ORIGINAL PAPER

Machine learning assisted detection and localization of mechanical discontinuity

Rui Liu · Siddharth Misra

Received: 16 January 2022 / Accepted: 1 June 2022 / Published online: 27 June 2022

Abstract Accurate detection and localization of mechanical discontinuities are essential for industries dependent on natural, synthetic and composite materials, e.g. construction, aerospace, oil and gas, ceramics, metal, and geothermal industries, to name a few. In this study, a physics-informed machine learning workflow is developed for detecting and locating single, linear mechanical discontinuity in homogeneous 2D material by processing the full-waveforms recorded during multi-point compressional/shear transmission measurements. This work is based on fundamental aspects of simulation of wave propagation, signal processing, feature engineering, and data-driven model evaluation. k-Wave simulator is implemented to model the compressional and shear wave transmission through the 2D numerical model of a material containing single mechanical discontinuity. For a specific source-sensor configuration, the newly developed data-driven workflow can detect and locate the mechanical discontinuity with an accuracy higher than 0.9 in terms of coefficient of determination. AdaBoost regressor with k-Nearest Neighbor as a base estimator significantly outperforms all other models. In terms of sensitivity to noise, k-Nearest Neighbor is the most robust to both gaussian and uniform distributed noise.

Keywords Machine learning · Regression · Wave propagation · Mechanical discontinuity · Signal processing · Feature engineering

R. Liu (✉)
Harold Vance Department of Petroleum Engineering, College of Engineering, Texas A&M University, College Station, TX, USA
e-mail: rui81@tamu.edu

S. Misra
Harold Vance Department of Petroleum Engineering, College of Engineering, Texas A&M University, College Station, TX, USA
e-mail: misra@tamu.edu

S. Misra
Department of Geology and Geophysics, College of Geosciences, Texas A&M University, College Station, TX, USA

1 Introduction

It is important to detect, locate, and characterize the mechanical discontinuities in natural, synthetic, composite and engineering materials. For purposes of energy security and environmental sustainability, the U.S. Department of Energy (DOE) Office of Basic Energy Science, Office of Fossil Energy, and Geothermal Technologies Office & U.S. National Science Foundation (NSF) Division of Earth Sciences emphasize the need to understand, predict, & control the mechanical discontinuities in subsurface (Pyrak-Nolte et al. 2015). From a geophysical standpoint, mechanical discontinuities refer to mechanical separation or interface, such as joints, fractures, and bedding planes (Osogba et al. 2020). Such discontinuities occur at various scales from interfaces of mineral to tectonic plates that are generated in various forms by several distinct processes. Mechanical discontinuities are important for producing the subsurface earth resources because they provide potential transport pathways and determine the bulk mechanical, physical and chemical behavior of the subsurface system. Various aspects of subsurface engineering rely heavily on robust characterization and control of mechanical discontinuities in subsurface.

In the petroleum and geothermal industries, a wide variety of materials, tools and techniques are utilized to identify, map, and characterize the induced and natural fractures. Laboratory tests are carried out to detect discontinuities, observe the failure mechanisms during uniaxial compressive strength tests, and analyze the factors affecting their mechanical strength (Szwedzicki and Shamu 1999). Other techniques for detecting fractures use core data, borehole image log (Kabir et al. 2009), seismic section (Kanasewich and Phadke 1988), well logs (Shalaby and Islam 2017), in-situ stress data, and well flow tests. However, each technique has its own limitations, e.g. the study of conventional logs suffers from a low spatial resolution and direct study of cores and image logs is associated with high costs (Kosari et al. 2015).

For purposes of extraction of subsurface earth resources, the mechanical discontinuities are characterized at different scales. Well testing and seismic tomography is used to characterize the beddings, joints, and faults at meter to kilometer scale. Resistivity/dielectric imaging is a well logging technique used to quantify the beddings and fractures in the near-wellbore region at centimeter to meter scale. In the civil engineering discipline, sonic and ultra-sonic waves are utilized to measure the discontinuities in steel, concrete, and other materials at centimeter scale for purposes of structure health monitoring. Many non-destructive tests are developed for the characterization of cracks at millimeter scale such as acoustic emission (AE) monitoring (Godin et al. 2018), ultrasonic imaging, and computer tomography (CT) scanning have been developed and applied concurrently to detect defects and visualize the failure process. AE tools are designed for monitoring acoustic emissions produced during crack initiation (Godin et al. 2018). A major issue in the use of AE technique is that the signal discrimination is difficult. Ultrasonic wave imaging captures the multiple reflections of ultrasonic waves due to the presence of discontinuity. Ultrasonic imaging detects discontinuities at laboratory scale (Lee et al. 2009). CT scanning is a non-destructive imaging technique that utilizes X-ray technology and mathematical reconstruction algorithms to view cross-sectional slices of a material. Although CT-scanners are medical diagnostic tools they have been used extensively by the petroleum industry for studying reservoir cores for more than 20 years (Siddiqui and Khamees 2004). However, compared to other methods, CT scanning is relatively time consuming and expensive.

1.1 Motivation

Recent advances in machine learning methods have allowed us to process large, high-dimensional datasets for purposes of enhanced detection of anomalies and processes with high granularity at multiple scales. As a replacement for traditional experimental data analysis, in this paper, we develop a regressor-based machine-learning workflow to precisely locate mechanical discontinuity by processing multipoint compressional and shear wave transmission measurements. In an earlier study, we developed a classifier-based machine-learning workflow to categorically characterize certain bulk properties of the embedded crack clusters, such as orientation, dispersion, and spatial distribution, by processing multipoint compressional and shear wave traveltime measurements (Misra and Li 2019; Liu and Misra 2022). In the previous study, we did not account for the mode conversion, reflection, attenuation, and dispersion of the wave propagation. Unlike the previous study that processed only the traveltime measurements, we process the full waveforms in this study.

Elastic wave propagation is simulated using k-Wave, an implementation of the k-space pseudospectral method, which can handle reflection, scattering, and mode conversion. k-Wave implementation honors the fact that elastic waves in materials are subject to attenuation and dispersion in a broad range of frequencies. The configuration of source/transmitter and sensors/receivers used in this study is inspired by real-world laboratory experiments (Bhoumick et al. 2018; Chakravarty et al. 2020).

Moreover, this study suggests that the combination of regression model, synthetic data generation using k-Wave simulator, and single-source based multipoint wave-transmission measurements is an effective tool for visualizing/locating the discontinuity. The proposed method needs to be further developed to handle dynamic spatiotemporal evolution of the discontinuity and the structural complexity of the discontinuity.

2 Introduction to the workflow

First, the wave propagation phenomenon is simulated in 2D material containing various types of single, linear mechanical discontinuity. The embedded discontinuity varies in terms of the size, location, and orientation. Full waveform recordings are captured by multiple sensors located on the surface/boundary of the 2D material. Regressors are then developed to predict the location, orientation, and size of the discontinuity by training and testing the regressors to relate the full waveform recordings to the continuous states of the single, linear discontinuity in the 2D material. Two main modules in this workflow include the wave-propagation simulation model followed by the data preprocessing and regression analysis of the full waveform recordings at multiple locations (Fig. 1).

Physics-based open-source k-Wave toolbox is used to simulate the elastic wave propagation, originating from a single pressure source, through a 2D material containing single discontinuity. The full waveforms are recorded by 20 receivers placed on the surface of the material. The full waveform at each receiver is recorded for 30 microseconds, discretized into 1500 time steps. Overall, 20 waveforms lasting for 30 microseconds each are generated for 13,000 materials containing single discontinuity of varying length, orientation, and location. k-Wave simulator was used to generate 10,000 training samples and 3000 test samples that took 55 h on a traditional computing desktop. The simulation model used for generating the training/testing data will be explained in Sect. 3. The training dataset is used to construct the regressors, whereas the testing dataset is used to evaluate the built model. The training dataset and testing dataset should not have common samples.

It is necessary to perform feature elimination/selection as a preprocessing step to overcome the curse of dimensionality because not all the 1500 time steps recorded by each of the 20 receivers are relevant and useful for the desired data-driven prediction. The feature selection methods are discussed in Sect. 4. Several relevant, independent, and informative features are extracted from the discretized full waveforms to serve as the inputs for the regression models. The regressors learn to relate these waveform-derived features with the location, size and orientation of the discontinuity. The regressors detect and localize the primary discontinuity by learning from the simulated waveforms and corresponding state of discontinuity. The performance of regression model is evaluated using coefficient of determination, R square score, which provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model (Castagna et al. 1985). The results of this evaluation showed that the fracture could be identified by the regression models with above 0.9 accuracy. This will be discussed in Sect. 5. Good performance of the regressors strongly depend on well-formulated feature extraction and data pre-processing. Care must be taken to ensure that there is no leakage of information between the testing and training stages during the data preprocessing steps.

3 Simulation model

3.1 k-Wave simulation

k-Wave is an open-source MATLAB toolbox designed for time-domain acoustic and ultrasound simulations in complex medium (Treeby and Cox 2010a, b). The toolbox can handle elastic wave propagation based on two coupled first-order equations describing the stress and particle velocity within the medium. The elastic simulation functions (pstdElastic 2D and pstdElastic3D) are invoked to perform the desired simulation. The four input structures, including kgrid, medium, source location and sensor location, define the properties of the computational grid. In an isotropic elastic medium, the material properties can be characterized by the shear and compressional wave velocities, and the mass density. The medium parameters are defined at the level of computational grid. The distribution of medium

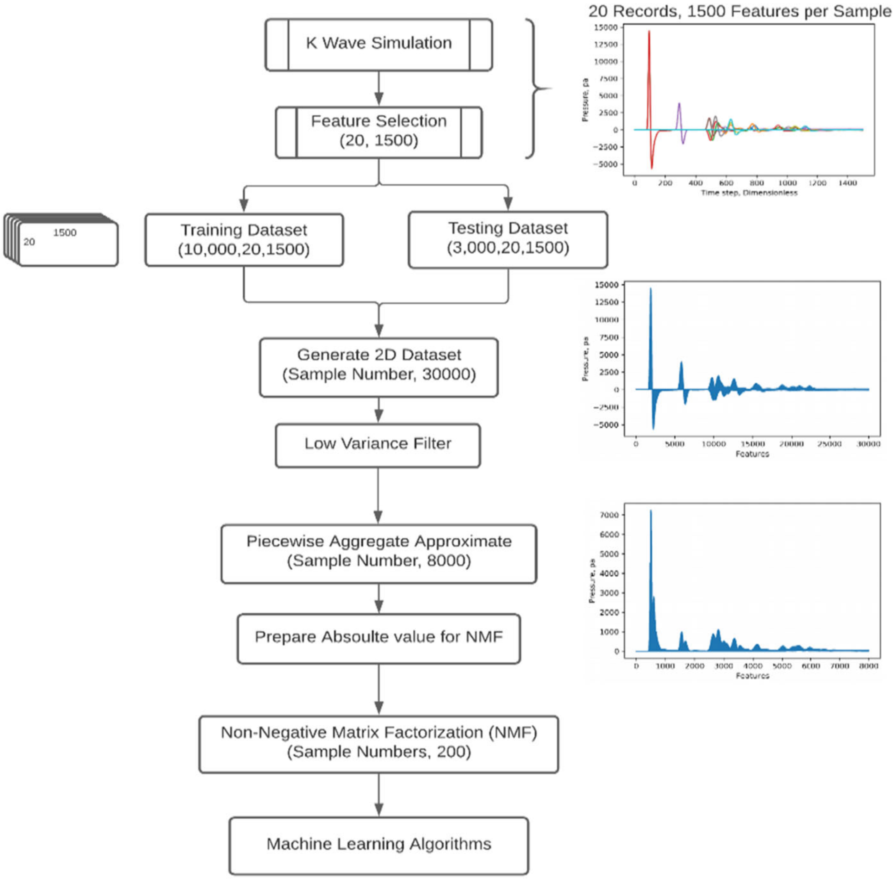
K Wave Simulation
Feature Selection (20, 1500)
20 Records, 1500 Features per Sample
Pressure, pa
Time step, Dimensionless
1500
20
Training Dataset (10,000,20,1500)
Testing Dataset (3,000,20,1500)
Generate 2D Dataset (Sample Number, 30000)
Features
Low Variance Filter
Piecewise Aggregate Approximate (Sample Number, 8000)
Prepare Absoulte value for NMF
Non-Negative Matrix Factorization (NMF) (Sample Numbers, 200)
Machine Learning Algorithms

Fig. 1 Data-driven workflow for learning to detect and locate mechanical discontinuities in materials by processing the full waveform measurements at multiple locations. This workflow clarifies the dimensions of the data set at each significant feature reduction step

properties, stress, and source determine the propagation of the wave field that can be captured at multiple sensor locations. The time array is defined by the user which must be evenly spaced and monotonic increasing. The elastic modeling supports three types of sources: an initial pressure distribution, time varying velocity or time varying stress sources (Treeby et al. 2014). In this work, our source is an initial pressure point located in the middle of the material's left boundary. The source and sensor locations are defined as a series of Cartesian coordinates within the computational grids. If the Cartesian coordinates do not exactly match the coordinates of a grid point, the output values are calculated from the interpolation.

Reflection and scattering are an important parameter of medium when simulating physical phenomena like wave propagation. k-Wave treats the medium as absorbing material for both compressional and shear waves. A split-field perfectly matched layer (PML) is used to absorb the waves at the edges of the computational domain. Without this boundary layer, the computation of the spatial derivates via the fast Fourier transform (FFT) causes waves leaving one side of the domain to reappear at the opposite side (Treeby and Cox 2010a). The use of the PML thus facilitates infinite domain simulations without increase the size of the computational grid. However, the computational time will still be dependent on the total size of the grid including the PML. For accurate simulation, it is crucial that all the source and sensor do not lie with in this layer. By changing the absorption and thickness of the PML, we could control the reflections and wave wrapping from the boundaries. The absorption within the layer is set by 'PMLAlpha' in units of Nepers per grid point, which is 2, by default.

k-Wave simulation has been experimentally validated in many studies (Martin et al. 2019; Treeby et al. 2012). The maximum supported frequency in k-Wave vary based on the spatial grid size. If the grid spacing is not uniform in each Cartesian direction, the maximum frequency supported in all directions will be dictated by the largest grid spacing. Therefore, when setting up a simulation it is necessary to ensure that the grid spacing is sufficiently small that the highest frequency of interest can be supported by the grid. Moreover, stability of the simulation is highly depending on the spatial size of grid and timestep.

3.2 Experimental design

The 2D numerical models of crack-bearing material implemented in this study are inspired by the laboratory experiments conducted at the Integrated Core Characterization Center (Bhoumick et al. 2018; Misra et al. 2019; Misra and Li 2019). In those studies, they placed multiple sonic wave sources and receivers around a porous cylindrical rock samples to visualize crack distribution inside the rock samples. Our 2D simulation model in k-Wave is inspired by real experimental equipment. k-Wave simulator was used to simulate wave propagation through 13,000 materials containing single discontinuity of varying length, orientation, and location.

3.2.1 Transmitter-receiver (source-sensor) configuration

In our experiment, we built a square-shaped crack-bearing material with dimension of 64 mm by 64 mm discretized using 320 by 320 grids. 10 additional grids representing the PML is added to all the four boundaries. In total, the numerical model is discretized using 340 by 340 grids. This partially effective PML absorbs some of the waves approaching the boundaries and majority of wave will be reflected back into the material, which honors the real-world behavior of a material sample. One source and 20 receivers are located around this material to record the whole waveform for 30 microseconds. Six sensors are placed on each boundary are shown as black circles in Fig. 2. The red triangle denotes the pressure source placed at the center of the left boundary of the material.

The material is assumed to be sandstone with 20% porosity. Porosity describes the volume of pore space within the bulk volume of the material. The compressional and shear wave velocities of the crack-bearing material are set on the basis of commonly occurring porous sandstones in the subsurface earth. Elastic waves can be divided into body waves and surface waves according to the way they propagate through a material. Compressional (P-wave) and shear (S-wave) waves as body waves are most often used for inspecting defects (He et al. 2019). Without considering spatial variations of water saturation and pressure, the compressional wave velocity of the crack-bearing material is set to 3760 m/s, whereas the shear wave velocity is set to 2300 m/s representing a 20% water-filled porous sandstone (Hamada and Joseph 2020).

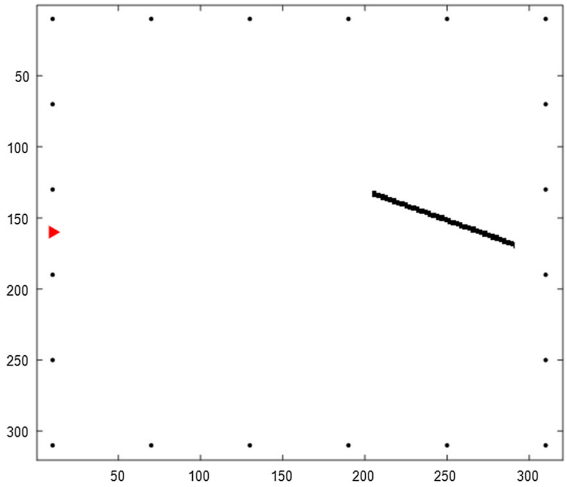

Fig. 2 Transmitter-receiver (source-sensor) configuration for a crack-bearing material containing single, linear discontinuity. One source/transmitter (red triangle) and 20 receivers/sensors (black circle) are placed on the boundary of the material

3.2.2 Mechanical discontinuity

Discontinuities include all types of mechanical break or plane of weakness in a material. For example, discontinuities can occur as joints, bedding plane, fractures and shear zones that weaken the strength of rock masses (Osogba et al. 2020). In our study, we will refer discontinuities as cracks. The length of crack is randomly selected from a uniform distribution ranging from 10 to 30 mm (50 to 150 grids). The crack is assumed to be filled with water and has a width around 0.6 mm.

3.2.3 Compressional and shear wave measurements

Both the compressional wave and shear wave can be used for characterizing the crack-bearing material (Klimentos and McCann 1990). The k-Wave simulations include both compressional- and shear-wave propagation. k-Wave can model the conversion of P-wave energy into S-wave and vice versa at a reflecting horizon. In addition, the normal and shear components of the stress field are also recorded. Figure 3 shows the waves recorded by the sensors placed on the left boundary of the material (shown in Fig. 2). The amplitude of the signal represents the pressure intensity in Pa. There exists approximate symmetry in the waves recorded above and below the single source due to their symmetric placement. The slight differences in the wave patterns between subplots 1 and 2 are influenced by the heterogeneity of the crack inside material and the presence of mechanical discontinuity. We hypothesize that the use of robust signal processing followed by machine learning can identify these differences in the waves recorded at multiple locations and then use those differences to interpret the location, orientation and length of the discontinuity. Notably, the wave measurements last for 30 microseconds which is long enough to capture the reflections from the boundaries, even those from the farthest border. An important requirement of the signal processing is to distinguish the signatures of reflection from the boundaries and those from the mechanical discontinuity.

3.3 Description of the dataset

In Machine Learning projects, we need a large dataset which is a collection of features and targets. k-Wave simulation is conducted on a numerical model of a crack-bearing material for simulating the elastic wave propagation through the material and for recording the full waveform at 20 sensors placed on the boundary of the material. Each simulation for a numerical model of crack-bearing material constitutes a sample. The features of the sample include the 20 full waveforms recorded at 20 sensors for 30 microseconds, comprising 1500 time steps. The waveform dataset recorded for each simulation has 20 rows representing the 20

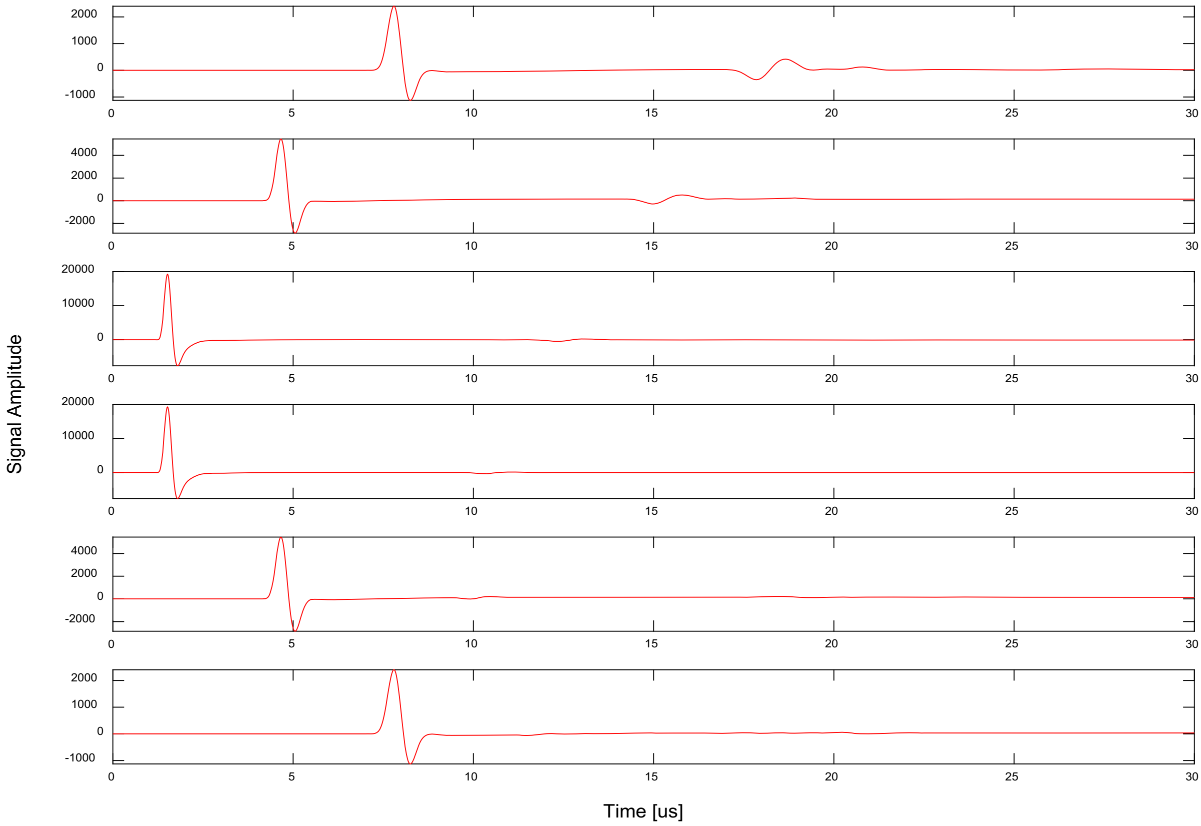

Fig. 3 Waveforms recorded at six sensors located on the left boundary (source side) of the material, as shown in Fig. 2

sensors and 1500 columns representing the equispaced time steps that discretize the 30 microseconds. The targets for each sample are the length, orientation, and location of the single, linear discontinuity in the material. As a rule of thumb, for training, a larger size of the dataset ensures higher statistical significance of the data-driven model. This study generates10,000 samples as training data and 3000 samples as testing data. Each sample represents a crack-bearing material with random length, orientation, and location of mechanical discontinuity. The regression model learns from training samples to relate the features (extracted from the 20 full waveforms) with targets (length, orientation, and location of single discontinuity). The regression models are evaluated on the testing dataset.

4 Feature reduction

Feature reduction, also called dimensionality reduction, is commonly applied as a data preprocessing step to overcome the curse of dimensionality. A data-driven model has low computation cost, low memory usage and low risk of overfitting when trained on low-dimensional data, i.e. when samples are described using less number of features. The full waveform recorded at 20 locations for 1500 steps is a very high dimensional data containing 30,000 features per sample. Such a dataset needs robust dimensionality reduction methods, such as Discrete Fourier Transform (DFT) (Allen 1977; Weinstein and Ebert 1971), Singular Value Decomposition (SVD) (De Lathauwer et al. 2000; Klema and Laub 1980), and Discrete Wavelet Transform (DWT) (Shensa 1992; Wu and Misra 2019) and Short-Time Fourier Transform (Chakravarty et al. 2020). There are methods for piecewise aggregate representations of complex waveforms, including Piecewise Aggregate Approximation (PAA) and Symbolic Aggregate approximation (SAX).

Feature reduction techniques can also be divided into feature selection and feature extraction. Feature extraction could be distinguished from feature selection. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features (Misra and Wu 2019). Feature selection is for filtering irrelevant or redundant features from the whole dataset that do not contribute to the prediction variable. Feature selection algorithms have three main categories: wrappers, filters, and embedded methods. This work tests the methods, such as low variance filter, mutual information, Pearson correlation, and F-regression to remove redundant information. In terms of comparing the effectiveness of each feature extraction technique, we have tried many commonly used approaches including, Principal Component Analysis (PCA) (Abdi and Williams 2010), Linear Discriminant Analysis (LDA) (Ye et al. 2004), Nonnegative Matrix Factorization (NMF) (Lee and Seung 2000), Sparse Random Projection (SRP) (Bingham and Mannila 2001), and Gaussian Random Projection (GRP) (Bingham and Mannila 2001). Finally, we get the optimum combination of dimensionality reduction algorithms for purposes of our study. The dimensionality reduction applied in our work starts with a low variance threshold to drop low variance features (i.e. the timesteps that do not have sufficient information/variance). Then, we apply PAA to reduce the number of noisy time points with a specific window size. In the end, NMF extracts sparse and meaningful features from a set of nonnegative data vectors. The following sections will describe these two algorithms in more detail.

4.1 Piecewise aggregate approximate (PAA)

Yi and Faloutsos (2000) and Keogh et al. (2001) independently proposed PAA, a popular and competitive basic dimensionality reduction method for high-dimensional time-series data. It transforms a time series $X = (x_1, \ldots\ldots x_n)$ into another time series $\tilde{X} = (\tilde{x}_1, \ldots\ldots \tilde{x}_m)$ with $m \leq n$, where each of $\tilde{x}_i$ is calculated as follows:

$$\tilde{x}_i = \frac{m}{n} \sum_{j=\frac{n}{m}(i-1)+1}^{\left(\frac{n}{m}\right)i} x_j \qquad (1)$$

The basic idea behind this algorithm is to reduce the dimensionality of the time series data by splitting them into equal-sized segments that are computed by averaging the values in these segments (Keogh et al. 2001). The window size of our study is 8000, which is still a high dimension for regression models.

4.2 Non-negative matrix factorization (NMF)

NMF was first introduced by Paatero and Tapper (Paatero and Tapper 1994) in 1994 and popularized in an article by Lee and Seung (Lee and Seung 1999) in 1999. NMF approximates a given matrix X with a low-rank matrix approximation such that $X \approx WH$. The three matrices are assumed to have no negative elements to speed up the factorization. For certain dataset, non-negativity is an inherent property. This results in a compressed version of the original matrix. A well-known cost function of measuring the quality of approximation WH is the Frobenius norm (Lee and Seung 2000):

$$\|X - WH\|_F^2 = \sum_{i,j} (X - WH)_{ij}^2 \tag{2}$$

This objective function is minimized by an alternating minimization of W and H with a square error expression. In the standard NMF algorithm W and H are initialized with random nonnegative values before the iteration (Berry et al. 2007).

4.3 Description of the dimensionally reduced dataset

Figure 4 compares the original dataset (a) and reduced dataset (b), used in this study. Subplot (a) aggregates the 20 waveforms recorded by the 20 sensors over 1500 time steps, in total 30,000 features. The number of retrieved features was decreased to 200 after using the recommended dimensionality reduction procedure, which included the low variance filter, PAA, and NMF. However, if input features are significantly correlated with each other, multicollinearity occurs. This can result in distorted or misleading regressor results. Therefore, we have always checked the correlation between different variables in our dataset with Pearson correlation filter. Any dimensionality reduction performed on training data must also be performed on new data and on the test dataset. First, dimensionality reduction approach is learnt from the training data. Following that, the learnt dimensionality reduction approach is applied on the training data, test data and new data. Machine Learning models perform better when the distributions of the features are approximately normally distributed and when the scales/ranges of the features are relatively similar. Therefore, feature transformation to Gaussian-like distribution and standard scaler are the final data preprocessing steps before the transformed features are fed into the regression models.

5 Regression models to detect and locate the mechanical discontinuity

5.1 Regression models used in this study

This section addresses methods for the detection and characterization of discontinuity by using regression

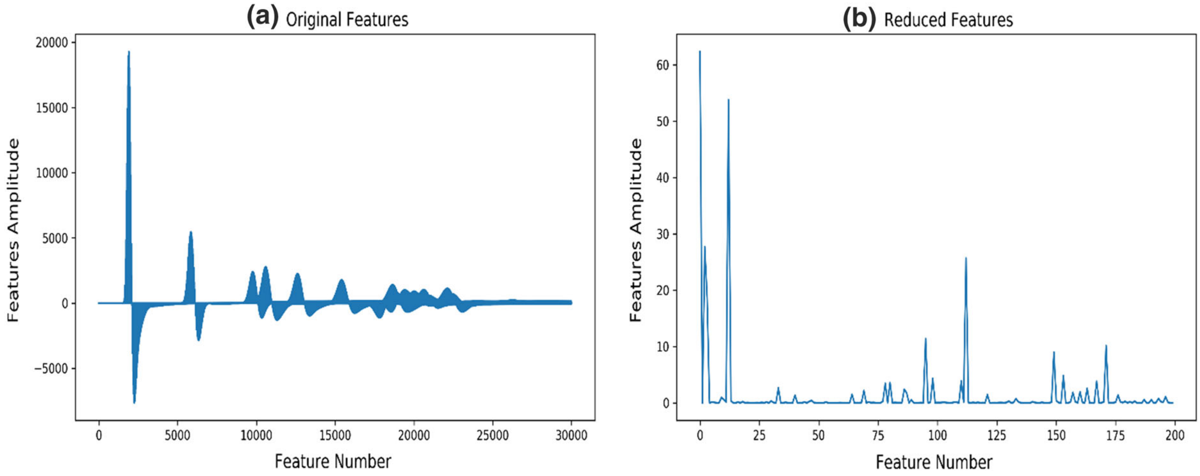

Fig. 4 **a** Original 30,000 features corresponding to a crack-bearing material sample; **b** 200 features extracted from the original 30,000-dimensional data corresponding to the crack-bearing material sample

models, such as Random Forest (RF), K-Nearest Neighbors (KNN), Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Support Vector Machine (SVM) and Bayesian Ridge Regression. RF is an ensemble learning method for classification and regression. It is a bagging technique where several decision trees are trained and deployed in parallel. RF model aggregates the predictions of all the trees to generate a final prediction based on a specific voting strategy. RF is known for its efficiency on large-sized databases. AdaBoost is another ensemble method which takes weaker learners and combines them in series to get a strong learner. The goal is grouping the weak learners to create a stronger, more generalizable and accurate model. Each subsequent weak learner improves on the predictions of previous weak learner. GB is also an ensemble method where multiple weak learners are combined in series for the final prediction. The major difference between AdaBoost and Gradient Boosting is how the two algorithms fix the shortcomings of pervious weak learners. AdaBoost improves the shortcomings by assigning higher weights to samples that were wrongly predicted by previous weak learner that guides the learning to best final prediction. Gradient boosting fixes the shortcomings of the previous weak learners by using gradient descent optimization of a loss function to reduce the errors of the previous weak learners.

KNN algorithm is a non-parametric machine learning method first developed by Evelyn Fix (1951) and Joseph Hodges that was later expanded (Altman 1992). KNN regression approximates the association between features and targets by averaging the target values of training samples within a specific neighborhood of a new sample for which the target needs to be predicted. The performance of KNN is based on the quality of the training dataset. SVM regression relies on kernel function to find the trend in data by using high-order transformations of the features. SVMs can efficiently perform a non-linear task by using a kernel transformation, which implicitly maps the samples from the original feature space to high-dimensional feature space, in which the data-driven modeling becomes tractable. The main idea is to minimize error in prediction based on an error-tolerance through an acceptable error margin. Another method, Bayesian Ridge regression estimates a probabilistic model by assuming the target is generated from a normal (Gaussian) distribution characterized by a mean and variance. Unlike Ordinary Linear Regression, Bayesian approach does not estimate a single optimal value of the model parameter but determines the posterior distribution for the model parameters. Ordinary linear regression assumes that there are enough measurements to find a meaningful model. Bayesian models are more flexible with better performance on smaller sized dataset. In the Bayesian approach, the data are supplemented with additional information in the form of a prior probability distribution.

This study compares the performances of six regression models capable of detecting and locating a single mechanical discontinuity embedded in a material. The prediction of each regressor can be evaluated by using the coefficient of determination, R2, which is a statistical measure of how well the regression predictions approximate the training/testing samples. This metric ranges typically from 0 to 1, such that 1 represents perfect model prediction and lower values denote poor predictions.

5.2 Performances of the regression models

In this section, we will discuss the performances of the six regression models for the task of detecting and locating the discontinuities. The performances are shown in Fig. 5. The models are trained on 10,000 samples and tested on 3000 samples. Grid search was used for hyperparameter optimization of each regressor. AdaBoost regressor with KNN as a base estimator performs the best, reaching an R2 of around 0.96 on the testing dataset. This is an exceptional

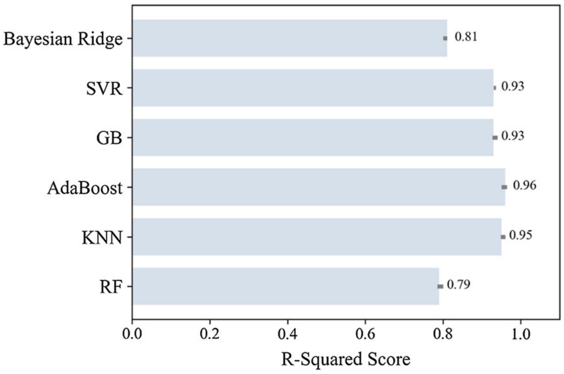

Fig. 5 Testing accuracies of the 6 regression methods with 95% confidence interval for the task of characterizing the orientation, location, and length of the mechanical discontinuity embedded in the material

generalization performance in detecting and locating embedded mechanical discontinuity. KNN, GB, and SVR also performed well with an accuracy in the range of 0.93 to 0.95. The average accuracy of these models is around 0.9. Figure 5 shows the 95% confidence interval for regressor performance as a grey error bar. This confidence interval is very small, the detail standard deviation of R-Squared is shown in the Table 1. The generalization performances of KNN regressor and SVR are visualized in "Appendix A". Figures 8 and 9 illustrates the exceptional performance of the regressors in detecting and locating the mechanical discontinuity in 2D materials.

To support the exceptional performance of most of the regressors, the generalization performance of KNN is presented in Fig. 6 (right), where the location, orientation, and size of a predicted discontinuity is compared against the known/true discontinuity. The shapes of the predicted and known discontinuity coincide; there is a slight difference is the lengths of the discontinuity. More such comparisons are presented in "Appendix A". The plot on the left contains the locations of the center of discontinuity for the 3000 testing samples and their predictions. The red dots are predicted x-coordinates of the center of discontinuity, and the blue dots are the y-coordinates of the center of discontinuity in the testing dataset. Overall, the predicted locations agree with the known locations because the dots lie on the X = Y line, indicating the good match between true and predicted responses. It is worth noting that predictions of x-coordinate are slightly less accurate than y-coordinate predictions. It seems possible that these results are due to the signal source is on the x-axis, resulting in more x-axis reflections in the signal as noise. It can be concluded that the KNN regressor is a reliable method for locating discontinuity.

6 Sensitivity of the regressors to noise in data

6.1 Noise generation

Noise is a common occurrence in all forms of measurements, especially when measuring wave propagation interactions with discontinuities because of energy loss and complex scattering and reflection processes. To investigate the robustness of the classifiers to the noise, numerical experiments are carried out to analyze the responses of the regressors to noisy data. The generation of noise can be characterized in different ways such as distribution or color of noise (Han and Misra 2018). Different colors of noise have significantly different properties. White noise is commonly used for impulse response which has equal intensity at different frequencies, giving it a constant power spectral density (Stein 2012).

In this work, experiments are conducted to analyze the regression model response with a noisy dataset in both training and testing data. The noise in the first experiment is generated using a Gaussian Distribution $\mathcal{N}(0,\theta)$. The level of noise is adjusted by changing the variance $\theta(0,10,50,100)$of the Gaussian distribution. Then, second experiment tests the regressor performance with noise characterized using Uniform Distribution $\mathrm{U}(0,\theta)$. The magnitude of noise in data varies with θ. When θ is equal to zero, the dataset is free of noise.

6.2 Impact of noise on algorithms

As evident in Fig. 7, the generalization performances of RF, KNN, AdaBoost, GB, SVR and Bayesian Ridge drop in the presence of noise and the drop in accuracy increases with increase in the noise level. Gaussian distribution of noise reduces the accuracy more than the uniform distribution of noise. KNN is the least sensitive to uniformly distributed noise, with its R-squared dropping

Table 1 Comparison of the generalization performances of 4 of the 6 regressors when all sensors on specific boundary are removed from the dataset. The number inside the parentheses are standard deviation of the regression accuracy

	KNN	AdaBoost	Gradient Boosting	SVR
All sensors	0.95 (0.003)	0.96 (0.007)	0.93 (0.003)	0.93 (0.002)
Remove left	0.96 (0.003)	0.96 (0.007)	0.93 (0.003)	0.94 (0.002)
Remove right	0.90 (0.002)	0.91 (0.005)	0.90 (0.004)	0.86 (0.003)
Remove upper	0.90 (0.002)	0.91 (0.005)	0.90 (0.004)	0.89 (0.002)
Remove lower	0.93 (0.003)	0.92 (0.005)	0.92 (0.003)	0.90 (0.002)

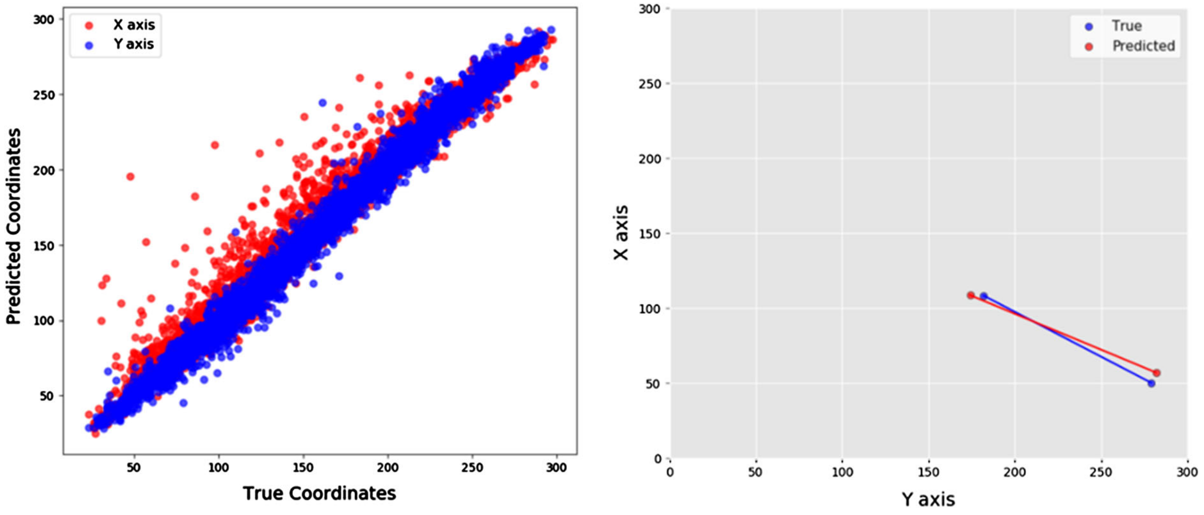

Fig. 6 Evaluation of the generalization performance of KNN regression on testing dataset **a** analysis of the predictions of centers of discontinuity and **b** predictions of location, orientation, and size of the discontinuity for a randomly selected sample from testing dataset. More such comparisons are available in "Appendix A"

by only 0.05. Both KNN and AdaBoost exhibit lower sensitivity to Gaussian distributed noise as compared to other regression techniques. As the noise level increases, the accuracy of KNN decreases from 0.95 to 0.75, while AdaBoost decreases from 0.96 to 0.77. The accuracies of KNN, AdaBoost, GB, and SVR is above 0.9 when the noise level has variance less than 10.

7 Analysis of sensor importance

Not all the sensors are significant for the characterization of discontinuity. For the desired characterization task, the importance of the 20 sensors placed on the 4 boundaries of the material (Fig. 2) can be quantified by performing sensitivity test. There are four sides/boundaries of the materials on which sensors are placed. The information from all sensors on each boundary is destroyed one boundary at a time by removing the information of all the sensors on that boundary. Following that, regressors were trained on the data without information from sensors on specific boundary. The newly trained models exhibited drop in the generalization performance quantified in terms of R2 (Table 1). The source and six sensors are located on the left boundary. Six sensors are located on the upper and lower boundaries. Six sensors are located on

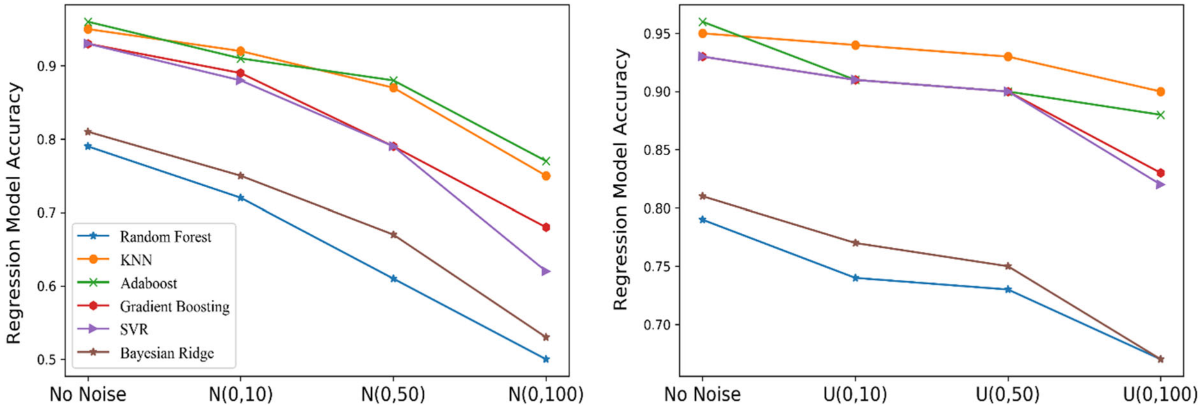

Fig. 7 Generalization performances of the 6 regressors when trained on dataset with different levels of noise exhibiting **a** Gaussian distribution and **b** uniform distribution

the right boundary opposite to the left boundary containing the source. The sensors located on the right boundary, opposite to the source, and the upper boundary are much more important than the remaining sensors on the left and lower boundaries. Without influencing the overall accuracy, sensors on the left boundary could be eliminated. This result can be explained by the fact that the signals received by the sensors on the left boundary do not actually pass through the crack. Similarly, because the signal travels through the entire material, the sensor on the right boundary contains more information. The sensor importance analysis strongly indicates that there is no need for sensors on the boundary containing the single source and all the sensors from one of the three other boundaries can be removed without drastically affecting the overall performances of the KNN, AdaBoost and Gradient Boosting regressors, whose generalization performances are above 0.9 despite the information loss.

8 Conclusions

The study is based on the hypothesis that the use of robust signal processing followed by machine learning can identify small differences and minute patterns in the waveforms recorded at multiple locations and then use those differences/patterns to predict the location, orientation and length of a single discontinuity embedded in a material. k-Wave simulation is used to model the elastic wave propagation in a 2D numerical model of material containing single, linear discontinuity. The modeling framework accounts for wave attenuation, reflection, mode conversion, and scattering including the effects of boundary. AdaBoost regressor with k-Nearest Neighbor as the base estimator significantly outperforms all other regression models and achieves an exception generalization performance of 0.96, in terms of the coefficient of determination. Overall, k-Nearest Neighbor (KNN) and Support Vector Regressor exhibit similar performances as AdaBoost. To achieve this exceptional performance, it is important that the high dimensionality of the multipoint time-series data be drastically reduced by two orders of magnitude using low variance filter, piecewise aggregate approximation, and non-negative matrix factorization in sequence. Moreover, feature transformation to Gaussian-like distribution followed by feature scaling positively contributes to the generalization performance of the regressors. Random Forest and Bayesian Ridge regression methods exhibit low performances. Adaboost and KNN regressors are relatively robust to uniformly distributed noise. However, increase in the variance of Gaussian noise adversely affects the performances of the regressors. Sensor importance study indicates that sensors are required on only two of the three boundaries/sides of the material, excluding the boundary containing the single source.

Data-driven workflow successfully predicted the location, orientation, and size of a mechanical discontinuity in a material by processing the full waveforms recorded by 20 sensors that originated from a single pressure-impulse source. However, the generalizability of these results is subject to certain limitations. For example, designed material is relatively small due to the restrictions of the simulation toolbox. Moreover, the crack generated in elastic wave simulator is linear and random. Notwithstanding these limitations, this study suggests that machine learning could be useful for crack characterization and detection. Further research is required to develop a new framework that incorporates machine learning and physics-based simulation models with the propagation of mechanical discontinuities.

Acknowledgements This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division, under Award Number DE-SC0020675. Dr. Hao Li provided some guidance to co-author Rui Liu during the initial phase of this work.

Data availability No data availability.

Declarations

Conflict of interest The authors declare that they have no conflict of interest

Appendix A: Visualization of the generalization performance

See Figs. 8 and 9.

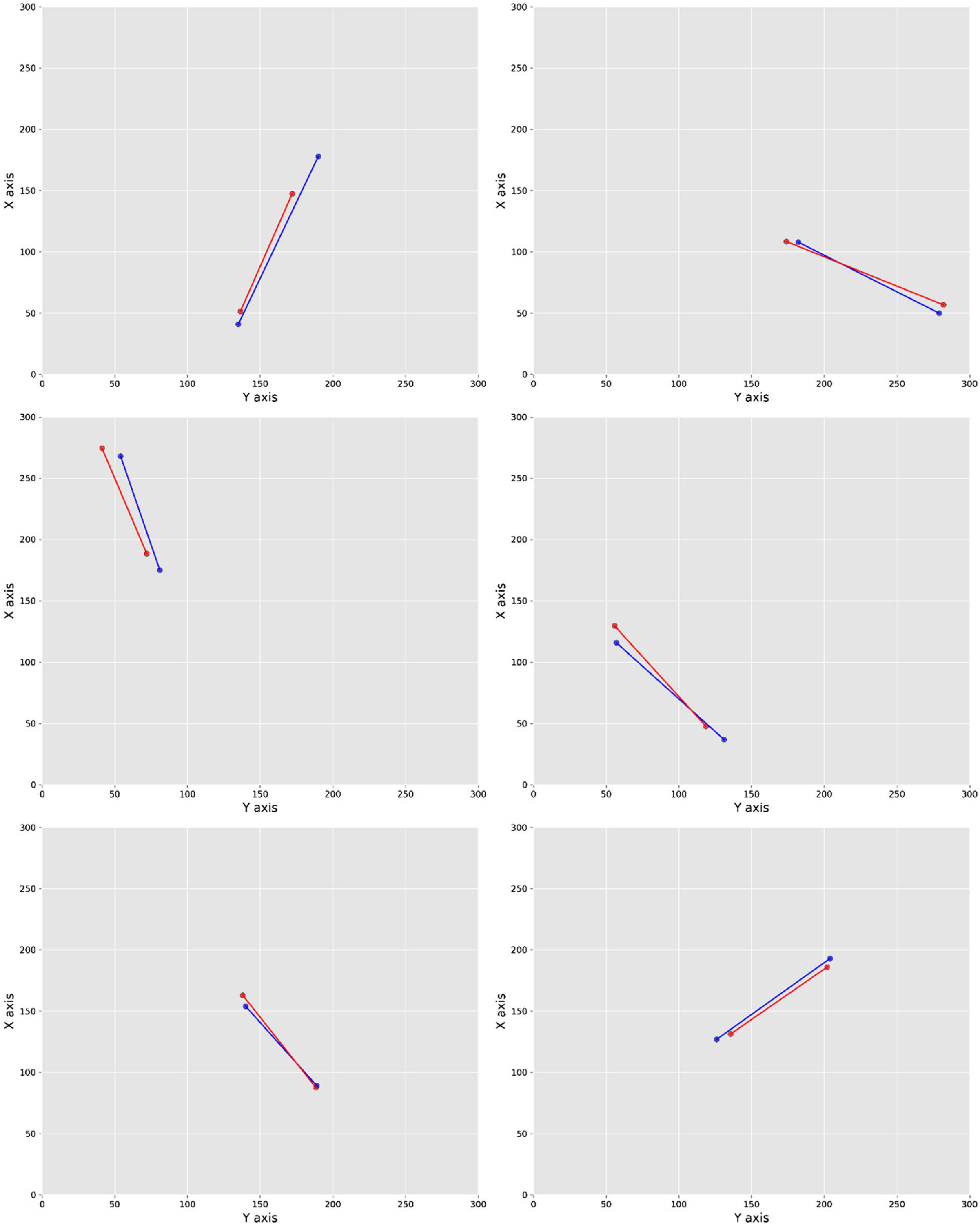

Fig. 8 Visualization of the predictions of location, orientation, and size of the discontinuity for six randomly selected material samples from the testing dataset. Predictions were obtained using the k-nearest neighbor regressor. Known discontinuity is shown in red and the predicted discontinuity is shown in blue

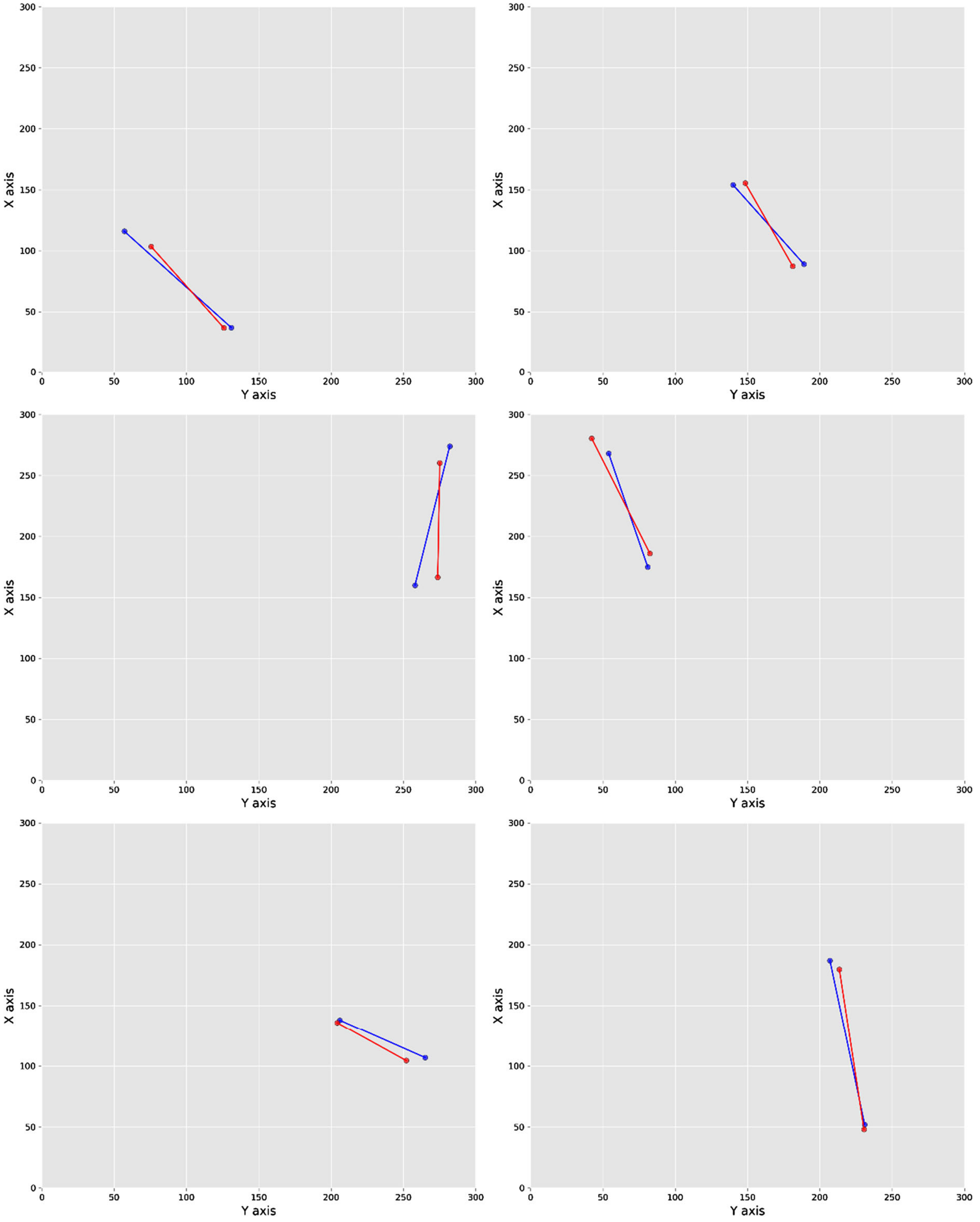

Fig. 9 Visualization of the predictions of location, orientation, and size of the discontinuity for six randomly selected material samples from the testing dataset. Predictions were obtained using the support vector regressor. Known discontinuity is shown in red and the predicted discontinuity is shown in blue

References

Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Rev: Comput Stat 2(4):433–459

Allen J (1977) Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans Acoust Speech Signal Process 25(3):235–238

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185

Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate non-negative matrix factorization. Comput Stat Data Anal 52(1):155–173

Bhoumick P, Sondergeld C, Rai C (2018) Mapping hydraulic fracture in pyrophyllite using shear wave. In: 52nd US rock mechanics/geomechanics symposium

Bingham E, Mannila H (2001) Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining

Castagna JP, Batzle ML, Eastwood RL (1985) Relationships between compressional-wave and shear-wave velocities in clastic silicate rocks. Geophysics 50(4):571–581

Chakravarty A, Misra S, Rai CS (2020) Hydraulic fracture visualization by processing ultrasonic transmission waveforms using unsupervised learning A

Chakravarty A, Misra S, Rai CS (2021) Visualization of hydraulic fracture using physics-informed clustering to process ultrasonic shear waves. Int J Rock Mech Min Sci 137:104568

De Lathauwer L, De Moor B, Vandewalle J (2000) A multilinear singular value decomposition. SIAM J Matrix Anal Appl 21(4):1253–1278

Fix E (1951) Discriminatory analysis: nonparametric discrimination, consistency properties. USAF School of Aviation Medicine, Dayton

Godin N, Reynaud P, Fantozzi G (2018) Challenges and limitations in the identification of acoustic emission signature of damage mechanisms in composites materials. Appl Sci 8(8):1267

Hamada G, Joseph V (2020) Developed correlations between sound wave velocity and porosity, permeability and mechanical properties of sandstone core samples. Pet Res 5(4):326–338

Han Y, Misra S (2018) Joint petrophysical inversion of multifrequency conductivity and permittivity logs derived from subsurface galvanic, induction, propagation, and dielectric dispersion measurements. Geophysics 83(3):D97–D112

He J, Li H, Misra S (2019) Data-driven in-situ sonic-log synthesis in shale reservoirs for geomechanical characterization. SPE Reserv Eval Eng 22(04):1–225

Kabir S, Rivard P, He D-C, Thivierge P (2009) Damage assessment for concrete structure using image processing techniques on acoustic borehole imagery. Constr Build Mater 23(10):3166–3174

Kanasewich ER, Phadke SM (1988) Imaging discontinuities on seismic sections. Geophysics 53(3):334–345

Keogh E, Chakrabarti K, Pazzani M, Mehrotra S (2001) Dimensionality reduction for fast similarity search in large time series databases. Knowl Inf Syst 3(3):263–286

Klema V, Laub A (1980) The singular value decomposition: its computation and some applications. IEEE Trans Autom Control 25(2):164–176

Klimentos T, McCann C (1990) Relationships among compressional wave attenuation, porosity, clay content, and permeability in sandstones. Geophysics 55(8):998–1014

Kosari E, Ghareh-Cheloo S, Kadkhodaie-Ilkhchi A, Bahroudi A (2015) Fracture characterization by fusion of geophysical and geomechanical data: a case study from the Asmari reservoir, the Central Zagros fold-thrust belt. J Geophys Eng 12(1):130–143

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791

Lee D, Seung HS (2000) Algorithms for non-negative matrix factorization. Advances in neural information processing systems 13

Lee I-M, Truong QH, Kim D-H, Lee J-S (2009) Discontinuity detection ahead of a tunnel face utilizing ultrasonic reflection: laboratory scale application. Tunn Undergr Space Technol 24(2):155–163

Liu R, Misra S (2022) A generalized machine learning workflow to visualize mechanical discontinuity. J Pet Sci Eng 210:109963

Martin E, Jaros J, Treeby BE (2019) Experimental validation of k-Wave: nonlinear wave propagation in layered, absorbing fluid media. IEEE Trans Ultrason Ferroelectr Freq Control 67(1):81–91

Misra S, Wu Y (2020) Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. Gulf Professional Publishing, Houston, pp 289–314

Misra S, Li H (2019) Noninvasive fracture characterization based on the classification of sonic wave travel times. Mach Learn Subsurf Charact. https://doi.org/10.1016/b978-0-12-817736-5.00009-0

Misra S, Chakravarty A, Bhoumick P, Rai CS (2019) Unsupervised clustering methods for noninvasive characterization of fracture-induced geomechanical alterations. Mach Learn Subsurf Charact, 39

Osogba O, Misra S, Xu C (2020) Machine learning workflow to predict multi-target subsurface signals for the exploration of hydrocarbon and water. Fuel 278:118357

Pyrak-Nolte LJ, DePaolo DJ, Pietraß T (2015) Controlling subsurface fractures and fluid flow: a basic research agenda. USDOE Office of Science (SC) (United States)

Paatero P, Tapper U (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5(2):111–126

Shalaby MR, Islam MA (2017) Fracture detection using conventional well logging in carbonate Matulla Formation, Geisum oil field, southern Gulf of Suez, Egypt. J Pet Explor Prod Technol 7(4):977–989

Shensa MJ (1992) The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Trans Signal Process 40(10):2464–2482

Siddiqui S, Khamees AA (2004) Dual-energy CT-scanning applications in rock characterization. In: SPE annual technical conference and exhibition

Stein ML (2012) Interpolation of spatial data: some theory for kriging. Springer, New York

Szwedzicki T, Shamu W (1999) The effect of discontinuities on strength of rock samples. In: Proceedings of the Australasian Institute of Mining and Metallurgy

Treeby BE, Cox BT (2010a) k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields. J Biomed Opt 15(2):021314

Treeby BE, Cox BT (2010b) Modeling power law absorption and dispersion for acoustic propagation using the fractional Laplacian. J Acoust Soc Am 127(5):2741–2748

Treeby BE, Jaros J, Rendell AP, Cox B (2012) Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using ak-space pseudospectral method. J Acoust Soc Am 131(6):4324–4336

Treeby BE, Jaros J, Rohrbach D, Cox B (2014) Modelling elastic wave propagation using the k-wave matlab toolbox. In: 2014 IEEE international ultrasonics symposium

Weinstein S, Ebert P (1971) Data transmission by frequency-division multiplexing using the discrete Fourier transform. IEEE Trans Commun Technol 19(5):628–634

Wu Y, Misra S (2019) Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix. IEEE Geosci Remote Sens Lett 17(7):1144–1147

Ye J, Janardan R, Li Q (2004) Two-dimensional linear discriminant analysis. Adv Neural Inf Process Syst 17:1569–1576

Yi B-K, Faloutsos C (2000) Fast time sequence indexing for arbitrary Lp norms

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.