Journal of Building Engineering 95 (2024) 110188

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Full length article

Data-driven evaluation of building materials using Ground Penetrating Radar

Ahmed Nirjhar Alam [a], Wesley F. Reinhart [a,b], Rebecca K. Napolitano [c,*]

[a] Department of Materials Science and Engineering, The Pennsylvania State University, United States of America
[b] Institute for Computational and Data Sciences, The Pennsylvania State University, United States of America
[c] Department of Architectural Engineering, The Pennsylvania State University, United States of America

ARTICLE INFO

Keywords:
Ground Penetrating Radar
Machine learning
Structural health monitoring
Non-destructive evaluation
Feature analysis

ABSTRACT

Ground Penetrating Radar is a widely used technology in the nondestructive evaluation and monitoring of structures during repair and maintenance phases. Within this domain, a critical focus lies in assessing layered structures such as building envelopes. Evaluating attributes like thickness and material type in layered structures traditionally demands labor-intensive manual efforts involving calculations of dielectric properties and extensive signal processing. This study addresses these challenges by leveraging machine learning data-driven models that harness the entire recorded waveform, presenting a potential breakthrough for expediting diagnostic monitoring within the construction industry. We evaluate both supervised and unsupervised machine learning models for classification and regression tasks predicting the properties of common building materials. Experimental GPR A-scans serve as input for these models, with a thorough evaluation of both instance-based and parametric modeling approaches. A detection accuracy of 100% is achieved for identifying outlier scans, while material classification accuracy is 85%. Layer thickness predictions also had a high accuracy, with a typical error of 5%. Additionally, we explore the impact of feature learning and other preprocessing strategies on model performance. Our findings demonstrate the suitability of standard data-driven models for a spectrum of supervised learning tasks for layered structure diagnostics. However, we emphasize the importance of careful attention to the distribution of training data and its relevance to intended use in the field.

1. Introduction

It is projected that 75% of buildings in the United States will be retrofitted for improved energy efficiency by 2035 [1]. Ensuring the success of these retrofits is vital to offset the additional carbon footprint of new constructions [1]. However, achieving this goal is challenging due to the unknown state of materials inside the building envelope. Interior material attributes, such as thickness and integrity of insulation layers, directly impact a building's energy efficiency and susceptibility to moisture penetration [2]. Various structural diagnostic techniques, such as acoustic emission, ultrasonic, radar, and thermal imaging, have been developed to monitor the state of materials inside building envelopes [3]. Presently, infrared (IR) inspections are one of the most common methods for identifying building envelope faults such as air leakage and inadequate insulation [4]. However, IR inspections have limitations, as they cannot penetrate the entire envelope depth and can only see through the cladding layer [5]. To address this limitation and examine more interior spaces, radar techniques, particularly Ground Penetrating Radar (GPR), have been employed [3].

* Corresponding author.
E-mail address: nap@psu.edu (R.K. Napolitano).

https://doi.org/10.1016/j.jobe.2024.110188
Received 26 February 2024; Received in revised form 17 June 2024; Accepted 9 July 2024
Available online 16 July 2024

GPR is a non-intrusive and non-destructive method commonly used for investigating underground utilities made of concrete, asphalt, and metals [6–8]. It utilizes electromagnetic radiation in Ultra High Frequency (UHF)/Very High Frequency (VHF) frequencies to detect reflected signals from underlying structures. GPR is widely used in structural health monitoring (SHM) and non-destructive evaluation (NDE) problems. A significant potential application of GPR lies in estimating building envelope thicknesses, a critical aspect of building health inspections. This study focuses on applying automated data-driven models, leveraging GPR waveform data, to enhance diagnostic monitoring within the construction industry.

Prior works involving layer thickness predictions using GPR have relied heavily on signal-processing steps to acquire accurate information. The two-way travel time method has been applied to asphalt pavement layers in highways [9]. The common midpoint method (CMP) has been applied to analyze GPR data collected from interstate highways, resulting in a mean error of 6.8% [10]. A similar CMP method was utilized to measure the dielectric permittivity and thickness of snow and ice on a water body and produced accurate results [11]. The common source method, a multi-offset measurement method, has been used successfully for asphalt layer thickness and EM wave velocity estimation [12]. Regularized deconvolution and non-linear optimization techniques have been applied to GPR signals to predict pavement layer thicknesses [13,14].

Such signal-processing methods are often quite application-dependent and require significant manual effort. For instance, the work of Ref. [14] on asphalt layer thickness prediction using non-linear optimization required the embedding of tunable coefficients corresponding to the electromagnetic properties of each medium in the analytical expression. This may be highly effective for specific tasks but can be too time-consuming for rapid and versatile field applications.

Artificial intelligence techniques [15], which leverage data to teach machines to discern relevant patterns and subsequently make useful predictions, have been successfully used in numerous structural diagnostic tasks. Conventional image data, captured using unmanned aerial vehicles, have been successfully interpreted by machine learning models for crack detection [16]. Moreover, ML models have also been used to interpret ultrasonic testing data for the detection of corrosion-induced deterioration [17]. Data-driven algorithms have also been employed to successfully interpret signals obtained from acoustic emission testing [18].

Moreover, data-driven methods such as machine learning [19–21] and deep learning models [22–25] have had tremendous success recently in interpreting GPR signals. Support Vector Machines (SVM) with specialized features have been used to detect and segregate hyperbolae arising in GPR scans from underground utilities such as pipes and cables [19]. Logistic Regression (LR) and neural networks (NN) have also successfully discriminated between eight different types of landmines and clutter in a homogeneous sandy soil environment [20]. A suite of machine learning models was employed to develop a framework for selecting the most suitable classification model for landmine detection, given criteria such as class label ratio and desired performance metrics [21].

NNs have been used to search for pipes buried in homogeneous media in GPR radargrams [22]. A priori knowledge that a buried cylinder produces a hyperbolic signature in GPR images has also been employed to create a custom NN for location detection of buried pipes [23]. The YOLO architecture [26] has also been adapted for detecting concealed cracks in asphalt pavement using 3D GPR data [24]. A Convolutional Support Vector Machine network has been employed to interpret GPR scans from various soil/material types and buried object shapes [25]. Deep neural network architectures have been used to invert entire GPR B-scans [27]. Convolutional neural networks have been used to segment defects in tunnel linings from GPR scans [28].

In this study, we present a data-driven methodology designed for predicting material properties and thicknesses directly from GPR radargrams, with minimal preprocessing. The framework developed demonstrates high accuracy and reliability, particularly in interpolation cases. Our approach offers a rapid and versatile alternative to traditional signal-processing-based methods, where prior information such as permittivity and other coefficients for different materials must either be carefully measured or estimated and later tuned for predictions.

To the best of our knowledge, this is the first study on the application of data-driven models for nondestructive evaluation of layered structures using experimentally obtained GPR traces. This application is especially interesting because the layers are thin relative to the signal wavelength, which presents challenges that can only be solved with data-driven methods. We also demonstrate a novel stochastic feature elimination process for data-driven models applied to GPR and demonstrate how the distilled features relate to conventional signal processing methods. We also discuss potential resolution issues and investigate how the proposed models overcome such issues, which contribute insights into the reliability and applicability of the models in this study.

2. Materials and methods

2.1. Data acquisition and preparation

Table 1
Sample materials, their typical permittivity values, and typical usage in building structures. We did not measure the relative permittivity of our material samples, we are only referencing published values [29,30].

Material	Relative permittivity	Typical usage in built environment
PVC(R5)	4.0	Pipes, cables, window profiles, flooring, and roofing
Birch-based plywood	2.4–2.5	Beams and hoardings, crates, bins.
Steel	∞	Pipers, girders and columns

One and two-layer samples of steel, plywood, and PVC-R5 (Polyvinyl Chloride of R5 insulation capacity) were placed between two plastic sawhorses 30 in. above a concrete floor. The samples included two materials with close permittivity values (R5 and plywood) and a third material with a much higher permittivity value (steel). PVC of varying insulation capacities (R-value) is ubiquitously

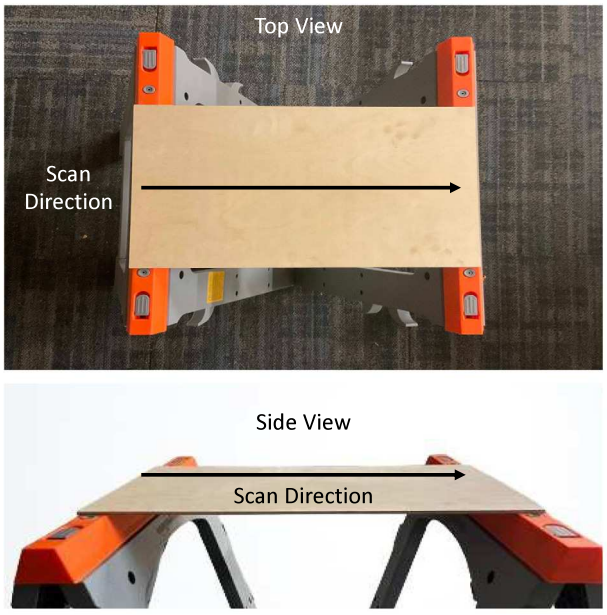

Fig. 1. Annotated photographs of our laboratory test apparatus with a plywood material sample. Material samples were placed between two plastic sawhorses, and GPR traces were taken in the scan direction indicated in the figure. Note the arrow represents the movement of the device, with signals being emitted perpendicular to this arrow (i.e., towards the floor).

used in building envelopes as thermal barriers [31]. Plywood is used frequently in wall sheaths [32]. Steel, on the other hand, is used as a cladding material in facades for both visual appeal and durability factors [33]. Thus in terms of material diversity in real-world construction scenarios, the training data covers a wide range of probable building envelope configurations. A commercial GPR machine (Proceq 8800) [34] was used for this work. It transmits Ricker wavelets in the 400–6000 MHz frequency range. The signal acquisition takes place over a 12 ns time window, with 655 samples evenly distributed in time, resulting in a 54.6 GHz sampling frequency. A single pass across the sample length (Fig. 1) was used to obtain radargrams (B-scans) corresponding to each material sample. Traces were removed from the sides of each B-scan to remove anomalies that appeared before and after full contact between the machine and the sample surface. Both B-scans and A-scans were used to train models with different tasks, as shown in Fig. 3.

All B-scans were resized to 15 traces for uniformity by removal of traces at either edge of the scan. Each trace was 12 ns long and recorded over 655 time samples. The B-scans and A-scans were represented as 2D and 1D arrays of size 655×15 and 655×1, respectively, with each row representing samples from the same time trace. In addition to the material layer samples, scans of interior and exterior building walls were included to create an out-of-domain data set for testing (Fig. 2).

2.2. Data set composition

The data set consisted of two broad categories; in-domain samples and outlier samples. In-domain samples are structures that fall within the problem domain and, hence, can be interpreted by the proposed models. Outliers are structures beyond the scope of the problem domain and, hence, cannot be interpreted by the proposed models.

In-domain configurations consisted of one material layer, and outlier configurations consisted of two material layers and building wall segments (Fig. 2). There were 80 in-domain scans corresponding to eight different material and layer thickness combinations. Generally, GPR A-scans contain signal noise which may vary significantly for scans of the same structure at different instances [35]. To account for such variations between otherwise identical scenarios, each configuration was scanned 10 times by the same operator using the same equipment.

2.3. Preprocessing

Typically, GPR data are subjected to preprocessing steps before interpretation [36]. Frequently used preprocessing techniques include frequency-based filters [37] and background clutter removal [38]. Moreover, signal deconvolution techniques have been successfully used for asphalt layer thickness predictions in prior works [39]. In this work, standard frequency domain filtration and clutter removal via mean subtraction did not improve results significantly. This is possibly due to the standardized environment in which the experiments were carried out. Each GPR scan was obtained in the identical EM scattering conditions present in the lab, with the same sources of background noise and induction effects.

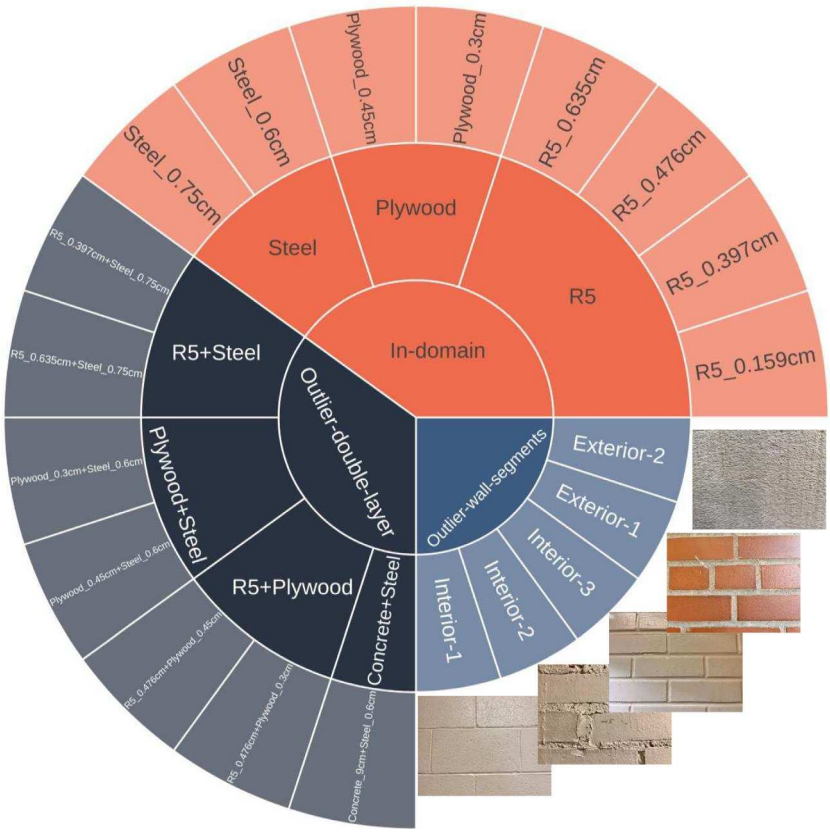

Fig. 2. Distribution of structural configurations in the data set. Sample names follow the convention [`top-material_top-thickness`] for one-layer samples and [`top-material_top-thickness + bottom-material_bottom-thickness`] for two-layer samples. Outlier configurations consist of two categories: (1) two-layer configurations and (2) Interior and exterior building wall segments consisting of brick, concrete, and plywood.

However, time-dependent gains were found to have a significant impact on model performance. This is perhaps due to the regularizing effect of such processing steps on the data set, which enabled relevant features to be significant in magnitude. Initial attempts at adjusting the gain using the principle of inverse amplitude decay [36] to achieve equal signal mean amplitude throughout the A-scan produced only minor performance boosts. This is likely due to relevant signals being present at a small time window in the A-scan, as seen in Section 5. Adjusting gain in this manner thus leads to irrelevant features being amplified (i.e., promoting a low signal-to-noise ratio), which hampers results. Subsequently, both linear and exponential time-dependent gains were explored, using model performance as the optimization objective. The application of gain involved the following operations:

$$A_g = A_0 \times t \tag{1}$$

$$A_g = A_0 \times 10^t \tag{2}$$

where t is the time-gain vector, A_0 is the initial A-scan without gain, A_g is the A-scan with gain A_g. Above, Eq. (1) implements linear gain, and Eq. (2) implements exponential gain. Exponential gain produced the best overall results, with the $t \in [0.00, 4.17]$ range being the most effective.

2.4. Workflow and models

Raw B-scans collected from all samples were first preprocessed by resizing and applying time-dependent gain (Fig. 3). The scans were then passed to a classifier, which characterized the scan as an outlier or in-domain. No further actions were taken if the passed B-scan was an outlier. If the scan was in-domain, only its central trace was passed to the material classifier and thickness predictor models. This is because the in-domain samples are uniform along the scan direction, therefore a single A-scan contains the same information as an entire B-scan (i.e., the comprising A-scans are effectively identical).

Various classification and regression models were used to perform the aforementioned tasks. All models were implemented in `scikit-learn` [40]. The `AX-Optimization` [41] library was used to optimize the classification model hyperparameters using Bayesian optimization, with the total number of trials set to 500. The default scikit-learn parameters were used as the starting

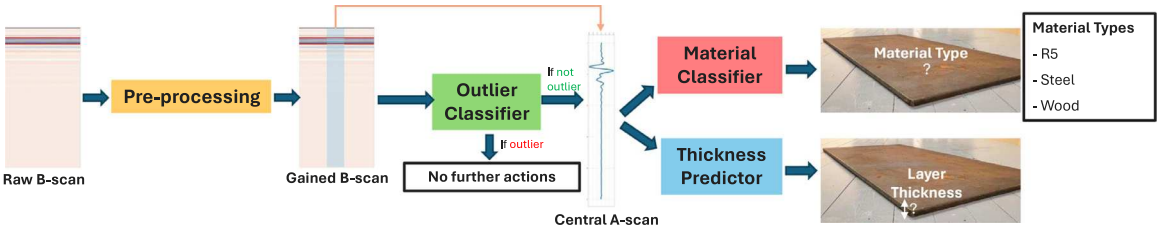

Fig. 3. Schematic of the data processing workflow. B-scans are initially processed with exponential time gain. A classification model subsequently determines whether the B-scan is from an in-domain or outlier sample. The central traces from in-domain sample scans are passed to the material classifier and layer thickness predictor models. No further actions are taken for scans from outlier samples.

Table 2
Summary of model types used for different tasks.

Tasks	Models
Outlier detection	**Supervised classifiers:** Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR) **Unsupervised classifier**
Material classification	**Classifiers:** RF, KNN and LR
Thickness prediction	**Linear parametric regressors:** Linear (Lin-R), Ridge (RR), and Lasso (Las-R) **Non-linear parametric regressors:** Support Vector (SVR), and Kernel Ridge (KR) **Non-parametric regressors:** KNN, RF, Gaussian Process (GP), and Gradient Boosting (GB)

point of the optimization process. The train–test data for each tuning process was selected based on specific tasks, as described in the corresponding sections. The hyperparameters were chosen to establish an optimal balance between bias and variance for each model and task. A systematic hyperparameter tuning procedure was also applied to regression models for material layer thickness predictions, as discussed in Section 3.5.

A wide array of parametric and non-parametric models were applied to the thickness prediction listed in Table 2. Parametric models assume a prior shape for the data distribution and hence often demonstrate better generalization capability and are computationally inexpensive to train [42]. There are two major classes of parametric models; linear and non-linear parametric models, which generalize to progressively complex functions. However, such models are often prone to biased estimations. Non-parametric models, on the other hand, make no prior assumptions about the data distribution. Hence, they are more versatile and can be fitted to more complex patterns [42]. However, such models can also be computationally expensive and may require more data to achieve the same performance as parametric models. Moreover, non-parametric models often overfit training data and display poor generalizability [42]. The performance of all three categories of models on different train-test splits was evaluated and compared in this work.

2.4.1. Metrics

We utilize the coefficient of determination according to its equivalence to explained variance in statistical analysis,

$$R^2 = \text{Explained Variance} = 1 - \frac{\sigma^2(y - \hat{y})}{\sigma^2(y)}, \tag{3}$$

where y indicates the true thickness, $\hat{y}$ indicates the predicted thickness, and σ^2 indicates the variance. Thus $\sigma^2(y-\hat{y})$ is the explained variance and $\sigma^2(y)$ is the total variance. Note that explained variance can be negative for predictions on unseen test data, indicating that the model predictions are worse than assuming the null hypothesis.

We also report results using the Root Mean Squared Error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

where i indicates a particular sample, and n indicates the total number of samples.

Percent Root Mean Squared Error (PRMSE) is a modified version of RMSE that considers the magnitude of each target label, reducing the priority towards high-magnitude labels. PRMSE was used to compare model performances for layer thickness prediction because RMSE showed significant bias to thicker layers. We define the PRMSE according to

$$\text{PRMSE} = \frac{\text{RMSE}}{\bar{y}} \times 100\%, \tag{5}$$

where $\bar{y}$ indicates the mean of the true thickness values.

2.5. Feature analysis

A Genetic Algorithm (GA) was explored to carry out feature elimination for the developed models. GAs are powerful and versatile optimization techniques inspired by natural selection and evolution [43]. The fundamental idea behind GA is to mimic the principles of natural selection, such as survival of the fittest and reproduction, to find a near-optimal solution to a problem. The process begins with a population of potential solutions (the model feature sets in this work), represented as chromosomes or strings of values, which are evaluated for their fitness based on a predefined objective function.

During each iteration of the algorithm (generation), individuals with higher fitness have a greater chance of being selected for reproduction. Selected individuals undergo genetic operations like crossover (recombination) and mutation, where parts of their genetic information are combined and altered to create new solutions.

3. Results

3.1. Outlier detection

For the proposed methodology to be useful as a diagnostic tool, it must have the capability to automatically screen out scans from configurations that are significantly different from training data and hence cannot be interpreted reliably by the trained models. This necessitates the development of an outlier detection model. Unlike in-domain configurations, which are all uniform one-layer structures, outlier configurations in this study may vary along the scan direction. Thus, B-scans were used as input data for outlier detection instead of A-scans. The experimental B-scans, resized to 655×15 arrays, were flattened to 1D vectors of length 9825 by appending each trace end-to-start sequentially along the scan direction.

A 10-fold cross-validation scheme was used to evaluate model performances. Unlike standard cross-fold validation schemes, the train–test split was not completely random since there are repetitions of each structural configuration in the data set (Section 2.2). Thus, a random splitting of the data set would result in radargrams from each configuration being present in both the test set and train set, which would not test the ability of the model to detect outlier configurations. For meaningful evaluation, the performance of different outlier screening models was tested by controlling the distribution of configurations in the train–test set. Classification accuracy was used both as training loss and as a general metric for comparing different models.

3.1.1. Unsupervised learning

Unsupervised neighbor-based classifiers have been successfully applied in structural health monitoring for anomaly detection. Many such algorithms have employed feature selection and feature extraction with tunable distance metrics [44,45] to enhance results. However, for generalizability, the outlier detector in this work was trained using the entire feature set on the standard Euclidean distance. The outliers in any diagnostic scenario may have near-infinite variations. Supervised models rely on the models of all class labels to create decision boundaries. Thus, it is necessary to have a representative sample of each class label in the training data. However, in this scenario, creating a comprehensive dataset of all possible outliers is impossible. Thus, instead of creating a supervised model that stores the attributes of all possible outliers and non-outliers, it is desirable if the outlier detector can simply use the characteristics of the in-domain scans and apply a similarity-based metric to classify new scans. Therefore, the model training data set consisted solely of in-domain scans. The test set, on the other hand, consisted of a mix of in-domain and outlier scans. Classification accuracy on a validation set is used as the performance metric to tune the number of neighbors considered in the unsupervised classifier (see Fig. 4).

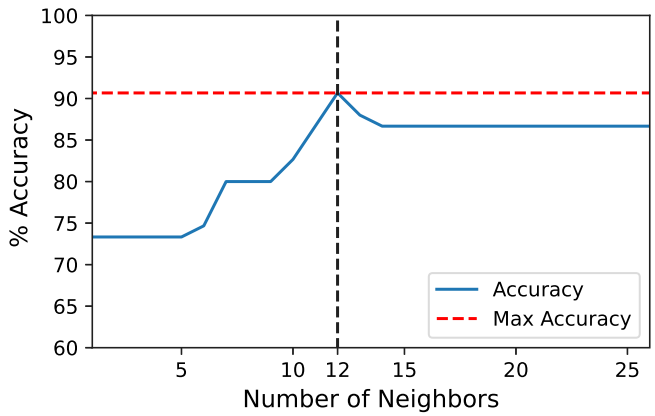

Fig. 4. Accuracy of the unsupervised classifier with different numbers of nearest neighbors.

The unsupervised learner calculated the 12 nearest neighbors (obtained through performance tuning) of each data point in the training set in Euclidean space and their corresponding distances.

Table 3
Composition of train and test set for the unsupervised scenario.

Train	Test
R5_0.159 cm	R5_0.497 cm
R5_0.397 cm	R5_0.635 cm
plywood_0.3 cm	plywood_0.45 cm
steel_0.6 cm	steel_0.75 cm
	All outliers

The decision boundary in the unsupervised model is simply defined by distance from neighbors. Thus, a balanced representation of all samples is not necessary for the unsupervised classifier. Indeed, creating a balanced representation for the unsupervised classifier involved trimming available data from the training set which led to a minor reduction in performance.

Predictions for the test set were made by calculating the distance of each test sample from the 12 nearest training data points. If this distance exceeded the reference maximum distance calculated using training points, the samples were labeled as out-of-domain. Otherwise, they were labeled to be in-domain. Using the unsupervised scenario defined in Table 3, an accuracy of 90.7% was obtained on the test set with 80.7% accuracy for in-domain detection and 89.2% for outlier detection.

A weighted version of this model was also implemented but failed to produce significantly different results. Only R5+wood and steel samples were misclassified. Steel sample A-scans have distinctive signatures that persist in double layers as well which can explain their misclassification as outliers. R5 + wood is likely to be misclassified as a single R5/wood layer due to similar permittivity values.

3.1.2. Supervised learning

Supervised models, which can utilize labeled data, were subsequently explored for outlier detection. RF, LR, and KNN models, with more varied and tunable decision boundaries, were trained on both outlier and in-domain scans. LR is a parametric model that assumes a non-linear yet simple decision boundary. KNN and RF, on the other hand, are non-parametric models. KNN is the simpler of the two models, operating solely based on votes from neighborhood data points. RF is an ensemble learning method that uses the average of predictions from a collection of decision trees. Thus, RF can have varied complexity, depending on hyperparameters such as the number of estimators (trees) in the model and the maximum depth allowed for each tree. The supervised learning models are tested using five different classes of train-test splits (Listed as different Scenarios in Table 4, each of which assesses the model performance in different distributions of the data. The hyperparameters of all models were tuned using the train and test set in Scenario 1 in Table 4, which best represents an adequate dataset in the real application scenario. The default and tuned hyperparmaeters of all models are listed in Table 5.

Table 4
Data splitting for different outlier classification Scenarios as discussed in the text.

Sample	Scenario 1		Scenario 2		Scenario 3a		Scenario 3b		Scenario 4	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R5 0.397 cm + steel 0.75 cm	x		x		x			x	x	
R5 0.635 cm + steel 0.75 cm		x				x		x		x
plywood 0.3 cm + steel 0.6 cm	x		x		x		x		x	
plywood 0.45 cm + steel 0.6 cm		x	x			x		x		x
R5 0.476 cm + plywood 0.3 cm		x	x			x	x			x
R5 0.476 cm + plywood 0.45 cm	x		x			x		x	x	
Interior 1	x		x					x	x	
Interior 2	x		x					x	x	
Interior 3		x		x						x
Exterior 1	x		x					x	x	
Exterior 2		x		x						x
In-domain (random split)	80%	20%	80%	20%	80%	20%	80%	20%	80%	20%
In-domain (even representation)									50%	50%

Table 5
Classification models and their default and tuned hyperparameter values for supervised learning based outlier detection.

Classification model	Default parameters	Tuned parameters

Random Forest (RF)	`n_estimators = 100,`
	`max_depth = None,`
	`min_samples_split = 2,`
	`min_samples_leaf = 1`
		`n_estimators = 50,`
		`max_depth = 100,`
		`min_samples_split = 11,`
		`min_samples_leaf = 2`

KNearest Neighbors (KNN)	`n _neighbors = 5,` `weights = uniform`	`n_neighbors = 3,` `weights = distance`
Logistic Regression (LR)	`tol = 0.0001,` `C = 1.0`	`tol = 0.0002,` `C = 1000`

Scenario 1: Stratified outlier sampling

This scenario explored the performance of the model when all possible categories of outliers are represented in the training data. There are five classes of outliers in the data set: interior wall, exterior wall, R5 + plywood, R5 + steel, and plywood + steel. A configuration from each class is included in the training and test sets. In-domain samples were split randomly between train and test sets to maintain as close to an 80%/20% train/test split as possible.

With stratified sampling, the classifiers perform very well, with near-perfect performance from all three models (Table 6). The misclassifications in KNN and LR models result soley from the R5 and plywood samples. R5 and plywood have similar permittivity values (Table 1), which may lead their two-layer to produce scans similar to a one-layer scan and thus be misclassified as in-domain scans and vice versa.

Table 6
Performance of each model type on the different outlier classification Scenarios (different train/test splits described in the text). Accuracy is reported from 10-fold cross-validation (reported as mean ± standard deviation). The best-performing model in each Scenario is displayed in bold.

Scenario	RF	KNN	LR
1	**100% ± 0.0%**	99.8% ± 0.7%	99.9% ± 0.3%
2	99.5% ± 0.2%	99.7% ± 0.1%	**100% ± 0%**
3a	74.2% ± 0.4%	**74.4% ± 0.0%**	**74.4% ± 0%**
3b	99.0% ± 1.1%	**100.0% ± 0%**	**100.0% ± 0.0%**
4	99.8% ± 1%	99.0% ± 0.3%	**99.9% ± 0.3%**

Scenario 2: Imbalanced outlier distribution

This scenario investigated the impact of data set imbalance on model performance. Interior 1–2, exterior 1, and all two-layer outliers were included in the train set. Interior 3 and Exterior 2 were included in the test set. The in-domain scans were randomly split between the test and train set, so the train–test split was as close to 80-20 as possible. This led to a training set that included 65.5% outliers and only 34.5% non-outliers.

Once more, a very high accuracy was obtained in the test set for all three classifiers, with LR slightly outperforming KNN and RF (Table 6). Notably, all errors in RF and KNN were false positives. The over-representation of outliers in the training set meant that the classifier could achieve a high training score by creating a decision boundary that is biased towards outliers, which leads it to err on the side of false positives.

For RF, the steel_0.75 cm sample was the only source of error, whereas, for KNN, the errors consisted of steel_0.75 cm and R5_0.635 cm. The very small percetage of misclassification indicated that the error is not systematic with respect to configurations (for instance, only a fraction of the steel_0.75 cm scans are misclassified). The small yet consistent misclassification of the steel samples was possibly due to their high reflectivity, which exacerbated the effect of irregularities such as surface roughness. Layer thickness predictions for steel configurations also showed significant variability, as seen in Section 3.5.

Scenario 3: Left-out outlier classes

The purpose of this scenario was to investigate whether the model can detect previously unseen outlier types. There are three classes of two-layer outliers in the data set: R5 + plywood, R5 + steel, and plywood + steel. Each class has two different configurations. To that end, one configuration each from R5 + steel and plywood + steel is included in the training data set, while no R5 + plywood configurations are included in the training set. This is called Scenario 3a. The test set contains configurations from all two-layer outlier types. As before, in-domain samples are added to each set to achieve as close to 80-20 train–test splits as possible. Interior-exterior wall outliers were left out from both sets.

The classification models detected all non-outliers accurately, but outlier detection accuracy was low, with roughly 50% of the outliers being misclassified for KNN and RF. All the R5 + plywood configurations, which were not included in the training set, are misclassified as non-outliers. Both outlier and non-outlier detection accuracies were worse for LR. The result suggested that layer stackings not included in the training set are misclassified. However, an alternate explanation could be that the similar permittivity of R5 and plywood makes it difficult to distinguish R5 + plywood scans from one-layer R5 or plywood scans.

To test this hypothesis, the models are re-run with R5 + steel being left out from the training set instead of R5 + plywood, called Scenario 3b (Table 4). A high accuracy (99%–100%) is obtained for RF and KNN again, as opposed to the 74% accuracy for RF and KNN with the R5 + plywood left out from the training set (Table 6). This supports the hypothesis that the similar permittivity of R5 and plywood makes R5+plywood configurations difficult to distinguish from one-layer in-domain scans.

Scenario 4: Stratified outlier sampling with equal representation

The in-domain dataset contains twice as many R5 samples as plywood and steel. To correct the over-representation of R5 samples in the training set, Scenario 1 is modified to exclude half of the R5 samples in the in-domain scans (Scenario 4).

As seen in Table 6 the loss in data leads to slightly reduced accuracy for the RF classifier (99.8%), which was the best-performing model for Scenario 1. All misclassifications arise from R5 and wood misclassification.

To summarize, RF and KNN classifiers, trained on a data set with a complete and balanced representation of outliers and non-outliers, can differentiate between outliers and non-outliers with over 98% accuracy. The prediction bias can lean towards either outliers or non-outliers, depending on the composition of the data set. Moreover, we hypothesize that samples of high-reflectivity materials like steel will likely have non-systematic misclassification due to erratic scattering effects. Lastly, layer stackings containing materials with similar permittivity values will likely be misclassified as one layer and, therefore, in-domain. Thus, for building envelope diagnostics, construction plans may need to be used as a supplement to the ML-based interpretation.

3.2. Single layer material classification

Previous works have reported various values for material classification using GPR signals. For instance, a classification accuracy of 95% has been reported using synthetic data composed of iron, aluminum, and limestone [46]. Similarly, accuracy ranging from around 78%–100% (depending on the EM environment) has been reported for classifying materials in underground utilities [47]. In this work, KNN, RF, and LR, with model hyperparameters tuned on an evenly split dataset (Section 3.2.2), are applied to material classification. The default and tuned hyperparameters are listed in Table 7. Three different train-test splits with a 10-fold cross-validation scheme were explored to evaluate the performance.

Table 7
Classification models and their default and tuned hyperparameter values for material classification in single-layer configurations.

Classification model	Default parameters	Tuned parameters

Random Forest (RF)	`n_estimators = 100,`
	`max_depth = None,`
	`min_samples_split = 2,`
	`min_samples_leaf = 1`
		`n_estimators = 544,`
		`max_depth = 79,`
		`min_samples_split = 8,`
		`min_samples_leaf = 3`

KNearest Neighbors (KNN)	`n _neighbors = 5,` `weights = uniform`	`n_neighbors = 7,` `weights = distance`
Logistic Regression (LR)	`tol = 0.0001,` `C = 1.0`	`tol = 0.025,` `C = 729`

3.2.1. Random split

A random split of all in-domain configurations, with an 80-20 train-test ratio, made up the initial data set. KNN and RF successfully classify the different materials, with accuracy scores between 98%–100% for test cases and 100% for the train cases (see Fig. 5).

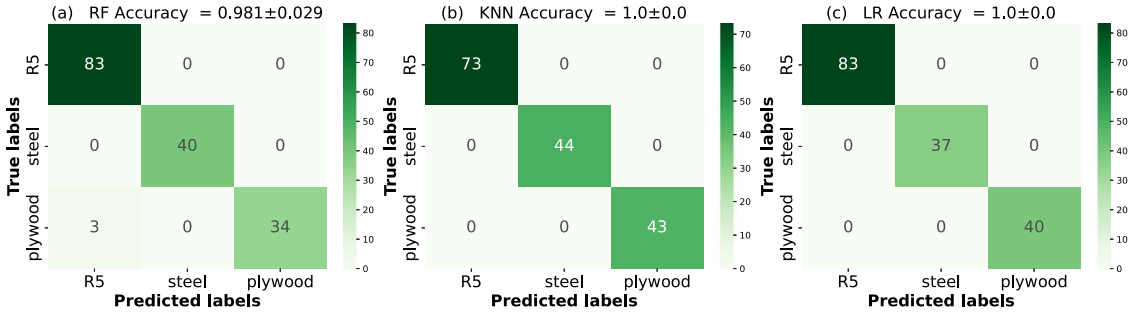

Fig. 5. Test performance for RF, KNN, and LR with randomly selected train/test sets displayed as confusion matrices.

However, a random split of in-domain scans is not a complete evaluation of the material classifier models. As before with outlier detection, the 10-scan repetition of each configuration means that, by random test–train split, each configuration's scans are present in both the training set and test set. The near 100% accuracy of the RF and KNN model simply demonstrates that the classifier is not derailed by the random noise components present in the A-scan signal.

3.2.2. Stratified split

To evaluate the capability of the classifier to detect material types for different configurations, the data set is split in a stratified manner (including an equal number of samples from each material class in the train and test set) (Table 8). This results in a test–train split of 50–50, which is not ideal, but unavoidable if all materials are to be equally represented in the test and train set. The performance of the classification models with stratified split are listed in Fig. 6.

Table 8
Composition of train and test sets with stratified split.

Train	Test
plywood_0.3 cm	plywood_0.45 cm
R5_0.159 cm	R5_0.497 cm
R5_0.635 cm	R5_0.397
steel_0.6 cm	steel_0.75 cm

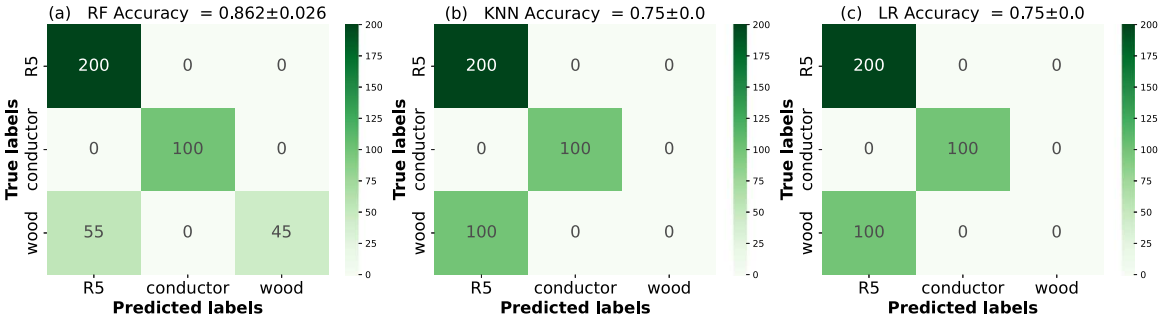

Fig. 6. Test performance for RF, KNN, and LR with a stratified (balanced) train/test split, displayed as confusion matrices.

Training accuracies of RF and KNN are both 100%. LR also has a training accuracy of 100%, most likely due to the hyperparameter tuning process being carried out on the stratified split dataset. In terms of test set performance, RF performs the best, with an accuracy of 86.2% while KNN and LR perform at an accuracy of 75%. In both models, the misclassifications result from plywood configurations being misclassified as R5, likely due to their similar permittivity values.

Only a single configuration, the plywood_0.45 cm, is misclassified by the RF model. 59% of the plywood_0.45 cm samples are misclassified as R5, whereas the rest 41% are correctly classified as plywood.

The significant discrepancy between the training and test performance indicated that there may be model overfitting. To test the possibility of overfitting, the RF, KNN, and LR models were re-trained with hyperparameters adjusted for overfitting (fewer number estimators/depth for RF, higher number of nearest neighbors for KNN, and a stronger regularization term for LR). However, this further deteriorated performance. RF, which had the best performance with tuned hyperparameters, had a reduced accuracy of 78%, with 66% plywood samples misclassified as R5.

3.2.3. Stratified split with equal representation

In the previous train-test split, plywood samples were misclassified as R5, likely due to the similar permittivity of the two materials. However, R5 samples were not misclassified as plywood. This was thought to be due to the over-representation of R5 in the training set (there are twice as many R5 samples as plywood in the training set). To investigate, the three models were re-trained with an equal number of samples from each material in the training set (see Table 9).

Table 9
Composition of train and test sets with an equal number of configurations from each category.

Train	Test
plywood_0.3 cm	plywood_0.45 cm
R5_0.159 cm	R5_0.635 cm
steel_0.6 cm	steel_0.75 cm

Training on a dataset with equal representation of each material caused the misclassification to occur in the reverse direction, with R5 samples being misclassified as plywood (see Fig. 7). This confirmed the strong impact of training data composition on the classifier output. However, unlike the previous split, 100% of the R5 samples were misclassified plywood, resulting in a reduced model accuracy of 66.7%. Removing data from the training set for a more balanced representation thus resulted in poorer performance.

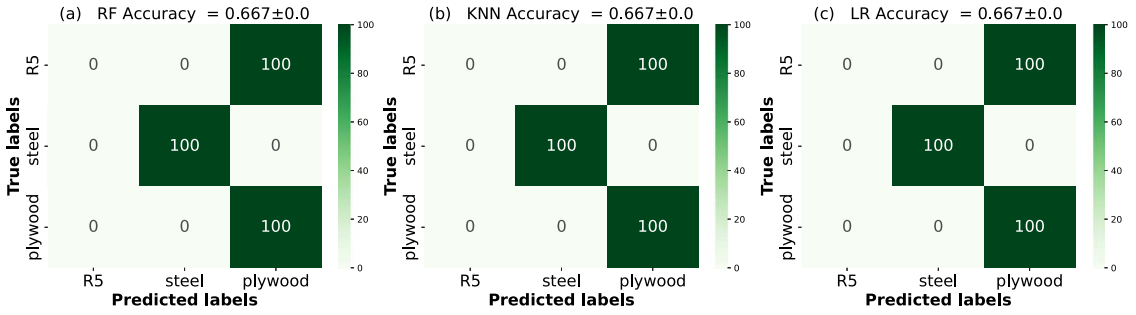

Fig. 7. Test performance for RF, KNN, and LR with a stratified (balanced) train/test split with equal representation of all materials, displayed as confusion matrices.

In conclusion, similar to outlier detection, material classifier performance is also highly dependent on training data composition, with over/under-representation of different material classes strongly impacting the bias in predictions. Moreover, materials with similar permittivity are more likely to be mistaken for one another, with the misclassification direction being dictated by the composition of the training data set. The failure of both balanced data sets and reduced model complexity to improve performance suggests that more samples are needed to enable the classifier models to differentiate between materials with similar permittivity values such as R5 and plywood. However, significant differences in attributes such as interior/exterior roughness, porosity, and moisture content may produce appreciably different reflected signals from materials with similar permittivity values. The homogeneous nature of the samples used in the lab experiments may have prevented such distinctive features from appearing in the received signals. Moreover, higher-resolution receivers may also capture the signal perturbations arising from such geometric irregularities, enabling their differentiation via ML models.

3.2.4. Uncertainty quantification

To further assess the quality of the material classification predictions, the uncertainty associated with individual predictions is investigated. The probability of each class label for a representative collection of test sample predictions is displayed in Fig. 8 with the tuned RF classifier model for the stratified split (Section 3.2.2). The prediction probability inaccuracies for R5 and steel samples are smaller, which is in line with the classifier performance. Plywood predictions, on the other hand, display R5 with a significant probability due to the similar permittivity of the two materials and the relative dominance of R5 in the training dataset.

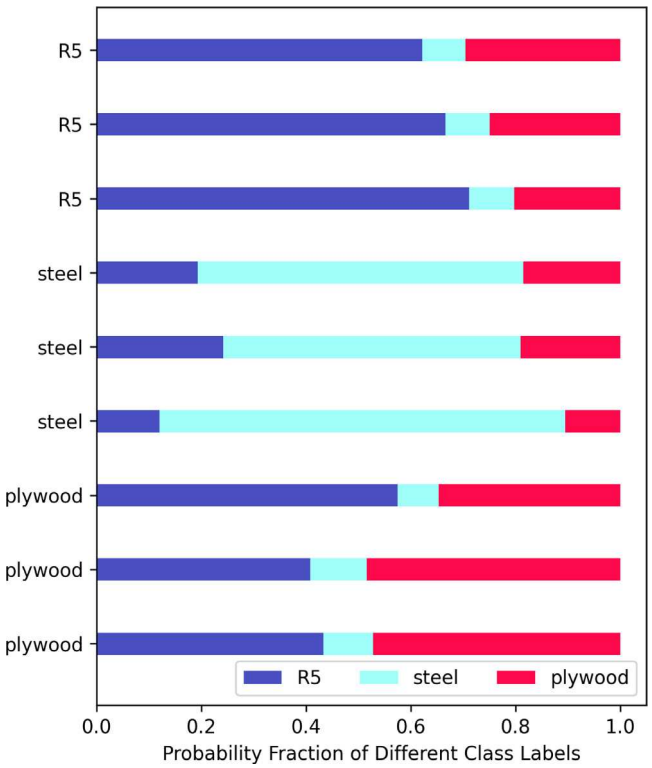

Fig. 8. Probability of each class label for a representative collection of sample predictions on unseen test data.

3.3. Impact data distribution and potential mitigation strategies

A recurring occurrence with all classification tasks was the impact of material over/under-representation dictating the direction of misclassification in the outlier detection and single-layer material classification tasks. Using balanced datasets impacted the direction of misclassification. For instance, plywood samples were misclassified as R5 for the stratified split in single-layer classification (Section 3.2.2). However, for the balanced data with equal representation of all samples (Section 3.2.3), R5 samples were misclassified as plywood. Moreover, in all such instances, overall model performance deteriorated due to the data omission from creating a balanced representation.

Thus, one potential mitigation strategy would be data augmentation, such as creating additional samples from existing data by applying signal processing steps such as varying time-dependent gain. Such steps would not fundamentally alter the A-scans but instead provide additional variability to the dataset that can be used to offset bias in data distribution. Moreover, varied scan

repetitions can also be applied to bolster samples from under-represented materials in the dataset. This exploration is kept as a possible future work in this paper, for the purpose of brevity.

3.4. Double layer material classification

Two classifiers connected end to end are used for double-layer material classification. The first classifier is used to distinguish between double-layer and single-layer configurations. The second classifier is used to identify the materials present in the double-layer configuration. Initially, a general classification model, with class labels for each double-layer and single-layer configuration present in the dataset, was implemented. This scheme produced very poor results, particularly for single-layer configurations. This may be due to the highly precise decision boundary that is required to separate all class labels efficiently.

3.4.1. Single-layer vs. double-layer classifier

Table 10
Composition of train and test for single-layer vs double-layer classifier.

Train	Test
plywood_0.3 cm	plywood_0.45 cm
R5_0.159 cm, R5_0.397 cm	R5_0.476 cm, R5_0.635 cm
steel_0.6 cm	steel_0.75 cm
R5_0.397 cm + Steel_0.75 cm	R5_0.635 cm + Steel_0.75 cm
Plywood_0.3 cm + Steel_0.6 cm	Plywood_0.45 cm + Steel_0.6 cm
R5_0.476 cm + Plywood_0.3 cm	R5_0.476 cm + Plywood_0.45 cm

As with single-layer material classification, RF, KNN, and LR with tuned hyperparameters were employed for distinguishing between single and double layer samples. The train and test splits for this task are listed in Table 10. RF displayed the best performance, with a mean accuracy of 89.6%. All misclassifications result from single-layer configurations misclassified as double-layer configurations, with R5 making up about 88% of the misclassifications. R5 and plywood have similar permittivity, which would cause double/single layers to have similar traces. Steel has distinctive traces with sustained long amplitudes in both single- and double-layer configurations, which possibly contribute to the misclassification.

3.4.2. Material classifier

Table 11
Composition of train and test for the double layer material classifier.

Train	Test
R5_0.397 cm + Steel_0.75 cm	R5_0.635 cm + Steel_0.75 cm
Plywood_0.3 cm + Steel_0.6 cm	Plywood_0.45 cm + Steel_0.6 cm
R5_0.476 cm + Plywood_0.3 cm	R5_0.476 cm + Plywood_0.45 cm

This classifier implements the one-vs-all strategy by using a separate model for detecting the presence of each material in the double layer. The train and test splits for this task are listed in Table 11. The combined output from all three classifiers fully determines the double-layer configuration since there are no different stacking combinations. If the models indicate all three materials are present, the result is invalid and is thus counted as misclassification. The performance of each material classifier is listed in Table 12.

Table 12
Accuracy of each double layer material classifiers.

Material classifier	Accuracy
R5	$89.3\% \pm 8.3\%$
Plywood	$100.0\% \pm 0.0\%$
Steel	$100.0\% \pm 0.0\%$

Combining the results, the double layer classification accuracy is 89.33%. All the misclassification stems from R5 classification with Plywood_0.45 cm+Steel_0.6 cm samples misclassified as containing R5. This is consistent with Plywood and R5 having similar permittivity values and steel having characteristic trace patterns, which possibly leads to Plywood+steel and R5+steel having similar A-scans.

3.5. Thickness prediction

Regression models presented in Table 2 were used to predict material thickness from preprocessed in-domain A-scans. However, the default scikit-learn model hyperparameter values for certain regression models produced poor results. This was especially true for RR, LR, SVR, and KR. Thus, a basic optimization was performed on each model to improve performance. The tuning process was carried out for all models, as opposed to simply the poor-performing models. The purpose of this was to standardize the performances across different models for comparison purposes by employing an identical number of optimizer iterations toward hyperparameter tuning.

The `AX-Optimization` [41] library was used to optimize the regression model hyperparameters using Bayesian optimization, with the total number of trials set to 500. The default scikit-learn parameters were used as the starting point of the optimization process. The defualt and tuned hyperparameters for each model are listed in Table 13. The train-test data were divided according to the interpolation split illustrated in Table 14. The computation time for optimization ranged from around 10 min to 12 h for each model on a remote Intel(R) Xeon(R) 2.3 GHz CPU dual-core CPU.

Table 13
Regression models and their default and tuned hyperparameter values.

Regressor model	Default parameters	Tuned parameters
Linear (Lin-R)	`None`	`None`
Ridge (RR)	α = 1	$\alpha = 1e^{-5}$
Lasso (LR)	α = 1	$\alpha = 1e^{-5}$
Support Vector (SVR)	C = 1, $\epsilon = 1e^{-1}$	C = 100, $\epsilon = 1e^{-5}$
Kernel Ridge (KR)	`kernel = linear`, α = 1,	`kernel = polynomial`, α = 0.01
KNearest Neighbors (KNN)	`n_neighbors = 5, weights = uniform`	`n_neighbors = fit to train, weights = uniform/distance`
Random Forest (RF)	`n_estimators = 100, max_depth = till pure leaf`	`n_estimators = 100, max_depth = 100`
Gaussian Process (GP)	`kernel = custom RBF`	`kernel = RBF+White`
Gradient Boosting (GB)	`n_estimators = 100, max_depth = 3 loss = 'absolute_error', learning_rate = 0.1`	`n_estimators = 1000, max_depth = 10 loss = 'absolute_error', learning_rate = 0.1`

Table 14
Data splitting for different regression performance tests as discussed in the text.

Samples in order of thickness	Random		Interpolation		Extrapolation	
	Train	Test	Train	Test	Train	Test
R5_0.159 cm	x	x	x		x	
plywood_0.3 cm	x	x	x		x	
R5_0.397 cm	x	x	x		x	
plywood_0.45 cm	x	x		x	x	
R5_0.476 cm	x	x		x	x	
steel_0.6 cm	x	x	x		x	
R5_0.635 cm	x	x	x			x
steel_0.75 cm	x	x	x			x

Linear regression is the simplest of all the models applied to the problem, with a single bias coefficient and weight coefficients corresponding to each time sample in the A-scan. Kernel and ridge regressions are regularized versions of the linear regression model where the coefficient α is the regularization parameter and controls the strength of the regularization. Likewise, for SVR, C is the regularization parameter, and ϵ is the minimum value of prediction error with which penalty is associated for the training loss function. In the KR model, the `kernel` is a non-linear transformation that maps data points to a higher dimensional space. Hyperplanes in this higher dimensional space can order data points more effectively for continuous-valued predictions, which enables simpler decision boundaries to produce better results. Several kernel shape functions are available for KR, GPR, and GB, and their choice is a major model tuning choice for all three models. Besides choosing the kernel shape, continuous-valued optimizations were also carried out to tune the exact functional shape of kernels for performance.

`n_neighbors` in KNN is the number of nearest neighbors whose output values are averaged for predictions. This is five by default in scikit-learn. In this work, the KNN model's number of neighbors is adjusted to best fit the training data. A weighted KNN model was also implemented. RF and GB are ensemble learning models and pool results from multiple models to make predictions. `n_estimators` is the total number of models employed. `max_depth` is the maximum depth a tree can reach and indicates the maximum number of conditions that can be applied to a thread from the root node to the leaf node. The 'loss' parameter in GB indicates the loss function that is used to evaluate model loss for parameter update.

Three types of train–test splits: (1) random split, (2) interpolation split, and (3) extrapolation split, were used to evaluate model performances. For random split, the data set was randomly split into train–test sets with an 80%–20% ratio. Thus, the training and test sets contained repeated scans of the same configurations solely for the random split case. Interpolation split evaluates the ability of the model to predict layer thickness values between the extremes encountered in the training set. Samples with thickness values between 0.4 cm–0.5 cm (plywood_0.45 cm and the R5_0.497 cm) are used in the test set, and the rest are used as training data. Extrapolation split evaluates the ability of the model to predict layer thicknesses that are outside the range of the training data. Samples greater than 0.6 cm (steel_0.75 cm and R5_0.635 cm) are used as test data, with the rest being in the training set.

To investigate model performance variability with different training data combinations, the training set was randomly shuffled and reduced by 10% during each model run. This process was repeated 10 times to evaluate each model and data split combinations.

A large range of accuracy values for layer thickness predictions has been reported in previous works, with a wide variation in functional areas and specimens. For instance, errors of 4.2% and 6.38% have been reported when predicting pavement top layer thicknesses using regularized deconvolution [39] and a non-linear optimization [48] technique respectively. In addition to previous works, tolerance standards in construction also provide a guide map for defining acceptable model performance. For vertical in-plane brick walls, a wall thickness tolerance of around 8%–11% is recommended [49,50].

3.5.1. Random splitting scenario

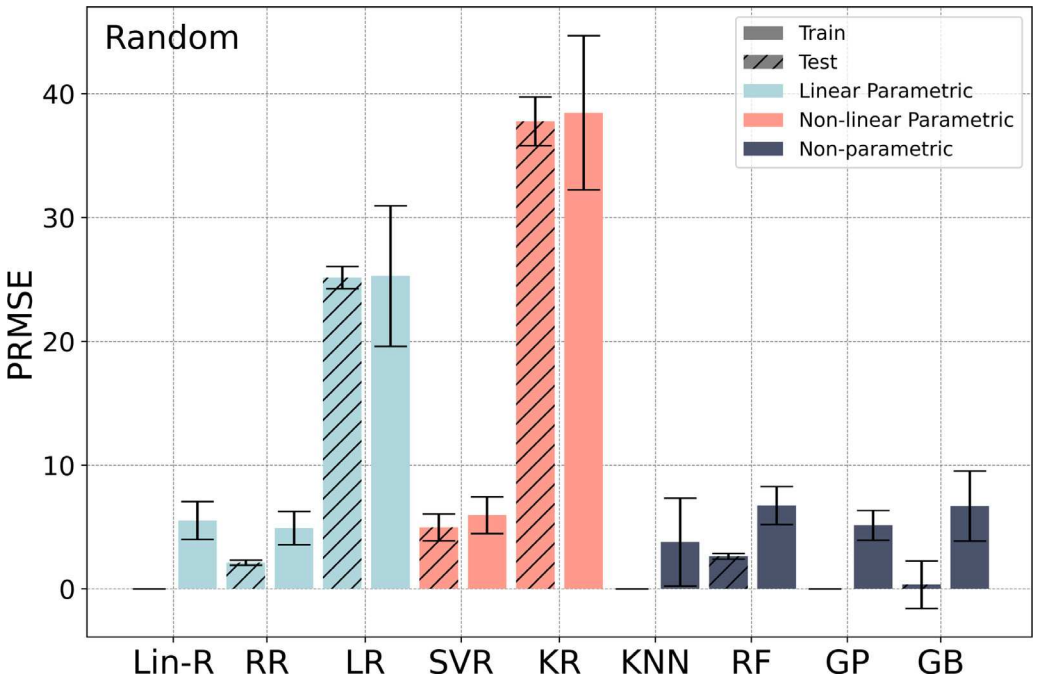

Fig. 9. Performance of regression models in the random splitting scenario. The colors correspond to linear parametric, non-linear parametric, and non-parametric models, respectively. The cross-hatched bars represent performance on training data, whereas the plain bars represent test data performance.

As seen in Fig. 9, Lin-R and RR are the better performers among the linear parametric models. While Lin-R has better train performance than RR (nearly 0 PRMSE), RR demonstrates superior test performance. Thus, Lin-R suffers from overfitting, suggesting that there may be a large number of features with low predictive power present in the A-scan. By using regularization, RR eliminates some of this excess model variability and disregards the uninformative features that are present.

However, LR, which like RR, is a regularized version of Lin-R, performs poorly on both the training and test sets. This is likely due to the nature of the problem. Firstly, A-scan time signals are highly correlated, with multiple adjacent data points corresponding to the interaction of the EM wave with overlapping regions in the material. The fundamental difference between RR and LR is that while RR scales down the coefficients of features it deems unimportant, LR can eliminate coefficients and, in effect, perform feature elimination. As a result, RR performs better on data sets that contain many correlated variables by scaling down the correlated features to reduce model overfitting. LR, on the other hand, chooses certain correlated features over others, creating a sparser model. Having a high number of correlated features means this leads to excessive feature elimination, which creates a less stable model with poor predictive capabilities, as is observed in Fig. 10. Secondly, RR performs better when the number of samples is small compared to the number of features since fewer non-zero parameters must be learned.

As seen in Fig. 10, Lin-R train predictions tightly fit the reference line, but the test predictions are more widely scattered, indicating model overfitting. RR test predictions adhere more closely to the reference line, demonstrating that its regularization

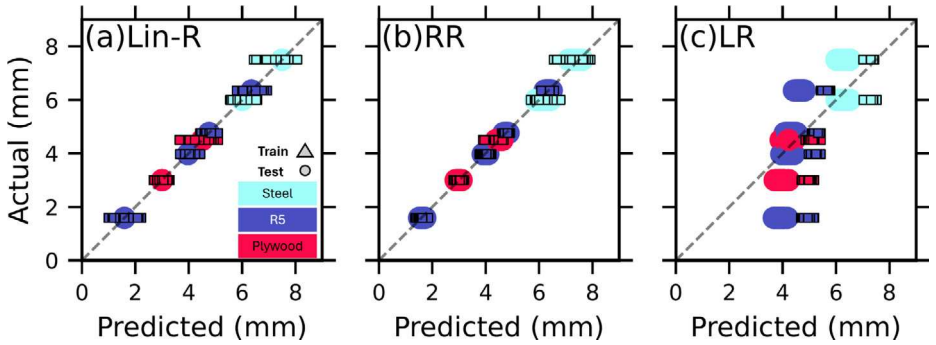

Fig. 10. Performance of linear regression models in the random splitting scenario. RR has lower test error with regularization, but predictions for steel configurations show little improvement. LR's highly biased predictions indicate over-elimination of features.

effectively removes the impact of certain uninformative features. However, the exception is the steel configurations (steel_0.6 cm and steel_0.75 cm), whose layer thickness predictions maintain a significant spread with RR. This is consistent with the results observed for outlier detection and classification, where the same steel configuration scan repetitions were both correctly and incorrectly classified by the same model (Sections: 3.1.2 - Scenario 2 and 3.2). This is possibly due to reflection from steel being noisier due to its greater reflectivity values, which possibly exacerbates surface roughness effects and makes the second steel–air interface difficult to identify. LR train and test prediction are both nearly identical, which suggests that the model is overly biased from the elimination of too many features.

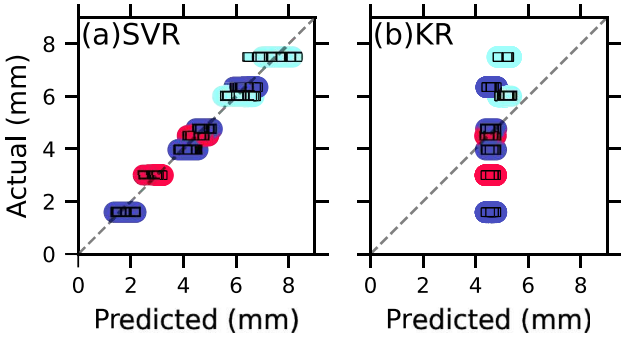

Fig. 11. (Refer to Fig. 10 for legend) Performance of non-linear regression models in the random splitting scenario. KR predicts all configurations to be near the average layer thickness from the training data.

Among the non-linear parametric models, SVR is far superior to KR, though SVR's test performance is still below standard with a mean performance of 8 PRMSE. The poor performance of KR was possibly due to the bare-bones hyperparameter optimization process failing to tune the KR model's kernel shape parameters. The results indicate that the KR model simply averages the layer thickness values in the training data set as observed in Fig. 11 (average layer thickness in the training data set is 0.4735 cm with slight variations due to random shuffling). Like LR, KR test and train predictions are also nearly identical.

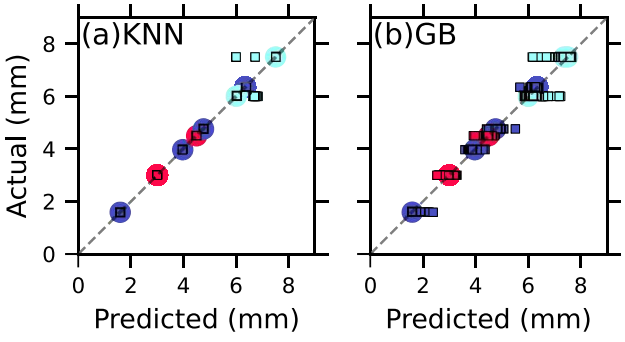

Fig. 12. (Refer to Fig. 10 for legend) Performance of non-linear regression models with randomly split data set. KR predicts all configurations to be near the train set average layer thickness.

Overall, the non-parametric models appear to be the best performers, with RF and GP PRMSE being within the range of accepted value for test cases. KNN has near zero training error but exceeds the reference PRMSE with test data, indicating that the model may suffer from overfitting. However, closer inspection of the results (Fig. 12) reveals that the major contributors to the KNN test errors are the steel configurations. This suggests that neighborhood-based and regularized regression methods such as RR have difficulty interpreting the noisy A-scan signals of conductors such as steel. GB has a high amount of variability in model performance, with training performance variability well beyond the mean PRMSE. Most of this variability results from the steel configurations, as seen in Fig. 12.

3.5.2. Interpolation scenario

Model performances on the interpolation set are significantly worse than the randomly split set. This is expected since the models now encounter A-scans from previously unseen configurations. Surprisingly, training losses are greater than test losses for LR, SVR, and KR (Fig. 13).

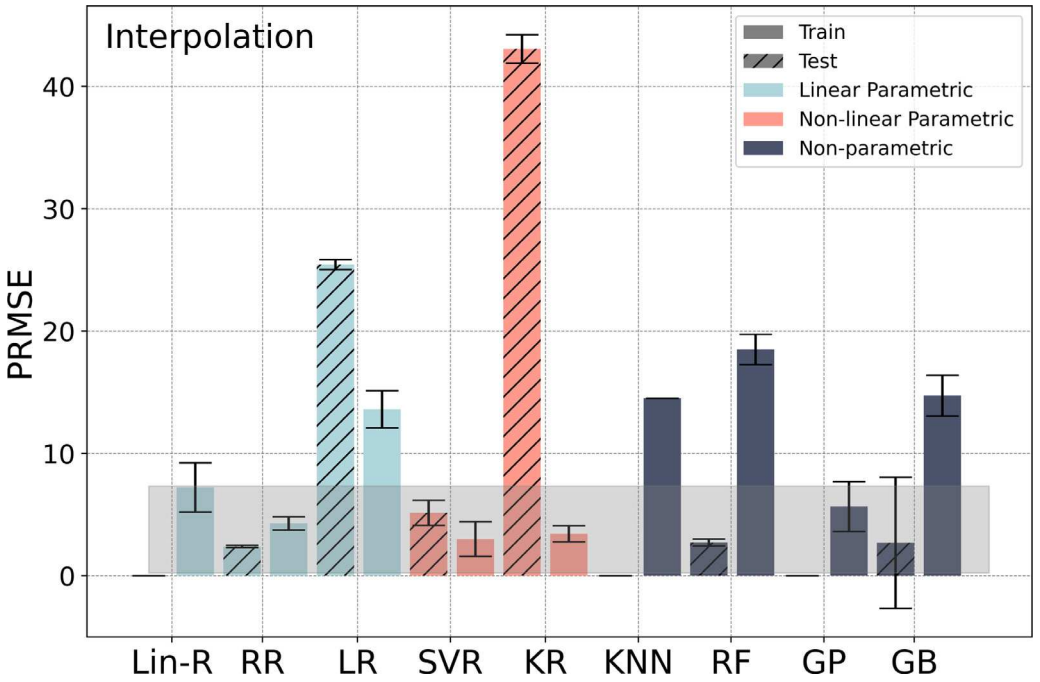

Fig. 13. Model performances for the interpolation scenario. Training losses are greater than test losses for LR, SVR, and KR, indicating no overfitting. Gray-shaded region indicates the best model performance from the preceding scenario, in this case, the random scenario.

Further investigation revealed that this is due to how the different material configurations are split between the train and test set. As seen in Table 14, the training set contains both of the steel configurations, which are hypothesized to produce high amplitude noises in the A-scan signals and have significantly varied predictions in the randomly split data set (Figs. 10, 11).

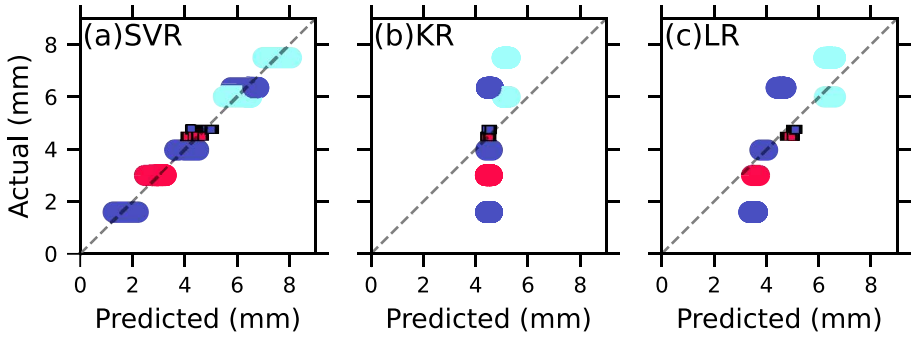

Fig. 14. (Refer to Fig. 10 for legend) SVR training error stems from steel samples. KR and LR simply compute the data set average for prediction. The data set average closely corresponds to the test set region, leading to a misleadingly small test error.

For SVR, as seen in Fig. 14, steel_0.6 cm, steel_0.75 cm, and R5_0.635 cm are the biggest contributors to the training error. KR and LR, conversely, generalize predictions to the mean of the training data set (0.4735 cm), as with the randomly split data set. However, for the interpolation split, the test case layer thicknesses are close to the 0.4735 cm mean (plywood_0.45 cm and R5_0.476 cm), leading to a misleadingly small test error.

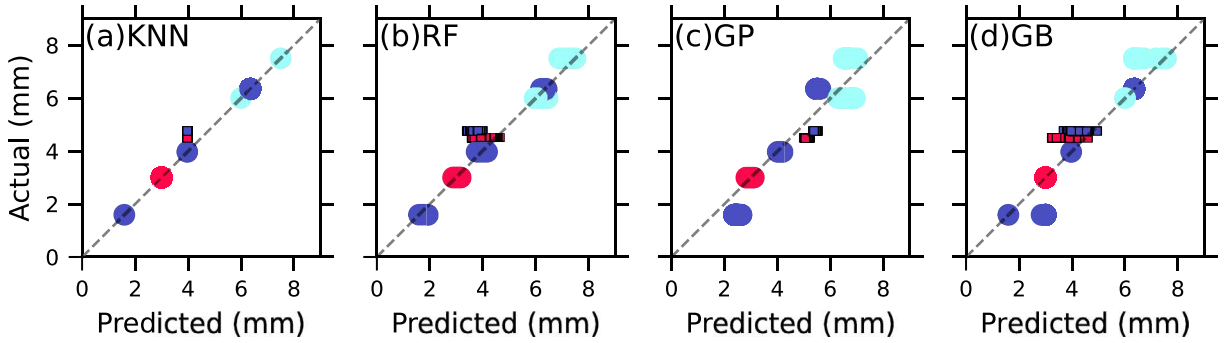

Fig. 15. (Refer to Fig. 10 for legend) Non-parametric model performances. Train predictions narrowly follow the reference line, but large magnitudes offset test predictions.

Overall, non-parametric models do not perform as well for the interpolation set as they do for the randomly split cases. A large discrepancy is observed between their train and test performances (see Fig. 15), indicating that all the non-parametric models may suffer from overfitting. This is probably because non-parametric models use more learned parameters, which makes them prone to overfitting in a small data regime such as this problem. The reason why overfitting occurs for the interpolation data set but not for the randomly split data set may be because the A-scans encountered in the test cases are not significantly different than the train set scans for the randomly split data set. They are generated by repeating scans from the same material configuration, meaning fewer aberrant features to overfit.

3.5.3. Extrapolation scenario

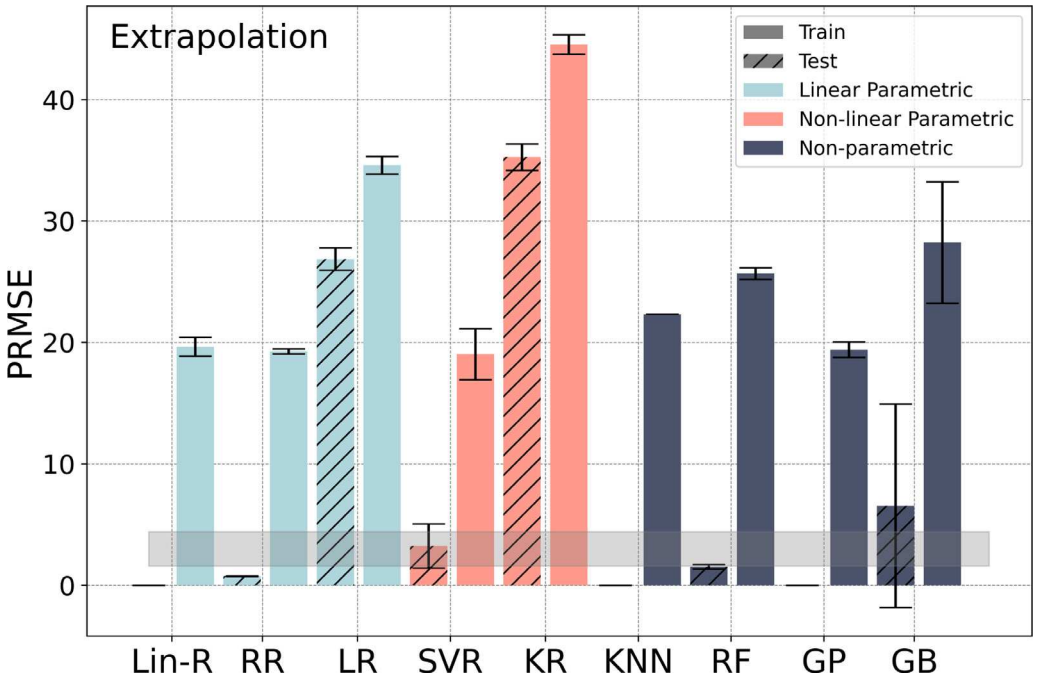

Fig. 16. Model performances on extrapolation scenario. All models show significantly worse performance compared to the random and interpolation scenarios. The shaded area represents the best performance range from the interpolation scenario.

Performance in the extrapolation scenario is worse than in the other scenarios (Fig. 16). This is expected because the modeled trends between A-scans and layer thickness values can only propagate to a certain extent without relevant training data. KNN, RF,

and GB, non-parametric models that performed satisfactorily for both the interpolation and random split scenarios, performed poorly for the extrapolation scenario (Fig. 17). It seems that the absence of adequate data at the extremes of the training data leaves these models unable to generalize effectively. Parametric models, on the other hand, can generalize trends more effectively with less data by making assumptions about the distribution of the data. LR, RR, and SVR, which had performed well for the random/interpolation sets, predicted the R5_0.635 cm layer thickness accurately but failed for the noisy steel_0.75 cm A-scans (Fig. 18).

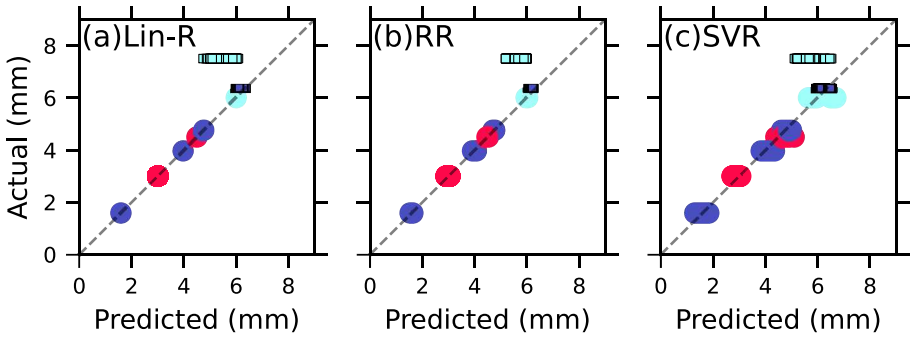

Fig. 17. (Refer to Fig. 10 for legend) Non-parametric model performances as parity plots.

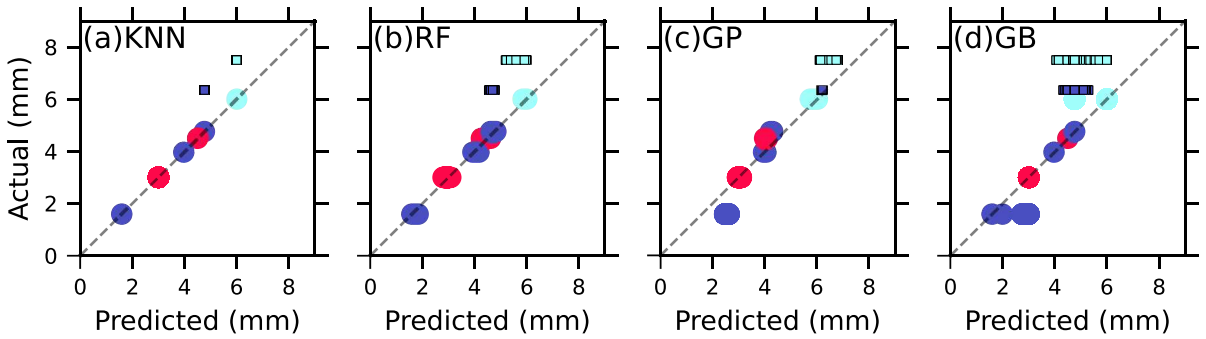

Fig. 18. (Refer to Fig. 10 for legend) Parametric model performances as parity plots. Train predictions narrowly follow the reference line, but large magnitudes offset test predictions.

Overall, all three models (linear and non-linear parametric, plus non-parametric) exhibited performances aligned with recent literature for layer thickness predictions. Certain kernel-based models show strikingly poor performance for all test–train splits, indicating that kernel shape tuning strongly impacts the performance of non-ensemble models. Predictions for steel are consistently inferior compared to other material samples across all models, suggesting that high reflectivity may obscure useful interfacial signals. Greater surface and subsurface roughness and inhomogeneities, likely encountered during field applications, may exacerbate prediction inaccuracy for metallic conductors in general. Thus, additional signal processing steps targeted toward noise reduction may be applied to scans identified as conductors by the material classifier. Lastly, extrapolation case results indicate that parametric models may be better than non-parametric models in cases of data sparsity (see Fig. 17).

3.5.4. Null hypothesis

The sample size in this work is small (80 in-domain scans) relative to the total number of features present in each data point (655 features). Likewise, the ML models, with a few exceptions, also have many parameters compared to the sample size. Thus, there is a possibility, given the "black-box" nature of ML algorithms, that the accurate predictions produced are not the result of any physically meaningful features that are captured in the A-scans but instead are a product of having a large number of parameters fit a small quantity of data. If that is the case, the layer thickness prediction results obtained in the previous section can be repeated for any possible combination of the data set. Thus, to test this null hypothesis, layer thickness values are randomly shuffled, and the models are retrained.

As seen in Fig. 19, Lin-R results were hugely out of bounds, while RR produces significantly worse predictions than all three split types discussed in the previous sections, and LR outputs the training set average as before. Similarly, poor performances are also observed for the interpolation split.

To further test the null hypothesis, layer thickness values were swapped between configurations(0.3 cm plywood was assigned a thickness value of 0.75 cm corresponding to steel_0.75 cm, 0.6 cm steel was assigned a thickness value of 0.159 cm corresponding to

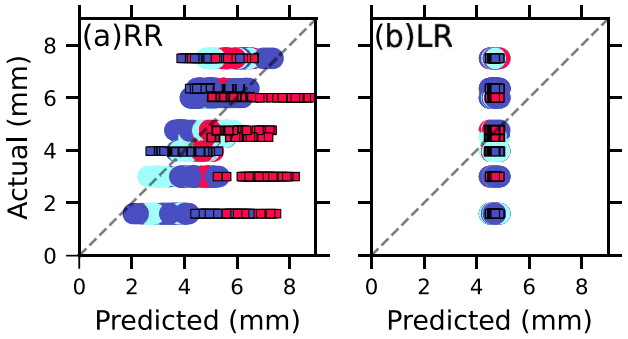

Fig. 19. (Refer to Fig. 10 for legend) Testing the null hypothesis with random shuffling — RR predictions have huge errors for both training and test sets. LR predicts the mean layer thickness for all samples. The stark contrast between these results and Fig. 10 counters the null hypothesis.

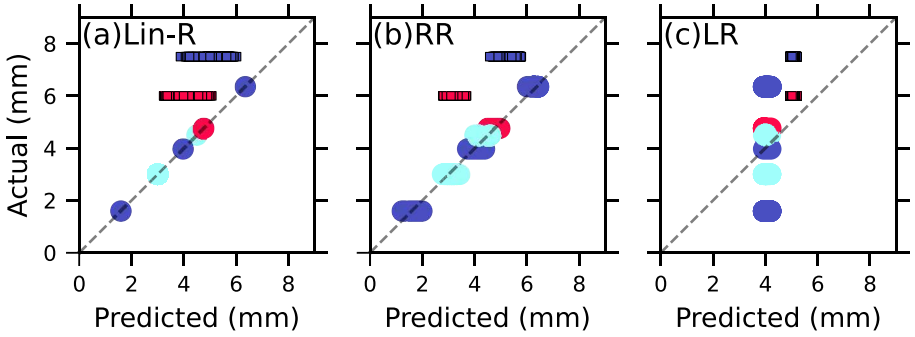

Fig. 20. (Refer to Fig. 10 for legend) Testing the null hypothesis with specific but erroneous swapping. Training error is low for Lin-R and RR, but test predictions are off by quite a large margin. LR outputs the mean of the training set as before.

R5_0.159 cm and so on). This ensured that scans produced by the same configuration had the same erroneous layer thickness value, mirroring the repetitions in the original data set. As seen in Fig. 20, training performance is much better than test performance; due to the small number of samples and a large number of features, the models can fit very well to the patterns in the training data. However, the random swapping of layer thickness values disrupts the general trends in the A-scans accompanied by thickness variations, which causes test predictions to be poor. This invalidates the null hypothesis, proving that the ML models operate on features salient to the material type and layer thickness values in the A-scan.

4. Model interpretation

The dataset comprised sample thicknesses ranging from 0.1 to 0.8 cm, which is notably smaller than the wavelength of the emitted GPR waves; the smallest emitted wavelength is about 5 cm (corresponding to a frequency of 6 GHz), which is an order of magnitude greater than the typical sample thickness. Despite this limitation, trained ML models could still make accurate predictions of the layer thicknesses, even on held-out test cases. One might argue that this indicates memorization within the models, i.e., that with sufficient trainable parameters, the models can obtain a good fit for any data. However, as seen in the null hypothesis test in Section 3.5.4, randomly swapping the input/output data leads to significant performance degradation. Moreover, the models also demonstrated generalization in their performance on unseen labels, particularly for the interpolation cases.

An inspection of the A-scans revealed that the difference in sample thickness values led to predictable variations in the A-scans, despite the small dimensions of the sample thicknesses compared to the wavelength of the incident wave. As seen in Fig. 21, the varying thickness values shift the A-scan waveform relative to the layer thickness. This shift becomes much more significant at 3.82 ns, where a secondary peak forms for all samples. A lower amplitude and, therefore, higher attenuation is also observed for the higher thickness samples.

In thicker samples, the second air–material interface is located further away from the receiver, and the waves propagate through the material longer than in thin samples. Since the dielectric permittivity values of our materials are much higher than air, waves propagate at a slower speed through the sample. Thus, signals reach the receiver later in time for thicker samples. The A-scan shifts in Fig. 21 fit this basic wave propagation model. However, the time difference between the initial oscillations and the second peak is too large for the secondary peak signals to be a product of primary reflections from the second air–material interface. Thus, it was hypothesized that the second reflection, highlighted in the above diagram, corresponds to incident reflections from the surrounding objects, such as the supports and floor, given the separation between the initial oscillation and the secondary peak.

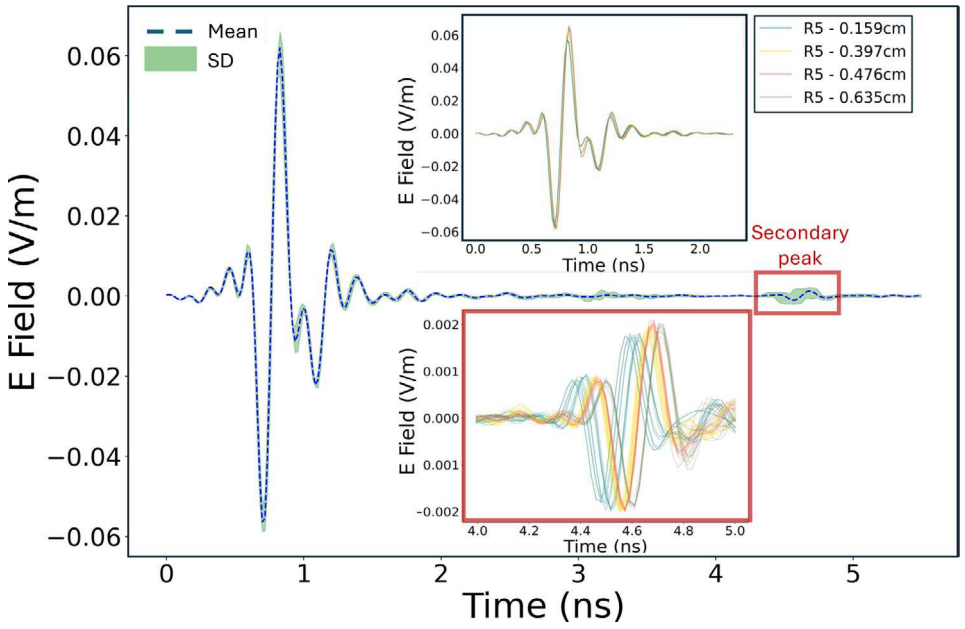

Fig. 21. A-scans of the R5 material samples, with different thickness values (up to 5.5 ns shown due to relevance). (Insets) A closer look at the A-scan signals at time windows 0–2.2 ns and 4–5 ns (secondary peak), respectively. The variations that correlate with layer thickness are visible here in close range.

To test this, the samples in the experimental setup were simulated using gprMax [51], a widely used open-source software for simulating GPR wave propagation. The electric field magnitude was rendered throughout the experimental domain as a function of time to visualize the wave propagation. Snapshots of the animation at different time windows are shown in Fig. 22.

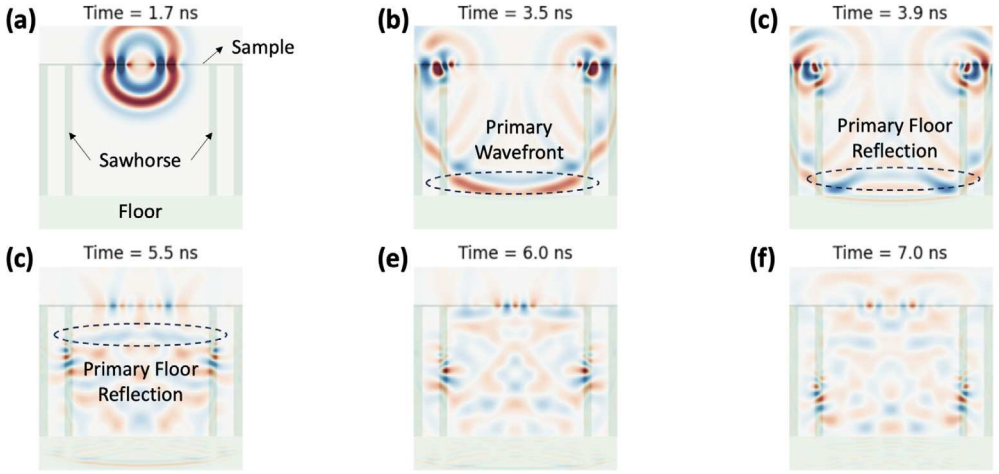

Fig. 22. R5 sample of 0.8 cm thickness was simulated in gprMax, with other configurations such as supporting sawhorses, floor distance, and surrounding materials representative of the experimental setup. The propagation of the wave throughout the domain at different times is described in each image. A pale green color indicates the sample, sawhorse, and floor geometry, whereas the electric field is shown in a red–white–blue color map (red positive, blue negative *E*).

As seen in Fig. 22, reflections from the sawhorses and the floor reach the receiver between 3.8 and 6.9 ns. The primary reflection from the floor reaches the transmitter at 6 ns, which corresponds loosely to the highest secondary peaks in the experimental A-scan at around 4.8 ns (Fig. 21). Several parameters vary between the experimental setup and the simulated setup, such as the true permittivity values, surface roughness, and material inhomogeneity and imperfections, which will necessarily explain the slight discrepancies between the signals.

Thus, the simulated animation supported the hypothesis that the second peak results from nearby reflections. Further confirmation was obtained by varying the sawhorse height above the floor for the experimental setup and observing the corresponding shifts

in the A-scan. Increasing the height of the sample above the floor leads to a delay in the second peak, as would be expected of the reflection from the floor.

However, a strong reliance on secondary sources of reflection is problematic, particularly when the diagnostic EM environment can be cluttered and inconsistent. To test whether the proposed ML algorithms are truly dependent on secondary reflections for predictions, a feature-trimming exercise using an evolutionary algorithm approach was implemented (Section 5). The linear regression model was used for this purpose due to its simplicity and interoperability.

4.1. Comparison to conventional methods

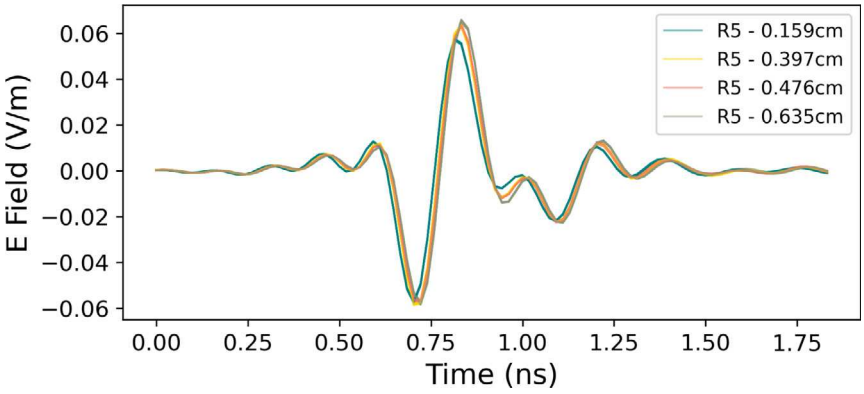

Fig. 23. Top inset enlarged from Fig. 21. No distinctly separated peaks appear in A-scans for different sample thicknesses. Instead, the signal peaks are shifted in time by much less than the difference in thickness would suggest.

The A-scans in the dataset do not contain distinct peaks corresponding to each material interface (such as the boundary between the material sample and air). This is illustrated in Fig. 21 inset, which is enlarged in Fig. 23. The sample thickness values investigated in this study, on the order of 1 cm or less, are too small compared to the input signal frequency (as well as the sampling frequency) of the GPR equipment for distinct peaks to occur at each interface. Instead, the signals display a shift commensurate to layer thickness due to increased propagation through the slower dielectric medium of the material layer. For instance, the waves should propagate at about 20 cm/ns in the plywood samples, passing entirely through a 1 cm sample in only 0.05 ns. Yet from Fig. 23, it is clear that there is no significant signal variation at these early times. Deconvolving the many reflections that can occur by the more discernible 0.5 ns time would impose a significant signal analysis challenge.

Layer thickness can be estimated using the two-way travel time, which involves calculating the distance between peaks corresponding to a material interface. In the absence of distinct peaks corresponding to each material interface, it would be problematic to predict layer thickness from A-scans of samples with such low thickness without additional signal processing. Moreover, the two-way travel time calculation requires knowledge of the material permittivity, which must be measured, such as via controlled experiments with a copper plate. However, it will not be reliable for material samples thinner than half the wavelength. In contrast, our data-driven models leverage the entire waveform instead of information at distinct peaks to utilize subtle shifts in A-scans to accurately predict the material and thickness predictions even for very thin material layers while imposing a smaller data collection burden.

5. Feature analysis

Each GPR A-scan consists of 655 data points, one corresponding to each time sample. Since adjacent time points share mutual information due to the physical constraint of wave propagation, not all 655 time samples are independent. Therefore, we suspect that only a small fraction of these points are necessary for predictive modeling. In addition, later segments of the A-scan will have a low signal-to-noise ratio, while other segments throughout the A-scan may be irrelevant for diagnostics due to being the product of interactions with unrelated geometric features. Thus, feature elimination can reduce unnecessary complexity in the models and potentially improve performance by ignoring spurious features.

5.1. Feature elimination

A GA is employed to obtain the optimum set of feature combinations that produce the best performance for the linear regression model. The fitness function F is constructed to penalize both poor model performance and a high number of model features:

$$F = \alpha \langle R^2 \rangle_{CV} - n_f \tag{6}$$

where α and β are hyperparameters to control the relative contribution from each term, $\langle R^2 \rangle_{CV}$ is the average R^2 over cross-fold validation with a given set of features, and n_f is the number of time samples in the A-scan that the GA selects for the model. The

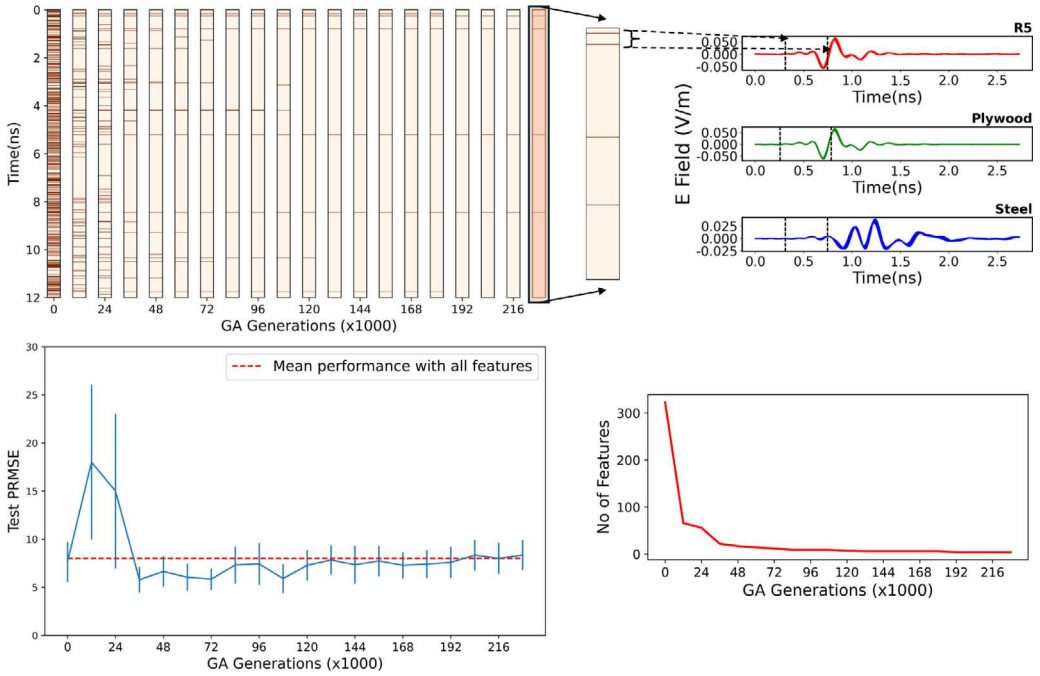

Fig. 24. (Top left) GA optimized features, where dark indicates "on" and light indicates "off", and (Bottom left) corresponding PRMSE scores over GA generations. (Top right) Position of the two dominant features in the A-scan distributions for all materials in the study. (Bottom right) Number of model features over GA generations.

GA is run for 240,000 iterations, where it tries to maximize the fitness function by improving the R^2 score of the model and/or decreasing the number of features. The model performance on varying interpolation splits is used to calculate the fitness function since random split performance may not indicate true performance in the field, and extrapolation performances were unstable for certain cases.

The GA optimizer selects four features among the 655 time samples that produce similar model performance compared to the entire A-scan, shown in Fig. 24. Among these four, through the process of elimination, two key features (shown inset) were critical for performance. Unfortunately, the GA failed to narrow down to these two features by itself due to the very large solution landscape with so many features (i.e., 2^{655} possible solutions). More sophisticated methods designed for high-dimensional spaces, such as Monte Carlo Tree Search, may be warranted.

The two key features occur at 0.26 ns and 0.79 ns, respectively. According to simulated results in Section 4, these signals are too early to be a product of nearby reflections from irrelevant objects, providing evidence that our models can perform well even with a cluttered far field, based on subtle signal variations due to reflections in the near field (i.e., within the material sample).

Typically, analytical calculations for thickness involve two signal samples from the dominant reflected waveform, each corresponding to an interface of the material of interest, along with permittivity calculations. Thus, the position of the first feature is unexpected, since there are no significant perturbations at 0.26 ns (Fig. 24 - Top right). It may be that because the depth of the top interface is unchanged relative to the GPR receiver antenna position in the training data, only the impact of the second interface on the waveform is required for calculating layer thickness. Thus, the first feature may simply be a reference point for the second feature.

5.2. Feature importance

The GA-based feature selection is a highly stochastic process and hence provides no guarantee that the optimized features will be robust in real-world applications. To evaluate their significance, the stability of model performance with the GA-selected feature sets using different interpolation train-test splits was explored. The linear regression model was run 100 times for each feature set, with a random selection of the interpolation boundary and random shuffling of the training data.

As seen in Fig. 25 (Bottom - No noise), the model loss remained low and relatively consistent for all listed generations, indicating that the salience of the selected features is robust to small random variations.

The uniqueness of the feature sets was tested further by randomly perturbing the GA-selected A-scan features by varying amounts (Fig. 25 - Bottom - ±1, ±2, and ±3) and re-training the models. All feature sets produced worse and more unstable performance with increased random shifts. This indicated that the selected features are significant and unique since shifting them even slightly heavily impacts performance.

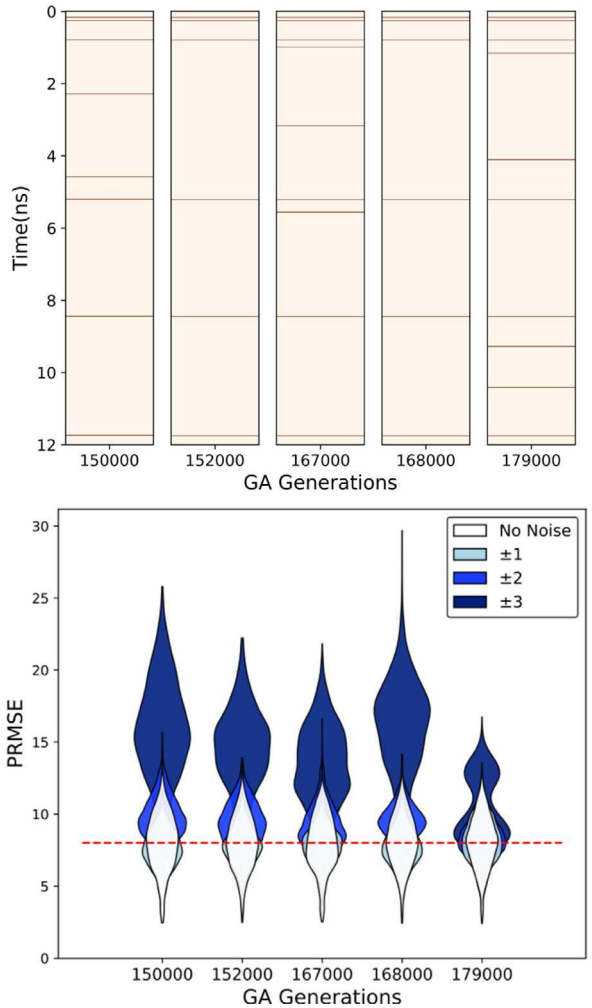

Fig. 25. (Top) Positions of GA optimized features in the A-scans at generations near convergence. (Bottom) The white violin plots with no noise represent corresponding PRMSE and their variability to different train–test splits. PRMSE variance for different levels of random shifts (±1, ±2, and ±3) in the GA-selected feature sets are demonstrated by the violin plots with varying shades of blue. All GA-selected feature sets show increasingly poor and erratic model performance with increased random shifts. The dotted red line represents the mean performance of Lin-R model on the interpolation set.

6. Conclusion

In this work, a data-driven framework for predicting one-layer material type and thickness from GPR A-scans was proposed and systematically evaluated. An accuracy of 99% was achieved for the detection of outlier samples and 86% for in-domain material classification. Significant variations in performance were observed as a function of the training protocol. Sample thickness prediction with performance similar to the current best in literature (around 5 PRMSE) was also achieved for randomly shuffled and interpolation scenarios. For extrapolation cases, the minimum model error was about 20 PRMSE, although this value may differ systematically between materials.

Unlike classification, the layer thickness regression performance was found to be highly dependent on model hyperparameters. Lastly, a stochastic search method was used to prune features from the A-scan that do not contribute to the model performance. It was found that peak model performance can be achieved with just two time samples, which constitute only 0.3% of the A-scan. Simulations indicated that these critical signals were not the by-product of any nearby reflections but were purely dependent on material sample thickness.

The model performances indicate that the proposed models would be highly effective for investigating building envelope conditions, given that adequate reference data are available to train the models. We have proposed a straightforward procedure for gathering data, training suitable models, and evaluating their performance in a variety of test scenarios. These procedures have a relatively low data collection burden and can be feasibly implemented in real-world situations. However, users should be cognizant of caveats such as the low level of certainty regarding the classification of materials with similar permittivities and the relatively low

accuracy of layer thickness prediction for conductive materials. The model interpretation in Section 4 analyzed the training data and model performance in detail and found that the signals crucial to model performance occur too early to be the result of surrounding clutter within reasonable proximity, indicating the validity of the proposed approach in cluttered environments. However, the applicability of training data for instances with significant environmentally-induced variation, such as moisture penetration and corrosion, can be a part of future work, though we believe it can be addressed via collection of a few representative examples in the training set.

Additional future work in this vein should consider including new construction material samples, such as concrete and fiberglass. Moreover, it is presently unclear how well similar models would perform on multi-layer or even more complex structures. Also, given the homogeneity of the material samples, it was assumed that permittivity is the sole attribute that distinguished materials. However, this may not always be the case in real-world scenarios where geometric irregularities such as surface/subsurface roughness and porosity create unique signals. Thus, the current work can also be expanded by applying the proposed methodology to more classes of building materials as well as material samples with varying properties such as moisture content, age, porosity, and so forth. However, this would greatly expand the complexity of the study, as additional labels would need to be measured in the experiment and then predicted by supervised learning. In this case, it may be difficult to control model overfitting without a much larger dataset.

Moreover, applying higher resolution signals and multi-offset GPR may enable the successful identification of such materials with only slight changes to model architectures and training protocols. Additional future work can also include the investigation of spectral analysis to distinguish between materials with similar permittivity by leveraging how each material interacts with radar signals of varying frequency. In field application scenarios, defects such as cracks, voids, and delamination may likely be present in the building envelope materials. The presence of such defects would lead to unpredictably perturbed signals that would heavily interfere with the prediction methods in the proposed methodology. Thus, another course of future work could involve modifying the present work to make predictions robust to such disturbances.

CRediT authorship contribution statement

Ahmed Nirjhar Alam: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wesley F. Reinhart:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Rebecca K. Napolitano:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This material is based upon work supported by the National Science Foundation, United States under Grant IIS-2123343 and by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Building Technologies Office, Award Number DE-EE0009748.

References

[1] Ingy El-Darwish, Mohamed Gomaa, Retrofitting strategy for building envelopes to achieve energy efficiency, Alexandria Eng. J. 56 (4) (2017) 579–589.
[2] Ö. Altan Dombaycı, Mustafa Gölcü, Yaşar Pancar, Optimization of insulation thickness for external walls using different energy-sources, Appl. Energy 83 (9) (2006) 921–928.
[3] A. Hussain, S. Akhtar, Review of non-destructive tests for evaluation of historic masonry and concrete structures, Sci. Eng. 42 (2017) 925–940.
[4] Ayca Kirimtat, Ondrej Krejcar, A review of infrared thermography for the investigation of building envelopes: Advances and prospects, Energy Build. 176 (2018) 390–406.
[5] Angeliki Kylili, Paris A. Fokaides, Petros Christou, Soteris A. Kalogirou, Infrared thermography (IRT) applications for building diagnostics: A review, Appl. Energy 134 (2014) 531–549.
[6] D. Andersson, N. Bjorsell, P. Ottonson, D. Ronnow, M. Sandberg, Radar images of leaks in building elements, Energy Procedia 78 (2015) 1726–1731.
[7] N. Bjorsell D. Ronnow, B. Laporte-Fauret, Determination of elongation of electrically small objects in building structures by polarimetric synthetic aperture radar, in: EEE International Instrumentation and Measurement Technology Conference, 2017.
[8] R. et al Agliata, On-invasive estimation of moisture content in tuff bricks by GPR, eConstr. Build. Mater. 160 (2018) 698–706.
[9] Imad L. Al-Qadi, Samer Lahouar, Amara Loulizi, Successful application of ground-penetrating radar for quality assurance-quality control of new pavements, Transp. Res. Rec. 1861 (1) (2003) 86–97.
[10] Samer Lahouar, Imad L. Al-Qadi, Amara Loulizi, Trenton M. Clark, David T. Lee, Approach to determining in situ dielectric constant of pavements: Development and implementation at interstate 81 in virginia, Transp. Res. Rec. 1806 (1) (2002) 81–87.

[11] Hai Liu, Motoyuki Sato, In situ measurement of pavement thickness and dielectric permittivity by GPR using an antenna array, NDT E Int. 64 (2014) 65–71.
[12] Hai Liu, Motoyuki Sato, In situ measurement of pavement thickness and dielectric permittivity by GPR using an antenna array, NDT E Int. 64 (2014) 65–71.
[13] Shan Zhao, Pengcheng Shangguan, Imad L. Al-Qadi, Application of regularized deconvolution technique for predicting pavement thin layer thicknesses from ground penetrating radar data, NDT E Int. 73 (2015) 1–7.
[14] Shan Zhao, Imad L. Al-Qadi, Siqi Wang, Prediction of thin asphalt concrete overlay thickness and density using nonlinear optimization of GPR data, NDT E Int. 100 (2018) 20–30.
[15] Ziaul Hasan, Hong Jie Xing, M. Idrees Magray, Big data machine learning using apache spark mllib, Mesop. J. Big Data 2022 (2022) 1–11.
[16] Hafiz Suliman Munawar, Fahim Ullah, Amirhossein Heravi, Muhammad Jamaluddin Thaheem, Ahsen Maqsoom, Inspecting buildings using drones and computer vision: A machine learning approach to detect cracks and damages, Drones 6 (1) (2022).
[17] Jinrui Zhang, Mengxi Zhang, Biqin Dong, Hongyan Ma, Quantitative evaluation of steel corrosion induced deterioration in rubber concrete by integrating ultrasonic testing, machine learning and mesoscale simulation, Cem. Concr. Compos. 128 (2022) 104426.
[18] Giuseppe Ciaburro, Gino Iannace, Machine-learning-based methods for acoustic emission testing: A review, Appl. Sci. 12 (20) (2022).
[19] T. Noreen, Umar S. Khan, Using pattern recognition with HOG to automatically detect reflection hyperbolas in ground penetrating radar data, in: 2017 International Conference on Electrical and Computing Technologies and Applications, ICECTA, 2017, pp. 1–6.
[20] Xavier Núñez-Nieto, Mercedes Solla, Paula Gómez-Pérez, Henrique Lorenzo, GPR signal characterization for automated landmine and UXO detection based on machine learning techniques, Remote Sens. 6 (10) (2014) 9729–9748.
[21] Nitin Ajithkumar, P. Aswathi, Rao. R. Bhavani, Identification of an effective learning approach to landmine detection, in: 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech, 2017, pp. 1–5.
[22] E. Costamagna, P. Gamba, S. Lossani, A neural network approach to the interpretation of ground penetrating radar data, in: IGARSS '98. Sensing and Managing the Environment. 1998 IEEE International Geoscience and Remote Sensing. Symposium Proceedings. (Cat. No.98CH36174), Vol. 1, 1998, pp. 412–414.
[23] P. Gamba, S. Lossani, Neural detection of pipe signatures in ground penetrating radar images, IEEE Trans. Geosci. Remote Sens. 38 (2) (2000) 790–797.
[24] Shuwei Li, Xingyu Gu, Xiangrong Xu, Dawei Xu, Tianjie Zhang, Zhen Liu, Qiao Dong, Detection of concealed cracks from ground penetrating radar images based on deep learning algorithm, Constr. Build. Mater. 273 (2021) 121949.
[25] Umut Ozkaya, Farid Melgani, Mesay Belete Bejiga, Levent Seyfi, Massimo Donelli, GPR B scan image analysis with deep learning methods, Measurement 165 (2020) 107770.
[26] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, 2016, arXiv:1506.02640.
[27] Bin Liu, Yuxiao Ren, Hanchi Liu, Hui Xu, Zhengfang Wang, Anthony G. Cohn, Peng Jiang, GPRInvNet: Deep learning-based ground-penetrating radar data inversion for tunnel linings, IEEE Trans. Geosci. Remote Sens. 59 (10) (2021) 8305–8325.
[28] Senlin Yang, Zhengfang Wang, Jing Wang, Anthony G. Cohn, Jiaqi Zhang, Peng Jiang, Lichao Nie, Qingmei Sui, Defect segmentation: Mapping tunnel lining internal defects with ground penetrating radar data using a convolutional neural network, Constr. Build. Mater. 319 (2022) 125658.
[29] What is dielectric constant of plastic materials. https://passive-components.eu/what-is-dielectric-constant-of-plastic-materials/.
[30] Miscellaneous dielectric constants. https://www.microwaves101.com/encyclopedias/miscellaneous-dielectric-constants.
[31] Dileep Kumar, Morshed Alam, Patrick X.W. Zou, Jay G. Sanjayan, Rizwan Ahmed Memon, Comparative analysis of building insulation material properties and performance, Renew. Sustain. Energy Rev. 131 (2020) 110038.
[32] Pierre Blanchet, Cedric Perez, Matheus Roberto Cabral, Wood building construction: Trends and opportunities in structural and envelope systems, Curr. Forestry Rep. 10 (2024).
[33] Building envelopes. https://www.steelconstruction.info/Building~envelopes.
[34] Screening eagle. https://www.screeningeagle.com/.
[35] Yilmaz Ö, Seismic data analysis, Soc. Explor. Geophys. (2001).
[36] A.P. Anana, Sensors & Software Inc, Practical processing of GPR data, in: Proceedings of the Second Government Workshop on Ground Penetrating Radar.
[37] Zhongming Xiang, Ge Ou, Abbas Rashidi, Robust cascaded frequency filters to recognize rebar in GPR data with complex signal interference, Autom. Constr. 124 (2021) 103593.
[38] Deniz Kumlu, Isin Erer, Improved clutter removal in GPR by robust nonnegative matrix factorization, IEEE Geosci. Remote Sens. Lett. 17 (6) (2020) 958–962.
[39] Shan Zhao, Pengcheng Shangguan, Imad L. Al-Qadi, Application of regularized deconvolution technique for predicting pavement thin layer thicknesses from ground penetrating radar data, NDT E Int. 73 (2015) 1–7.
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[41] AX adaptive experimentation platform. https://ax.dev/.
[42] Jason Brownlee, Master Machine Learning Algorithms, Melborne, Australia, 2016.
[43] John H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–73.
[44] Hassan Sarmadi, Alireza Entezami, Filipe Magalhães, Unsupervised data normalization for continuous dynamic monitoring by an innovative hybrid feature weighting-selection algorithm and natural nearest neighbor searching, Struct. Health Monit. 22 (6) (2023) 4005–4026.
[45] Hassan Sarmadi, Alireza Entezami, Bahareh Behkamal, Carlo De Michele, Partially online damage detection using long-term modal data under severe environmental effects by unsupervised feature selection and local metric learning, J. Civ. Struct. Health Monit. 2022 (2022) 1043–1066.
[46] Nairit Barkataki, Ankur Jyoti Kalita, Utpal Sarma, Automatic material classification of targets from GPR data using artificial neural networks, in: 2022 IEEE Silchar Subsection Conference, SILCON, 2022, pp. 1–5.
[47] Mohamed S. El-Mahallawy, Mazlan Hashim, Material classification of underground utilities from GPR images using DCT-based SVM approach, IEEE Geosci. Remote Sens. Lett. 10 (6) (2013) 1542–1546.
[48] Shan Zhao, Imad L. Al-Qadi, Siqi Wang, Prediction of thin asphalt concrete overlay thickness and density using nonlinear optimization of GPR data, NDT E Int. 100 (2018) 20–30.
[49] Building tolerances. https://www.engineeringtoolbox.com/building-tolerances-d{_}1790.html.
[50] Glen-gery - brick sizes. https://www.glengery.com/brick-sizes.
[51] Craig Warren, Antonios Giannopoulos, Iraklis Giannakis, gprMax: Open source software to simulate electromagnetic wave propagation for ground penetrating radar, Comput. Phys. Comm. 209 (2016) 163–170.