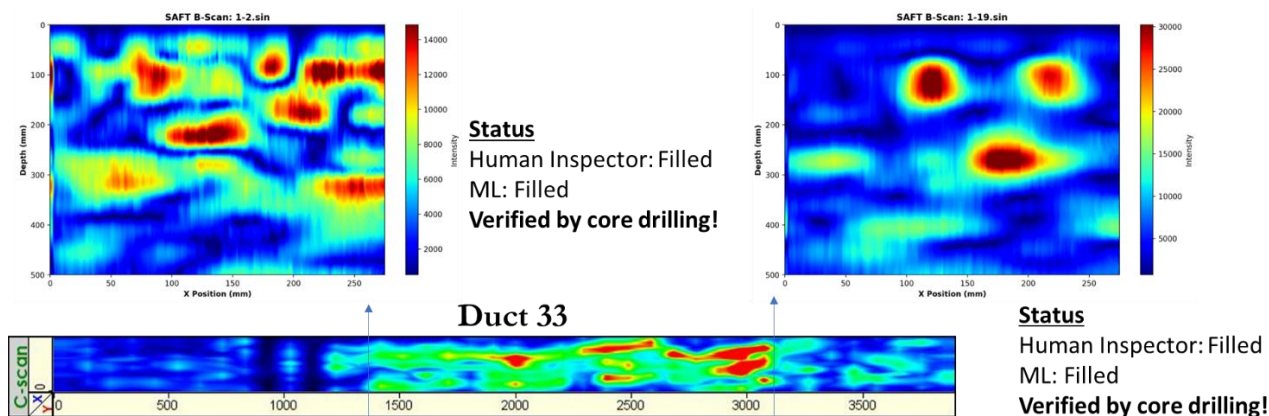


Figure 59 Comprehensive comparison of human inspector and ML classifications at verification drilling locations. Top: D-scan showing complete Tendon 33 profile with verification opening positions marked

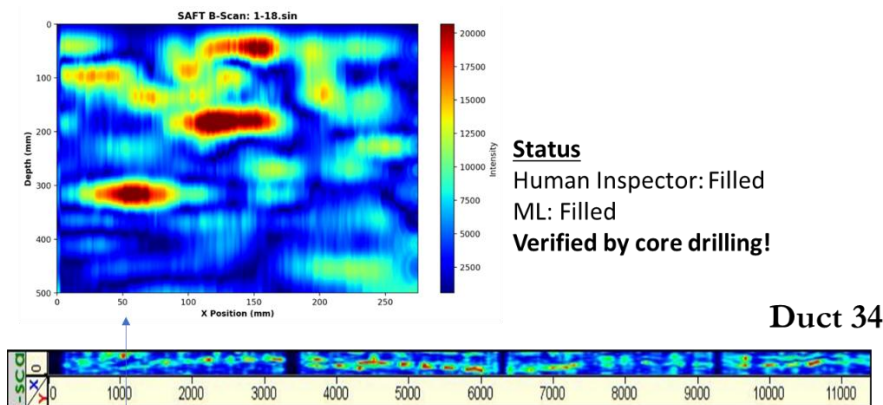


Figure 60 Comprehensive comparison of human inspector and ML classifications at verification drilling locations. Top: D-scan showing complete Tendon 34 profile with verification opening positions marked

9.4 Key Validation Outcomes

1. Perfect Agreement Across Both Conditions: Both human inspectors and the ML model correctly identified EMPTY or FILLED conditions at all six verification openings (Table 9-1). This validates the performance of ML application for both void detection and filled duct confirmation under challenging field conditions. Since the model requires fine-tuning with additional field data from diverse concrete type/surface, these validation results from a model trained on limited laboratory datasets demonstrate promising generalization capability and provide a strong foundation for continued development.

2. Spatial Localization and Longitudinal Pattern Analysis: The current ML implementation analyses the entire MIRA scan region rather than isolating the duct area specifically. This means defects in the surrounding concrete—ASR-induced microcracking, delaminations, casting defects, cold joints, or repair interfaces—can generate ultrasonic reflections interpreted as duct voids. Spatial masking to isolate only the duct region remains a priority for future development.

However, a practical workaround exists through longitudinal pattern analysis. Since MIRA inspections cover the entire tendon duct length continuously, the classification pattern distinguishes between actual grouting voids and local defects in the surrounding concrete:

- **Series of consecutive EMPTY classifications** along multiple scan positions → Genuine ungrouted tendon duct section
- **Isolated EMPTY classification** surrounded by FILLED predictions → Likely local defects in the surrounding concrete (ASR damage, delamination, repair interface, or casting defect) rather than duct void

This longitudinal context enables inspectors to differentiate duct condition from surrounding concrete condition, providing reliable grouting assessment even without automated spatial masking. The Gisund Bridge validation demonstrates this principle: Duct 31 shows consistent EMPTY classifications across all three verification locations spanning several meters, confirming genuine grouting deficiency rather than isolated concrete damage. Conversely, Ducts 33 and 34 show predominantly FILLED classifications with strong confidence scores, accurately reflecting their grouted condition.

3. Confidence Degradation Quantified: ML confidence scores for empty duct predictions (54-64%) fell approximately 20-30 percentage points below typical laboratory performance (80-95% on Specimen 2 data). This confidence degradation reflects the combined effects of limited training data diversity and ASR-induced signal scattering. Interestingly, filled duct predictions showed more variable confidence levels (67-84%, with one location reaching 89.9%), suggesting that grout presence creates more robust signal patterns that partially overcome ASR complications.

4. Robust Feature-Based Classification: Despite reduced confidence on empty ducts, the ML model's physics-informed features—particularly coherence analysis—maintained correct classification across all conditions. Empty duct coherence coefficients consistently fell in the void range ($\gamma < 0.6$), while filled duct locations showed elevated coherence ($\gamma > 0.75$ -0.85), providing clear physical evidence supporting the classifications.

5. Validation of Raw Data Approach: The success under ASR-affected conditions validates the raw-data processing philosophy. Commercial SAFT visualizations were severely degraded by scattering artifacts, creating ambiguous interpretation challenges for human inspectors. Yet the ML model analyzing raw I/Q coherence successfully distinguished ASR-induced distributed scattering (which preserves coherence across channels) from void-induced reflections (which destroy coherence at specific duct locations).

6. Field Deployment Readiness: The successful real-world validation demonstrates that the system transitions effectively from laboratory proof-of-concept to practical bridge assessment. Correct classification accuracy across both empty and filled conditions, despite challenging ASR degradation, confirms the technology is ready for field deployment with appropriate inspector training and quality assurance protocols. The feedback loop, Section 8.2.4—where inspectors validate uncertain predictions and these verifications continuously improve model performance—will be essential for adapting to diverse bridge conditions encountered in operational use.

10 Conclusions and future development

This final chapter brings the conducted research together by summarizing what was accomplished, acknowledging what didn't work perfectly, and outlining where the technology should go next. The research successfully demonstrated that machine learning models analysing raw NDT measurements can detect voids in post-tensioned bridge ducts with accuracy rates around 85-90%—performance that matches or exceeds experienced human inspectors while providing consistent, repeatable results. The Gisund Bridge pilot proved the system works on real infrastructure under difficult conditions, not just in controlled laboratory settings.

But the chapter doesn't just celebrate successes. It honestly discusses limitations that remain: the training datasets are still relatively small, the models struggle with concrete that differs significantly from what they were trained on, and practical deployment faces organizational hurdles beyond just technical performance. False alarm rates, while reduced, still require human review. Adapting to unusual bridge designs or severely deteriorated concrete remains challenging. These aren't failures—they're realistic boundaries that define where the technology works well and where it needs further development.

The future development section looks ahead to what comes next: vehicle-mounted scanning systems that could map entire bridge decks in hours instead of days, advanced machine learning techniques that need less labelled training data, and physics-based simulations that could generate synthetic training examples for rare defect types. The closing emphasizes that this technology isn't about replacing engineers—it's about giving them better tools to do their jobs more effectively and safely.

This research demonstrates the effectiveness of integrating ML with multiple NDT modalities—GPR, UPE, and IE—for detecting grouting defects in post-tensioned concrete structures. By performing trained ML models including SVM, RF, and 1D-CNN on raw measurement data rather than processed visualizations, the framework achieves robust performance with F1-scores ranging from 0.84 to 0.90 across modalities on held-out test sets. The pilot validation on the Gisund Bridge confirmed practical utility: ML predictions aligned closely with expert visual inspections and destructive verifications, identifying ungrouted ducts correctly while reducing false positives through confidence-weighted multi-modal fusion.

10.1 Discussion and limitations

10.1.1 Advantages of the raw data approach

Processing NDT signals directly from raw measurements—GPR A-scans, UPE I/Q complex data, IE time-domain responses—offers several compelling advantages over the conventional approach of working with preprocessed visualizations.

Enhanced generalization across measurement conditions

Models trained on raw data learn intrinsic physical signatures rather than artifacts introduced by preprocessing operations like gain application or filtering. The Gisund Bridge deployment demonstrated this clearly: the Random Forest model maintained consistent performance despite substantial variations in transducer coupling conditions and surface roughness that severely degraded visual interpretation of processed B-scans. An inspector examining commercial software's processed images struggled to distinguish genuine defects from coupling artifacts, while the ML model analysing raw signals successfully extracted the underlying physics through the noise.

Complete information preservation

Raw data processing retains phase information, absolute amplitude relationships, and high-frequency details that visualization-oriented processing discards or distorts. These subtle characteristics often provide the clearest signatures of defects. Consider GPR signals: phase relationships at the grout-void interface encode information about impedance contrast that gets completely eliminated during envelope detection and time-varying gain application in commercial processing software. B-scan images show bright and dark bands appealing to human visual interpretation, but the mathematical relationships that distinguish filled from void conditions have been irreversibly destroyed. Raw A-scan analysis preserves everything, enabling more accurate defect characterization.

Physical interpretability

Features extracted from raw data—spectral centroids, coherence coefficients, transmission amplitudes—maintain direct connections to fundamental wave propagation physics. This isn't merely an academic consideration; it's essential for practical deployment. When a model classifies a duct as void or filled, inspectors need to understand why. Unlike end-to-end deep learning approaches that offer limited insight into their reasoning process, physics-informed features can be traced back to first principles. The model predicted a void because inter-channel coherence dropped to 0.42 when theory predicts 0.85+ for filled conditions—that's an explanation an engineer can verify, defend, and trust.

Robustness to material degradation

Field structures rarely match laboratory specimens in their pristine simplicity. The Gisund Bridge case study illustrated this challenge: extensive alkali-silica reaction (ASR) in the concrete matrix introduced severe ultrasonic scattering that completely obscured conventional B-scan interpretations. Inspectors examining processed tomograms saw scattered, noisy images throughout and struggled to distinguish ASR-related scattering from genuine voids. Yet the RF model analysing multi-channel coherence features from raw I/Q data maintained correct classifications at all three verification locations, successfully distinguishing between the two phenomena. ASR affects all channels similarly, reducing absolute amplitudes but preserving coherence patterns. Voids destroy coherence selectively at specific locations. This distinction remains visible in raw multi-channel data but gets lost in single-image visualizations.

10.1.2 Limitations and challenges

Despite the advantages demonstrated, several limitations affect current system performance and applicability. Understanding these limitations guides future development priorities and sets realistic expectations.

A. Data and model limitations

The datasets compiled for this study, while sufficient for proof-of-concept demonstration, remain small by contemporary machine learning standards, introducing class imbalance (voids <20% of measurements). Training data originates from a limited number of Mock up samples, potentially encoding lab-specific characteristics. The classical ML approaches—SVMs and Random Forests—demonstrate strong performance but depend on manual feature engineering requiring deep domain expertise. Ground truth labeling carries inherent ambiguity, as reality offers a continuum from fully grouted to completely void rather than binary categories.

B. Generalization challenges

Performance metrics reflect model capabilities on concrete structures sharing reasonably similar characteristics with the training set. Confidence in generalization decreases when confronting markedly different conditions: varying aggregate mineralogy, water-cement ratios, concrete age, moisture content, reinforcement layouts, and tendon duct geometries (50-150mm diameter, steel vs. HDPE material, 10-50cm depth). The Gisund Bridge illustrated these challenges: ASR-induced microcracking degraded signal-to-noise ratios substantially, requiring the model to distinguish distributed scattering from localized voids through coherence analysis.

C. Practical deployment challenges

Field deployment demands rigorous data acquisition protocols because measurement variability directly propagates to model performance. Inspectors need training not just in equipment operation but in recognizing when poor coupling produces unreliable data requiring remeasurement. False alarm rates, though reduced from 11-13% to approximately 5.6% through multi-modal fusion, still necessitate human review of flagged locations. Finally, practical deployment confronts organizational barriers: established proprietary formats, legacy reporting requirements, and risk-averse cultures favouring proven methods. Overcoming this resistance requires building trust through pilot deployments, clear documentation of capabilities and limitations, and personnel training.

Despite these challenges, results from this project show that the developed AI tool considerably supports inspectors and increases the probability of avoiding false negatives. The system is planned to be used in Invator's assessment procedures of concrete structures, facilitating additional data collection for training and continuous improvement of the developed model.

10.2 Opportunities for future development

10.2.1 Immediate priority: Invator implementation and inspector training

The most critical near-term development involves transitioning the ML-enhanced NDT framework from research prototype to operational tool within Invator's inspection workflow. This requires systematic integration of the web-based ML interface with existing practices, comprehensive inspector training, and establishment of quality assurance protocols.

Current workflow integration

The ML system serves as a **decision-support tool** providing quantitative second opinions on NDT measurements, not as a replacement for inspector judgment. The implementation strategy positions the ML interface to complement rather than disrupt existing workflows:

Standard Invator inspection, current workflow:

1. Site preparation and equipment setup
2. GPR area scanning for duct/rebar mapping
3. MIRA ultrasonic scanning at targeted locations
4. Impact Echo grid measurements (when applicable)
5. Visual interpretation of processed B-scans/tomograms by inspector
6. Manual documentation and reporting
7. Client deliverables with inspector findings

Enhanced workflow with ML integration:

1. Site preparation and equipment setup (unchanged)
2. GPR area scanning → raw data automatically uploaded to ML interface
3. MIRA ultrasonic scanning → raw .sin/.cos files uploaded alongside
4. Impact Echo grid measurements → raw time-domain data uploaded
5. ML analysis runs in parallel: Interface displays classification results with confidence scores
6. Inspector reviews both traditional visualizations AND ML predictions
7. Decision protocol:
 - Green segments (ML: filled, confidence >80%) → Accept, routine monitoring
 - Red segments (ML: void, confidence >80%) → Prioritize for detailed review
 - Yellow segments (uncertain, confidence 50-80%) → Inspector applies expert judgment
 - Disagreement between ML and inspector → Inspector documents reasoning
8. Enhanced reporting with ML confidence metrics alongside traditional findings
9. Feedback loop: Inspector validates uncertain predictions, system improves over time

Inspector training program

Successful deployment requires training inspectors to understand ML capabilities, interpret confidence scores, and integrate predictions into decision-making. The training program comprises three levels:

Level 1: ML Fundamentals

- What is machine learning and how does it work?
- Understanding Random Forest, SVM, and 1D-CNN approaches
- Physics-informed features: coherence, amplitude, spectral characteristics
- Why raw data analysis outperforms processed visualizations
- AUC, F1-score, and confidence metrics explained
- Limitations: when to trust ML, when to rely on experience

Level 2: Interface Operation

- Uploading raw data files (SEG-Y, MIRA .sin/.cos, IE .dat)
- Navigating the web-based interface
- Interpreting B-scan visualizations and ML predictions
- Understanding probability bars and confidence scores
- Extracted features panel: what the model "sees"
- 3D visualization: color-coded segments and interactive inspection
- Generating reports with ML annotations
- Practical exercises using mock-up specimen data

Level 3: Field Application

- Data acquisition best practices for ML analysis
- Quality control: recognizing poor coupling, artifacts, noise
- Real-time ML feedback during inspections
- Decision protocols: accepting, questioning, or overriding ML predictions
- Documenting disagreements and rationale
- Feedback loop participation: validating uncertain segments
- Case studies: Gisund Bridge and other pilot projects
- Troubleshooting common issues

Web-based interface enhancement roadmap

The current interface provides core functionality but requires several improvements for optimal field deployment:

Current capabilities:

- Upload and process GPR, MIRA, IE raw data
- Display SAFT B-scans with ML classifications
- Show confidence scores and probability bars
- Present extracted features (coherence, amplitude, etc.)
- Generate 3D visualizations with color-coded segments

Phase 1 improvements (0-6 months):

- **Mobile-responsive design** for tablet use in field (scaffolding, cramped spaces)
- **Offline mode** for remote sites without reliable internet
- **Batch processing** for multiple scan lines simultaneously
- **Integrated positioning** to auto-align GPR/MIRA/IE using metadata
- **Comparison view** to display inspector interpretation alongside ML prediction
- **Quick feedback buttons** for validating/correcting predictions (feeds training loop)

Phase 2 improvements (6-12 months):

- **Uncertainty visualization** showing which features drive uncertain predictions
- **Historical comparison** tracking same structure over multiple inspections
- **Client portal** for bridge owners to view results with simplified explanations
- **Integration with Invator's existing reporting system** (auto-populate templates)
- **LLM chatbot support** for interactive communication with database

10.2.2 Machine learning model improvements

Beyond practical implementation, several ML advancements offer performance improvements as larger datasets accumulate through routine field use.

Transfer learning and domain adaptation: Pre-training deep learning models on diverse NDT datasets could provide robust feature representations that generalize across varied concrete types and structural configurations. Fine-tuning on site-specific data would require substantially fewer labelled examples than training from scratch—potentially reducing annotation burden from hundreds to dozens of samples per structure.

Semi-supervised and active learning: The vast majority of NDT data collected during routine inspections remains unlabelled. Semi-supervised learning techniques can leverage this abundant unlabelled data to improve generalization. Active learning complements this by having the model identify the most informative unlabelled examples—those with high prediction uncertainty—and requesting human expert labels only on these cases, optimizing information gained per labelled sample.

Synthetic training data generation: Finite-difference time-domain (FDTD) or finite-element method (FEM) simulations of electromagnetic and acoustic wave propagation can generate realistic synthetic GPR, MIRA, and IE signals. By systematically varying simulation parameters—concrete dielectric properties, elastic moduli, void geometry, depth—thousands of synthetic training examples can be created representing defect scenarios rarely encountered in field data, particularly rare but catastrophic failure modes like complete tendon rupture.

10.2.3 Automated data collection (longer-term development)

Looking further ahead, transitioning from handheld instruments to automated, vehicle-mounted systems represents a significant evolution enabling comprehensive structural mapping at dramatically accelerated speeds.

Vehicle-mounted GPR: Purpose-built inspection vehicles equipped with GPR antenna arrays operating at multiple frequencies (1.6 GHz for shallow features, 400 MHz for deeper penetration) can systematically survey bridge decks considerably faster than handheld inspection. Integration with RTK-GPS enables precise georeferencing to within 1-2 cm accuracy, supporting temporal change detection by comparing surveys conducted months or years apart. Beyond void detection, the ML framework offers additional capabilities: automated hyperbola fitting on detected rebars enables calculation of site-specific dielectric constants, improving depth estimation accuracy beyond the standard assumption ($\epsilon_r \approx 6-9$ for concrete). This calibrated dielectric constant can subsequently be correlated with material properties such as chloride content and moisture levels, providing a comprehensive structural assessment from a single GPR survey.

Risk-based inspection strategy: Automated GPR surveys generate comprehensive structural maps annotated with ML-derived anomaly scores, enabling intelligent deployment of time-intensive MIRA and IE measurements only at high-priority zones. This fundamentally alters inspection strategy from exhaustive grid-based sampling to focused investigation of ML-identified critical areas.

10.3 Market possibilities

The outcomes of this project create market possibilities within the inspection, assessment, and management of concrete infrastructure, particularly post-tensioned assets where internal defects cannot be visually verified. The developed framework combines multi-modal NDT measurements—Ground Penetrating Radar, Ultrasonic Pulse Echo, and Impact-Echo—with machine learning analysis of raw data to detect and classify voids and grouting deficiencies. This addresses a defined limitation in current practice, where interpretation depends heavily on individual operator experience and processed imagery rather than physics-based signal analysis.

The primary market consists of infrastructure owners and asset managers responsible for prestressed bridge inventories. Transport administrations, railway agencies, and municipalities manage aging structures where tendon duct voids pose durability and safety risks. Current inspection outputs are costly and time consuming and often lead to conservative decisions due to uncertainty. ML-assisted classification enables more consistent defect identification and supports risk-based maintenance planning. This influences intervention timing, repair budgeting, and prioritization across bridge portfolios.

Engineering consultancies and inspection firms represent the second major segment. These organizations already perform NDT investigations but rely on manual interpretation and manufacturer software. The developed system can be integrated as an advanced analysis layer within existing workflows. Commercial delivery may include software licensing, project-based analytics services, or cloud processing subscriptions. The value proposition is reduced interpretation time, standardized assessments across teams, and the ability to support less experienced inspectors with data-driven guidance.

A parallel market exists with NDT equipment manufacturers. Hardware suppliers seek differentiation through software capability rather than sensor hardware alone. The project outputs—feature extraction routines, trained classification models, and multi-modal fusion logic—can be embedded into proprietary analysis platforms or deployed as post-processing modules. Integration at acquisition stage could enable real-time defect probability indicators during scanning. Commercial structures may include OEM licensing or co-development agreements.

Repair contractors and strengthening designers also benefit commercially. Accurate void detection informs repair scope definition, targeted grout re-injection, and tendon replacement strategies. Improved pre-intervention diagnostics reduce contingency pricing in tenders and limit dispute risk between contractors and asset owners. Post-repair verification using the same analytical framework provides documented quality assurance.

This supports lifecycle modelling, deterioration forecasting, and investment planning rather than one-time inspection reporting.

Training and knowledge transfer form an additional service market. Because the system functions as decision support rather than full automation, structured training is required. Mock-up datasets, validated workflows, and interpretation guidelines developed in the project can be converted into certification courses or professional training modules.

The methodology is transferable beyond post-tensioned bridges. Comparable applications include tunnels, hydropower structures, marine quays, nuclear facilities, and large prestressed buildings. Any asset where internal voids, grout defects, or delamination influence durability can adopt similar analytics, provided NDT data is available.

Geographically, deployment potential aligns with regions holding large prestressed infrastructure inventories. Adoption depends on regulatory acceptance of ML-assisted interpretation and integration with established inspection standards. Initial commercialization is most feasible where multi-modal NDT inspections are already routine.

10.4 Closing remarks

This research demonstrates that ML applied to raw NDT signals can achieve detection performance exceeding traditional human interpretation while maintaining the physical interpretability and robustness required for safety-critical infrastructure inspection. The successful validation on the Gisund Bridge—a challenging operational structure exhibiting complex material behaviour due to alkali-silica reaction—provides confident potential that the methodology generalizes to diverse bridge inventories.

For Invator, the immediate priority involves transitioning this research prototype into an operational decision-support tool through systematic inspector training, interface refinement, and quality assurance protocols. The feedback loop built into the system—where inspectors validate uncertain predictions and these verifications continuously improve model performance—ensures the technology evolves through real-world use rather than remaining static.

The longer-term vision involves automated data collection and advanced ML techniques, but the foundation must first be established through successful integration with current inspection practices. Early adoption of this technology shows Invator's commitment to improving bridge safety while keeping engineers at the centre. It gives experienced inspectors better tools that help them detect problems they might miss, speed up routine inspections, and free up time for complex decisions that require deep expertise.

The future of bridge inspection isn't about replacing engineers with computers—it's about combining human expertise with machine learning. Together, they can achieve reliability that neither could reach alone. This research demonstrates that vision in practice, providing better tools for keeping infrastructure safe.

References

- Alqurashi, I., Alsulami, M., Alver, N., & Catbas, N. (2025). Ultrasonic tomography with deep learning for detecting embedded components and internal damage of concrete structures. *Developments in the Built Environment*, 23, 100742. <https://doi.org/10.1016/j.dibe.2025.100742>
- Belhadj, A., & Waldron, P. (1993). Investigation into cutting of bonded prestressing bars during demolition. *Proceedings of the Third International RILEM Symposium on Demolition and Reuse of Concrete and Masonry*, 295–306.
- Bore, T., Taillade, F., Placko, D., & Himbert, M. (2010). CAPACITIVE PROBE FOR NON DESTRUCTIVE INSPECTION OF EXTERNAL POST-TENSIONED DUCTS: MODELLING BY DPSM TECHNIQUE. *AIP Conference Proceedings*, 1211(4), 1983–1990. <https://doi.org/10.1063/1.3362353>
- Buchner, S., & Lindsell, P. (2004). Controlled Demolition of Prestressed Concrete Structures. In *Demolition methods and practice* (p. 10). CRC Press.
- Carsana, M., & Bertolini, L. (2015). Corrosion failure of post-tensioning tendons in alkaline and chloride-free segregated grout: A case study. *Structure and Infrastructure Engineering*, 11(3), 402–411. <https://doi.org/10.1080/15732479.2014.887736>
- Cavell, D. G., & Waldron, P. (2001). A residual strength model for deteriorating post-tensioned concrete bridges. *Computers & Structures*, 79(4), 361–373. [https://doi.org/10.1016/S0045-7949\(00\)00150-4](https://doi.org/10.1016/S0045-7949(00)00150-4)
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Dai, L., Chen, Y., Wang, L., & Ma, Y. (2021). Secondary anchorage and residual prestressing force in locally corroded PT beams after strand fracture. *Construction and Building Materials*, 275, 122137. <https://doi.org/10.1016/j.conbuildmat.2020.122137>
- Datta Rachuri, R., Liao, D., Sarikonda, S., & Kondur, D. V. (2024). A Multimodal Fusion Framework for Bridge Defect Detection with Cross-Verification. *2024 IEEE International Conference on Big Data (BigData)*, 3292–3300. <https://doi.org/10.1109/BigData62323.2024.10825867>
- Dethof, F., & Keßler, S. (2024). Design of Concrete Mock-Ups with Complex Defect Scenarios Using Numerical Simulations. *Journal of Nondestructive Evaluation*, 43(2), 59. <https://doi.org/10.1007/s10921-024-01074-9>
- Felstead, A. E., & LINDSELL, P. (1981). Controlled demolition of a post-tensioned beam. *Transport Research Laboratory*, 15(9), 20–23.
- Giannini, E., R., Rivard, P., Kodjo, S. A., Wood, S. G., & Guimaraes, M. (2020). DEVELOPMENT AND TESTING OF NDE MOCKUP SPECIMENS WITH ASR. *ICAAR 2020 - 16th International Conference on Alkali Aggregate Reaction in Concrete*. ICAAR 2020 - 16th International Conference on Alkali Aggregate Reaction in Concrete, Lisboa, Portugal.
- Hansen, B. (2007). Tendon Failure Raises Questions About Grout in Posttensioned Bridges. *MnDOT Post-Tensioned Bridge Preservation. Presentation, National Bridge Preservation Conference.*, 77(11).
- Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, 278–282 vol.1. <https://doi.org/10.1109/ICDAR.1995.598994>
- Holmqvist, M., Popescu, C., O'Sadnick, M., & Täljsten, B. (2024). *Special Inspection of PT – ducts on Gisund Bridge* (TEchnical Report 2024:01252; p. 47). Sintef Narvik AS.

- Im, S. B., & Hurllebaus, S. (2012). Non-destructive testing methods to identify voids in external post-tensioned tendons. *KSCE Journal of Civil Engineering*, 16(3), 388–397. <https://doi.org/10.1007/s12205-012-1295-0>
- Kang, M., Kaji, S., Lee, S.-Y., Kim, T., Ryu, H.-H., & Choi, S. (2025). *A Novel Shape-Aware Topological Representation for GPR Data with DNN Integration* (arXiv:2506.06311). arXiv. <https://doi.org/10.48550/arXiv.2506.06311>
- Kessler, S., & and Grosse, C. U. (2023). Hands-On Training of Non-destructive Testing Using a Mock-Up in the Curriculum of Civil Engineers. *Research in Nondestructive Evaluation*, 34(5–6), 205–221. <https://doi.org/10.1080/09349847.2023.2255538>
- Krause, M., Milmann, B., Mielentz, F., Streicher, D., Redmer, B., Mayer, K., Langenberg, K.-J., & Schickert, M. (2008). Ultrasonic Imaging Methods for Investigation of Post-tensioned Concrete Structures: A Study of Interfaces at Artificial Grouting Faults and Its Verification. *Journal of Nondestructive Evaluation*, 27(1), 67–82. <https://doi.org/10.1007/s10921-008-0033-5>
- Lau, K., Lasar, I., & Paredes, M. (2013). *Corrosion Failure of Post-Tensioned Tendons in Presence of Deficient Grout*. NACE-2013-2600.
- Lee, S., & Popovics, J. S. (2025). Ultrasonic defect detection in a concrete slab assisted by physics-informed neural networks. *NDT & E International*, 151, 103311. <https://doi.org/10.1016/j.ndteint.2024.103311>
- Liu, P., Ding, Z., Zhang, W., Ren, Z., & Yang, X. (2023). Using Ground-Penetrating Radar and Deep Learning to Rapidly Detect Voids and Rebar Defects in Linings. *Sustainability*, 15(15), 11855. <https://doi.org/10.3390/su151511855>
- Martin, J., Broughton, K. J., Giannopolous, A., Hardy, M. S. A., & Forde, M. C. (2001). Ultrasonic tomography of grouted duct post-tensioned reinforced concrete bridge beams. *NDT & E International*, 34(2), 107–113. [https://doi.org/10.1016/S0963-8695\(00\)00035-9](https://doi.org/10.1016/S0963-8695(00)00035-9)
- Menga, A., Kanstad, T., Cantero, D., Bathen, L., Hornbostel, K., & Klausen, A. (2023). Corrosion-induced damages and failures of posttensioned bridges: A literature review. *Structural Concrete*, 24(1), 84–99. <https://doi.org/10.1002/suco.202200297>
- Minh, H., Mutsuyoshi, H., & Niitani, K. (2007). Influence of grouting condition on crack and load-carrying capacity of post-tensioned concrete beam due to chloride-induced corrosion. *Construction and Building Materials*, 21(7), 1568–1575. <https://doi.org/10.1016/j.conbuildmat.2005.10.004>
- Nürnberg, U. (2002). Corrosion induced failure mechanisms of prestressing steel. *Materials and Corrosion*, 53(8), 591–601. [https://doi.org/10.1002/1521-4176\(200208\)53:8<591::AID-MACO591>3.0.CO;2-X](https://doi.org/10.1002/1521-4176(200208)53:8<591::AID-MACO591>3.0.CO;2-X)
- Oh, R.-O., Kim, H.-H., Choo, Y.-J., Park, S.-K., Rodrigazo, S. A., Yeon, J., & Park, C. G. (2024). Evaluation of Delaminations and Defects in Concrete Deck Using Non-Destructive Multi-Physical Scanning Technology. *Sustainability*, 16(21). <https://doi.org/10.3390/su16219225>
- Popescu, C., & Täljsten, B. (2020). *Design of mock-up specimen for void detection in tendon ducts*. Internal report (unpublished).
- Popescu, C., Täljsten, B., Holmqvist, M., Langås, I., & Antonsen, R. (2024). *Condition survey of grouted tendons of a prestressed concrete girder bridge* (Technical Report 2024:01598; p. 37). Sintef Narvik AS.
- Razak, N. A., Senin, S. F., & Hamid, R. (2015). Detection of Sizes and Locations Air Voids in Reinforced Concrete Slab using Ground Penetrating Radar and Impact-Echo Methods. *Jurnal Teknologi (Sciences & Engineering)*, 74(3). <https://doi.org/10.11113/jt.v74.4553>

- Sayyar-Roudsari, S., Shalbaftabar, A., Damirchilo, F., Taslimian, R., Abu-Lebdeh, T., Hamoush, S., & Yi, S. (2024). Defect Classification of Reinforced Concrete Structures with Nondestructive Tests Using Statistical and Machine Learning Methods. *American Journal of Engineering and Applied Sciences*, 17(3), 102–115. <https://doi.org/10.3844/ajeassp.2024.102.115>
- Sharma, S., Osmolska, M. J., Hornbostel, K., Bathen, L., Novakova, I., & Geiker, M. (2024). Repair Options for Corrosion-damaged Prestressed Concrete Bridges. *Nordic Concrete Research*, 70(1), 99–123. <https://doi.org/10.2478/ncr-2024-0004>
- Taylor-Phillips, S., & Stinton, C. (2019). Fatigue in radiology: A fertile area for future research. *British Journal of Radiology*, 92(1099), 20190043. <https://doi.org/10.1259/bjr.20190043>
- Tinkey, Y., & Olson, L. D. (2007). Sensitivity Studies of Grout Defects in Posttensioned Bridge Ducts Using Impact Echo Scanning Method. *Transportation Research Record*, 2028(1), 154–162. <https://doi.org/10.3141/2028-17>
- Trejo, D., Been Im, S., Pillai, R. G., Hueste, M. B. D., Gardoni, P., Hurlebaus, S., & Gamble, M. (2009). *Effect of voids in grouted post-tensioned concrete bridge construction: Inspection and repair manual for external tendons in segmental, post-tensioned bridges* (Technical Report FHWA/TX-09/0-4588-2; p. 62). Texas Transportation Institute The Texas A&M University System College Station, Texas 77843-3135. <http://tti.tamu.edu/documents/0-4588-2.pdf>
- Wang, L., Zhang, X., Zhang, J., Ma, Y., Xiang, Y., & Liu, Y. (2014). Effect of insufficient grouting and strand corrosion on flexural behavior of PC beams. *Construction and Building Materials*, 53, 213–224. <https://doi.org/10.1016/j.conbuildmat.2013.11.069>
- Wiggenhauser, H., Wöstmann, J., Schulze, S., Barry, K., Guimaraes, M., Scott, D., Lindberg, J., & L'Areau, J. P. (2015). Non-destructive evaluation of steel-concrete mock-ups. *10th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurized Components*, 969–975.
- Xu, X., Ran, B., Jiang, N., Xu, L., Huan, P., Zhang, X., & Li, Z. (2024). A systematic review of ultrasonic techniques for defects detection in construction and building materials. *Measurement*, 226, 114181. <https://doi.org/10.1016/j.measurement.2024.114181>
- Zhang, J.-K., Yan, W., & Cui, D.-M. (2016). Concrete Condition Assessment Using Impact-Echo Method and Extreme Learning Machines. *Sensors*, 16(4), 447. <https://doi.org/10.3390/s16040447>
- Zhang, J.-Y., Huang, L., & Guan, Y.-J. (2025). Real-time defect detection in concrete structures using attention-based deep learning and GPR imaging. *Scientific Reports*, 15(1), 35507. <https://doi.org/10.1038/s41598-025-19596-1>

Appendix - A

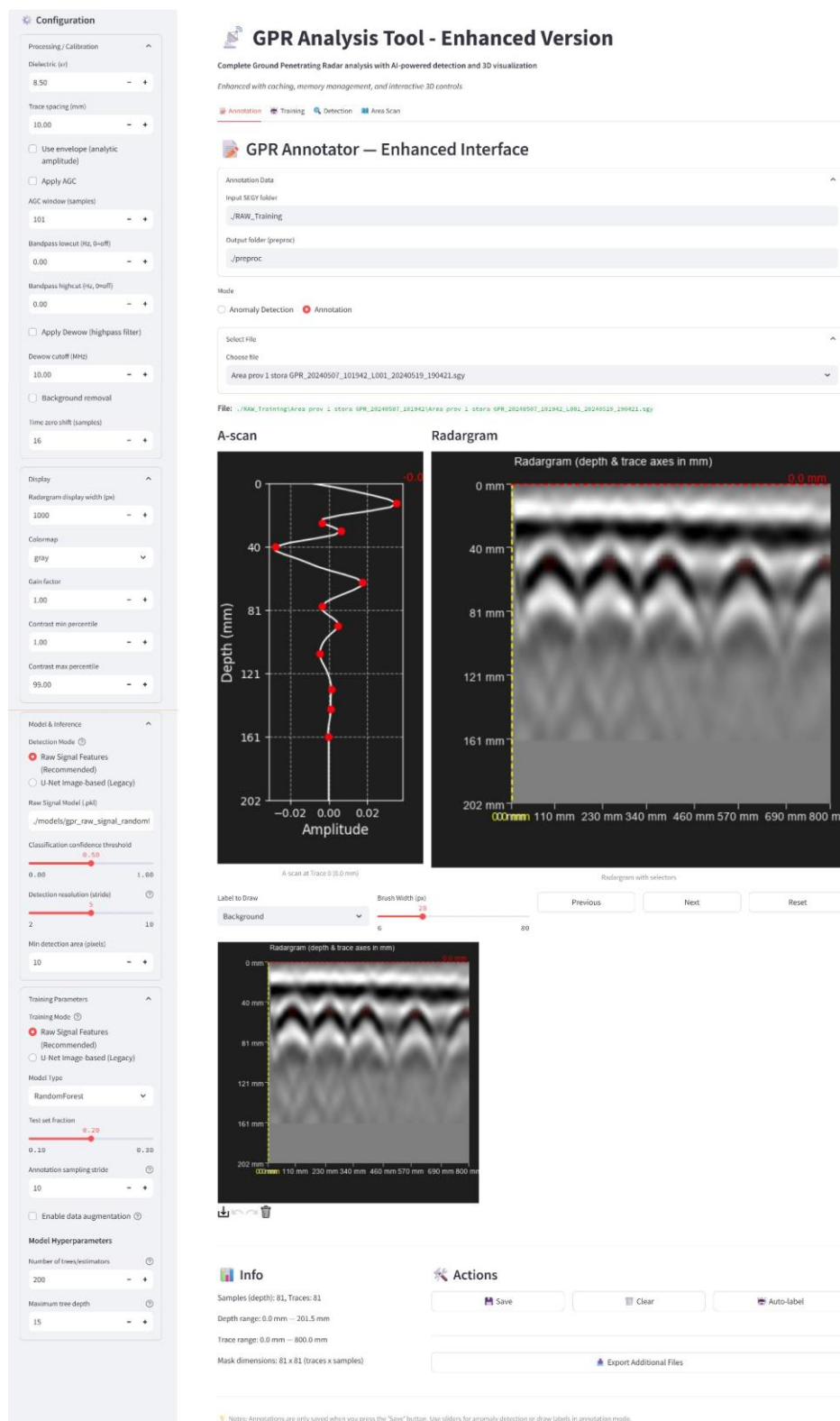


Figure A-1: Data annotation interface for GPR; Web-based annotation interface for labeling GPR A-scan data. The interface allows inspectors to mark rebar reflections (red), and tendon duct hyperbolas (blue), directly on radargrams but annotated on RAW data. Labeled data feeds both Random Forest and U-Net training pipelines for automated duct detection.

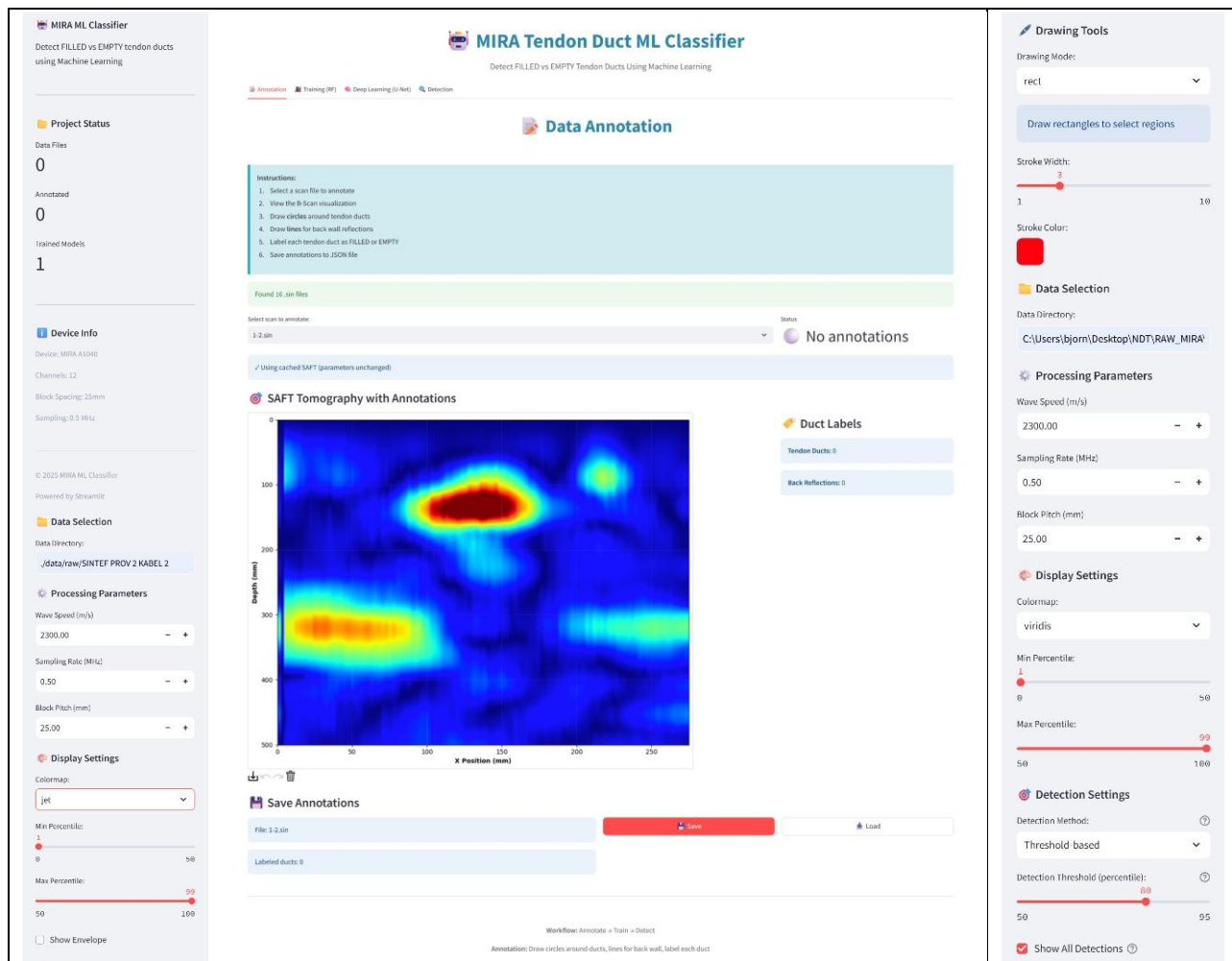


Figure A-2: Data annotation interface for UPE (MIRA)

Data annotation interface for MIRA ultrasonic measurements. Inspectors label individual B-scan segments as polygon and classes as EMPTY (void) or FILLED (grouted) based on SAFT reconstructions from raw I/Q data.

Impact-Echo ML Detector

Machine learning-based void detection in embedded tendon ducts

Project Status

Training Files
6

Test Files
10

Trained Models
3

Quick Actions

[Open Project Folder](#)

[View Documentation](#)

System Info

Python: 3.12.9

Working Dir: C:\Users\bjorn\Desktop\NDT\impact-echo-ml-project

© 2025 Impact-Echo ML Detector

Powered by Streamlit

Impact-Echo ML Detector

Machine Learning for Void Detection in Concrete Tendon Ducts

[Data Annotation](#) [Model Training](#) [Void Detection](#)

Data Annotation

Instructions:

- Select a .dat file to annotate
- The grid layout will be displayed
- Mark each point as **Filled (0)** or **Void (1)**
- Save annotations to CSV file

Select data source: ☒ Training Data ☐ Test Data ☐ Raw Data

Data Source: **Training Data**

Found 6 .dat files

Select .dat file to annotate:
Prov 2 kabel 1 sida A-2.dat

Annotating: **Prov 2 kabel 1 sida A-2.dat**

Grid size: 16 points across 16 rows

Quick annotation mode:
☒ Individual ☐ Batch - All Filled ☐ Batch - All Void

[Clear All Annotations](#)

Grid Annotation

Row 1	A1	Fill	Void	☐ Not Set
Row 2	A2	Fill	Void	☐ Not Set
Row 3	A3	Fill	Void	☐ Not Set
Row 15	A15	Fill	Void	☐ Not Set
Row 16	A16	Fill	Void	☐ Not Set

Annotation Statistics

Total Points	Annotated	Filled (0)	Void (1)
16	0	0	0
	↑ 0.0%	↑ 0.0%	↑ 0.0%

⚠ 16 points not yet annotated

Save Annotations

Save location:
data/training/labels/Prov 2 kabel 1 sida A-2_labels.csv

[Save Labels](#)

✓ New file will be created

Need help? Check the documentation files or hover over icons for tooltips

For best results: Annotate at least 200-500 points with balanced void/filled examples

Figure A-3: Data annotation interface for Impact Echo

Annotation interface for Impact Echo grid data. Labels distinguish between solid concrete (thickness frequency), and void conditions (anomalous high-frequency peaks).

MIRA ML Classifier

Navigation:

- ☐ Enhanced Annotation
- ☒ Training
- ☐ Detection

Data Configuration

Data Directory:

C:\Users\bjorn\Desktop\NDT\mira-ml-pr

Cache Directory:

./data/cache

Processing Parameters

Wave Speed (m/s)

2300.00 - +

Sampling Rate (MHz)

0.50 - +

Block Pitch (mm)

25.00 - +

Enhanced Model Training

Dual Training Pipeline:

- Random Forest:** Trained on global scan labels (FILLED/EMPTY) with extracted features
- U-Net:** Trained on polygon annotations mapped to raw I/Q data

Select Training Mode

Choose which model to train:

☒ Random Forest (Scan Classification) ☐ U-Net (Segmentation) ☐ Both

Dataset Summary

Total Scans	FILLED	EMPTY	Polygons
15	7	8	15

Random Forest Training

Training Strategy:

- Extract features from each annotated scan
- Use global scan label (FILLED/EMPTY) as target
- Train Random Forest classifier on extracted features

Model Parameters

Number of Trees:

100 - +

Max Depth:

10 - +

Min Samples Split:

5 - +

Train Random Forest

Training Settings

Test Set Size (%)

10 20 40

Random Seed:

42 - +

Figure A-4: Model training interface for UPE (MIRA)

Model training configuration interface for MIRA-based void detection. The interface allows selection of ML algorithms (Random Forest for classification, and U-Net for segmentation), hyperparameter tuning, and real-time visualization of training progress with cross-validation performance metrics (F1-score, AUC, confusion matrices).

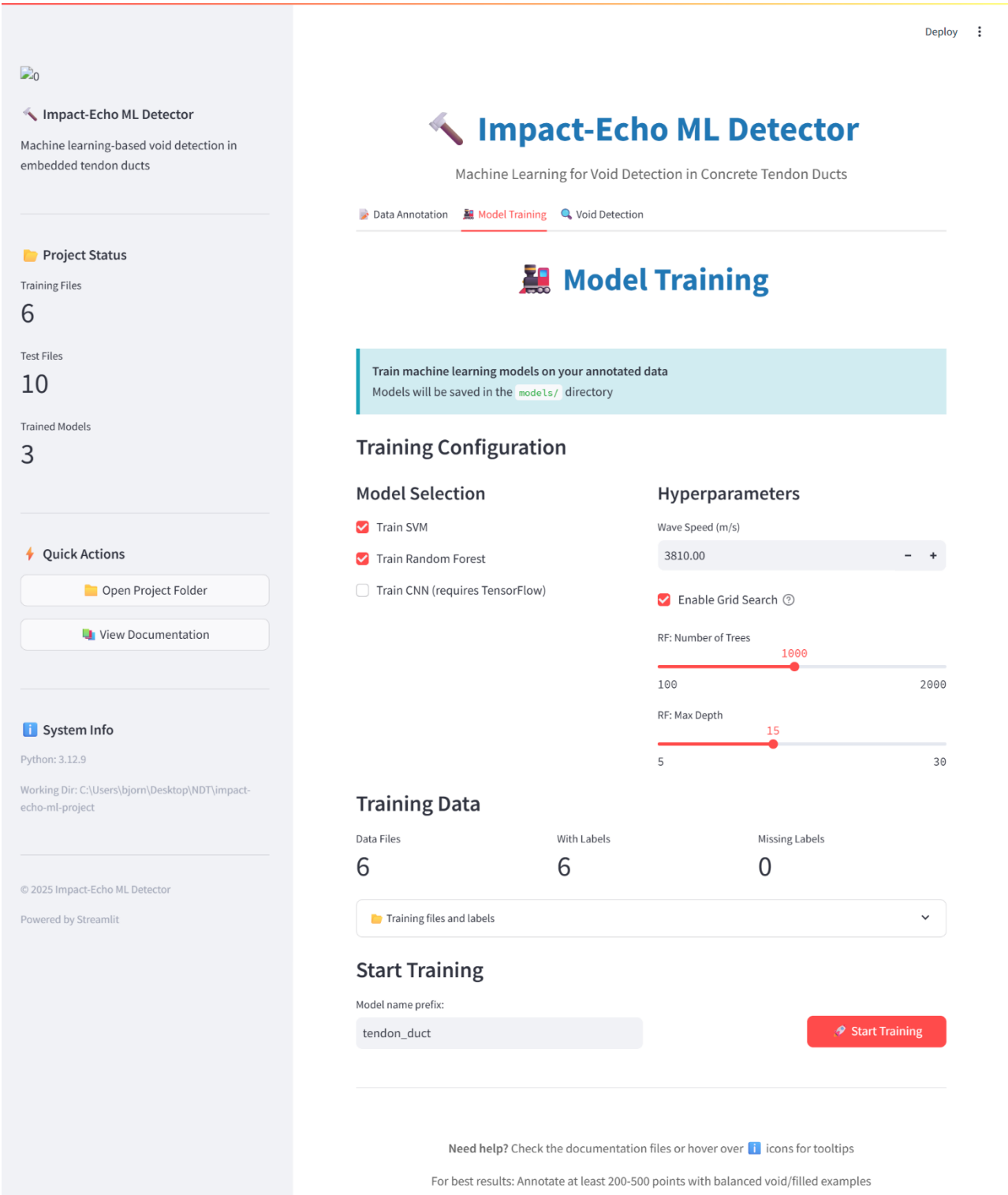


Figure A-5: Model training interface for Impact Echo

Training interface for Impact Echo resonance analysis models. Users configure model architecture, select spectral features (dominant frequency, peak width, amplitude ratios), set training/validation split ratios, and monitor convergence through loss curves and performance metrics. The interface compares Random Forest and 1D-CNN performance side-by-side to identify optimal model selection for deployment.