Data-Driven Classification of Materials with Open or Closed Mechanical Discontinuities Based on Multipoint, Multimodal Travel-Time Measurements

Rui Liu, Siddharth Misra

Texas A&M University, College Station, Texas, USA

1. Abstract

Wave propagation and diffusive transport phenomena could work as evidence of the mechanical discontinuities in material. For the problem of poor efficiency of the existing fracture simulation methods, this paper proposes crack-bearing material characterization approach by processing wave travel-time using seven data-driven classification techniques. To that end, we perform classification models to predict discontinuities orientation, dispersion, and spatial distribution prediction by learning from the different-waves simulation model. The travel-time measured by multiple sensors placed around the material perform as our input data of machine learning method. As a result, this work found that machine learning models exhibit best classification performance on classifying crack dominant orientations. Combination of compressional wave and shear wave are enough to capture the crack information in the material, however, the pressure diffusion also able to optimize our algorithms. Voting classifier and gradient boosting classifier perform the best for purposes of characterization. When compare the performance of different mechanical discontinuities, embedded closed discontinuities shows high accuracy than open discontinuities on the classification models.

2. Introduction

2.1. Motivation

"Discontinuity" is a general term denoting any separation in a rock mass having zero or low tensile strength (Zhang et al., 1998). Discontinuities include many types of mechanical breaks such as fault, joints and fractures than weakened the strength of the rock mass (Osogba et al., 2020).

Mechanical breaks are highly related to the capacity of reservoir rocks to contain or storage the fluids (Misra et al., 2020). Fractures are an important storage space in oil and gas reservoirs. This study focusses on mechanical discontinuity. Nevertheless, it is a challenge task to characterize the discontinuities in the rock. Observation and analysis the core data is one way obtain the crack characterization. Meanwhile, it is also possible to predict the distribution of cracks according to rock rupture criterion by rupturing the core. However, these methods always influenced by the sample number and experimental environments. As the development of the machine learning algorithms, the motivation of our study is that if the classification models can characterize crack-bearing material with high accuracy based on sonic waves and pressure diffusion travel times without destroy core data. This work is focus on the discussion of discontinuities orientation, dispersion and spatial distribution in formation by processing the different-waves travel-time to data-driven methods. The main purpose of classification algorithms is to learn from the existed training travel time and evidence, to be able to make predictions for testing data. Then, the developed models can be applied to any new travel time dataset to predict the crack information.

3. Workflow

Figure 1 is the detail description of the workflow to build the classification models. A good starting point is to accurately identify the current fracture system created by discrete fracture network (DFN). All cases in this study are start from model simulation. The model has a dimension of 150mm by 150mm discretized using 500 by 500 grids. The wave source and 28 sensors are located around a 2D squared crack-bearing material designed with 100 discontinuities. Crack length is selected randomly from an exponential distribution in the range of 0.3mm - 3mm. The wave source is located in the middle of the left boundary. Sensors are equally settled at the other three boundaries. Each boundary has 10 sensors, 28 receivers in total. The travel time of the front wave at 28 sensors will be recorded as our dataset.

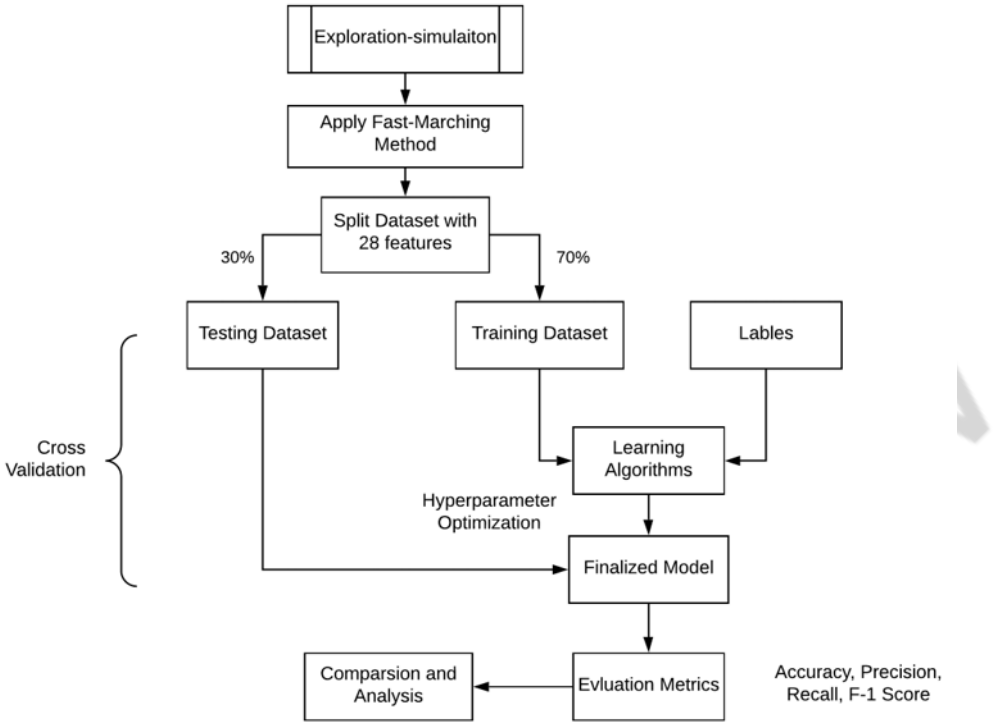

Figure 1. Model development and evaluation workflow

Then, the simulation travel time were divided into two parts: training and testing dataset. For example, our tests use 70% data as training data. Classification models will be trained on the training dataset. Cross validation approaches split the original training data into one or more training subsets to balance the response variances. The remaining 30% data is working as testing data to evaluate the classifiers accuracy.

The paper is structured as follows. We provide the physics-based simulation model we use as reference to be fed in our data-driven model. Then we followed the discussion of the classification models. The comparison presentation is followed by computational results supporting by our proposed model. We wrap up the paper with a few conclusions and further directions of work.

3.1.Properties of the material with embedded discontinuities

3.1.1. Properties of the Background material

In this work, we use sandstone with 20% porosity as the material background. Therefore, the velocity set of wave/diffusion is set on the basis of the behavior of the porous sandstone physical ground. The present work aims to learn more about the classification behavior when using P-wave,

S-wave, and pressure diffusion velocity in sandstone samples to characterize the fracture system. Without considering the water saturation and pressure effect in the real field, we assume the compressional wave velocity is 3760 m/s. The shear wave velocity values in a porous material will always be less than compressional wave. The shear wave in our cases are assumed to be 2300m/s. Then, our approach draws on the solution to the diffusive-pressure equation that mimics pressure front propagation phenomena. The travel time calculation will get from fast-marching method (FMM) described in section 3.3.

3.1.2. Variable Spatial properties of mechanical discontinuity

The spatial distribution classification of mechanical discontinuities in the formation are creating using intensity functions. These functions describe the crack occurrence probability in the investigated material. In the classification of spatial distribution, type 1 is the random distribution means cracks have an equal probability in the domain. Type 2 is the linear distribution followed by the linear probability function that related to y axis:

$$\boldsymbol{\lambda}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y} \tag{1}$$

Then, when gaussian function applied on the crack-bearing material as intensity function:

$$\boldsymbol{\lambda}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{c} * \boldsymbol{exp}^{\left(-\boldsymbol{d}*(\boldsymbol{x}-\boldsymbol{x_o})^2+(\boldsymbol{y}-\boldsymbol{y_o})^2\right)} \tag{2}$$

where $\boldsymbol{x}_o$ and $\boldsymbol{y}_\mathbf{o}$ is the center of the Gaussian distribution both set to 250, d controls the variance of the distribution is set to 0.00005, and c controls the minimum value of the intensity function equal to 1 in our case. When adding two Gaussian crack clusters in the material, the equation is similar to Eq.2.

3.1.3. Variable transport properties of mechanical discontinuity

Transport phenomenon, in physics, any of the phenomena involving the movement of various entities, such as mass, momentum, or energy, through a medium, fluid or solid, by virtue of nonuniform conditions existing within the medium. The transportation through pores media in our study is driven by the wave energy and pressure difference between matrix and the mechanical discontinuities. The travel time of the wave front are expected to capture the characteristic of the mechanical discontinuities.

3.1.4. Default transport properties

One aim of this study is to compare the classification model performance for open and embedded closed discontinuities. We assume the open cracks are filled with air; closed ones are cemented by 0 % porosity limestone. The following sections will discuss the wave velocity used for different discontinuities.

3.1.4.1. P-wavefront case

For embedded open discontinuities, the compressional wave is travelling through the air. The travel time calculation used 340 m/s as the fracture velocity. That shows a large contrast compared to the matrix velocity which is 3760 m/s. On the other hand, the wave velocity through the cemented discontinuities is the same as wave velocity traveled in the limestone without pores.

3.1.4.2. S-wavefront propagation case

The shear wave travels slower than P-wave because they do not change the volume of the material through when propagate. S-waves are waving that shear material. An important distinguishing characteristic of an S-wave is its inability to propagate through fluid or gas because they cannot transmit the shear stress. In our experiments, the open discontinuities are filled with air that not able to propagate S-wave. We set the S-wave velocity as a small value, 1m/s. For embedded closed discontinuities, S-wave travels around 4200m/s through limestone which is 1.6 times slower than compressional wave.

3.1.4.3. Pressure front propagation case

The pressure diffusion propagation was control by its diffusivity. Diffusivity is an important parameter indicative of the diffusion mobility. It explained the velocity of diffusion that related to porous media permeability, porosity, compressibility and fluid viscosity. The equation of diffusivity expressed in Eq.4. The mainly difference is the permeability of fracture and matrix. For closed discontinuities, the crack permeability is lower than matrix. On the contrary, open crack is more permeable than sandstone matrix.

3.2. Transmitter-receiver configuration

In our study, we focus on analysis the crack system inside 2-D crack-bearing material. The material size is 150mm by 150mm which divided into 500 by 500 grids. The 2D squared numerical model has 1 wave source and 28 receivers. The wave source is in the middle of the left boundary, and the remaining three boundaries have 28 receives equally located. Then the fast-marching method

(FMM) will apply to simulate the propagation of the wave/diffusion front from this single source to the receivers around the crack-bearing material as the initial dataset for classification.

3.3. Fast Marching Method (FMM) for First Arrival Simulation

The fast-marching method (FMM) is a front-tracking method created by James Sethian for solving boundary value problems of the Eikonal equation. Eikonal equation characterizes the evolution of a closed surface as a function of time with specified velocity on the given surface, express as:

$$\boldsymbol{f(x)|\nabla u(x)| = 1} \ \textit{for} \ \boldsymbol{x \in \Omega} \tag{3}$$

$$\boldsymbol{u(x) = 0} \ \textit{for} \ \boldsymbol{x \in \partial\Omega} \tag{4}$$

where $u(x)$ represents the travel time of the front wave to reach the location x, $f(x)$ stands for the speed at x known as velocity function, Ω is the open set with well-behaved boundary, $\partial\Omega$ is the boundary, and x is the coordinate system. As described, this equation is a non-linear partial differential equation to solve the wave propagation problems. In this paper, we use FMM to approximate the solution to the Eikonal equation. Scikit-fmm is a python extension module which implements the fast-marching method used in our program.

For pressure diffusion propagate front, the Eikonal equation is described based on diffusivity.

$$\sqrt{\boldsymbol{\alpha(x)}}\boldsymbol{|\nabla\tau(x)| = 1} \tag{5}$$

Where, the diffusivity $\boldsymbol{\alpha(x)}$ defined as:

$$\boldsymbol{\alpha(x) = \frac{k(x)}{\phi(x)\mu C_t}} \tag{6}$$

Eq. 5 tells that the pressure "front" propagates in the reservoir with a velocity given by the square root of diffusivity. For homogeneous reservoirs, $\boldsymbol{\tau(x)}$ is related to physical time through a simple expression of the form $t(x) = \frac{\tau(x)^2}{c}$ where the constant c depends on the specific flow geometry (Xie, 2015). For linear, radial, and spherical flows, c is 1⁄4 2, 4, and 6, respectively (Kim et al. 2009).

3.4. Dataset description – samples, features, target

Data collection is a critical step towards the development of machine learning models. Before beginning to train the model, we should transfer our data in a way that can be fed into a machine

learning model. The preprocessing in this work is to deal with features on the same scale by using the standardization approach. This study will include three tasks: (1) classify the dominant crack orientation with different fracture dispersion; (2) classify the crack dispersion with dominated orientation; (3) classify the crack spatial distribution. All 2D numerical models are using the discrete fracture network (DFN) method to various types of mechanical discontinuities system.

Task 1 will build classification model to identify four dominant crack orientations of 0°, 45°, 90° and 135°. Figure 2 shows the experimental configurations for different orientations at dispersion equal to 10. For each orientation, 10,000 sample's travel time are recorded as the model dataset. It took around 2 hours on a Dell workstation with 3.5GHz Intel Xeon CUP and 32GB RAM. The front wave travel times at different sensors are features of data-driven model.

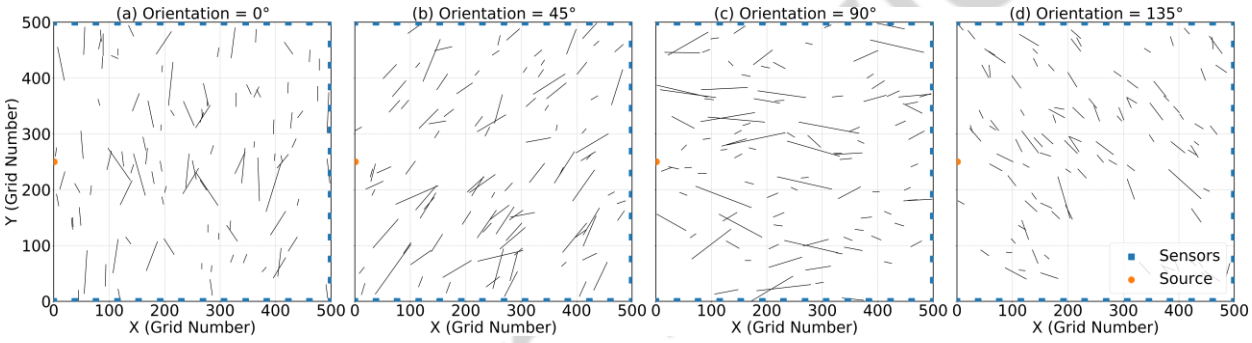

Figure 2. Experiment configuration for orientation classification with dispersion = 10

Features are columns in the input dataset. It can be further simplified by feature engineering to avoid overfitting. The samples from different orientation are labeled as 0, 1, 2, 3 which are our model targets. In a word, features briefly explained the input fed into the system and the label would be the output models are expecting. For the classification of dispersion, we built three crack-bearing models with distinct dispersions around dominant crack orientation set as vertical in our case. The outputs are labeled as 0, 1, 2 which stand for dispersion equal to 0, 500 and 1000. The simulation models are presented in figure 3. When the dispersion factor is set to 0, the crack orientations are equally distributed in all directions. As a comparison, when dispersion is set to 1000, orientations are nearly aligned along the direction of dominant orientation. The entire dataset

comprising 30,000 samples, such that each sample has 28 features and 1 target for only compressional wave.

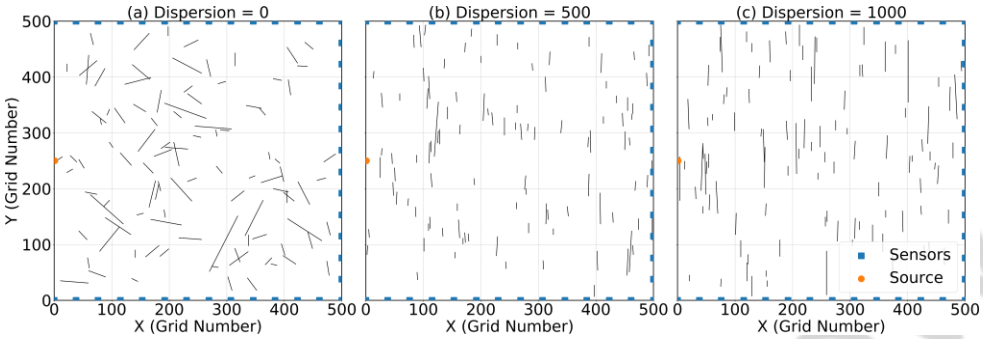

Figure 3. Experiment configuration for dispersion classification with vertical orientation

Final task is to identify four type of spatial distribution created using intensity functions that shows in figure 4. Each type of distribution model is embedded with 100 fractures with randomized length, orientation, and dispersion. The fracture systems are created by the DFN method. The dataset will have 40,000 samples, with 4 targets. The data will be divided to training and testing data to be applied in data-driven models.

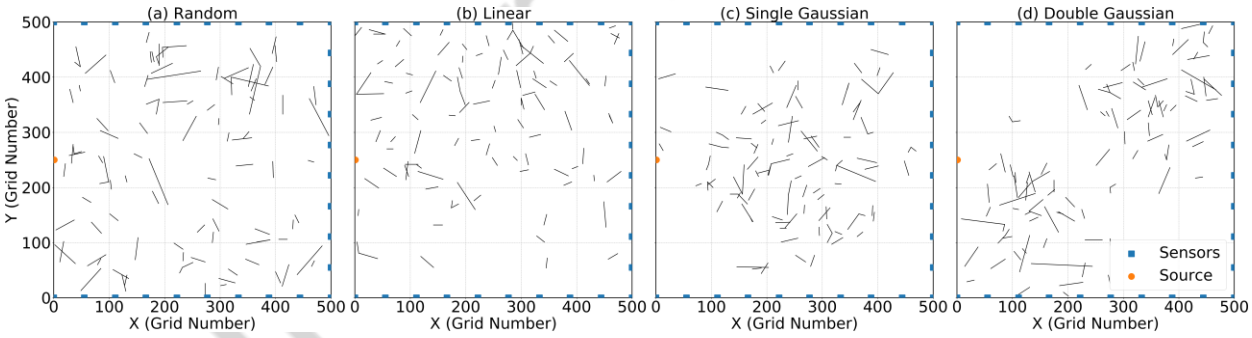

Figure 4. Experiment configuration for spatial distribution classification

4. Data-Driven Classification of Materials with Embedded Open Mechanical Discontinuities

4.1. Comparison of accuracy - P wave

In this section, we trained 7 data-driven classifiers on p-wave travel time recorded at 28 sensors and investigated their performances on three characterization tasks for material with embedded open mechanical discontinuities. The pores in the background material and the embedded

mechanical discontinuities (i.e. cracks) are assumed to be filled with air. The source-sensor configuration for generating the dataset is the same as the numerical experiment described in the previous section. The dataset used for the training and testing are the p-wave travel time recorded by the 28 sensors placed around the material. Hyperparameters of the 7 data-driven classifiers are tuned by performing grid search with 5-fold cross-validation. Our hypothesis is that multi-point measurements of p-wave travel time can be processed by data-driven classifiers to identify certain bulk aspects of the network/cluster of open (i.e air filled) mechanical discontinuities.

4.1.1. Classification of the Dominant Crack Orientation

Task 1 contains 4 different types of the crack clusters that differ in orientation with fixed dispersion around the dominant orientation. In order to compare the effects of dispersion on this task, we set one experiment at dispersion equal to 10, and the second experiment at dispersion equal to 50. We recorded p-wave traveltimes for 20,000 samples of materials with a specific orientation for 4 different orientations $(0°, 45°, 90°, 135°)$. The complete dataset has 28 features represents the sensor locations. 70 % of the entire data is set as training data, and the remaining is testing data. The four orientations are labeled as 0, 1, 2, 3 as our testing target. In other words, a target value of 0, 1, 2, and 3 represents crack orientation of 0°, 45°, 90° and 135°. The classification accuracy is presented as the bar plot in figure 5. The mean accuracy of 7 classifiers for dispersion equal to 10 is around 0.89. As a comparison, the higher dispersion which gives more uniform cracks shows better performance. The average accuracy of this experiment is around 0.98. However, the best classification performance for these two experiments are both from voting classifier.

4.1.2. Classification of the Crack Spatial Distribution

Task 2 exploiting the spatial distribution features (i.e., direction feature and density feature) of the cracks under classification models to identify different types of spatial distribution in crack-bearing materials. Four types of spatial distribution of cracks, as shown in the upper panel of **Error! Reference source not found.**. Spatial distribution of cracks is affected by the mechanical properties and the surrounding environment of the material. Each type of crack-bearing model is embedded with 100 fractures of randomized length, orientation, and dispersion. Similar with the

previous test, the dataset has 20,000 samples with four target and 28 features. The accuracy for the classifiers is about 0.81.

4.1.3. Classification of the Crack Dispersion with Dominated Orientation

In this part, a numerical experiment is conducted to classify the three diverse crack dispersion: 0, 500, 1000 with dominated orientation. The orientation for three dispersions is set as vertical (Orientation = 0°). The dataset includes 15,000 samples, 5000 for each dispersion. The target for each sample is either 0, 1 or 2, depending on the dispersion of the crack cluster. The overall accuracy is around 0.60 for the 7 models. The best classifiers are gradient boosting and voting classifier which gives the accuracy around 0.61.

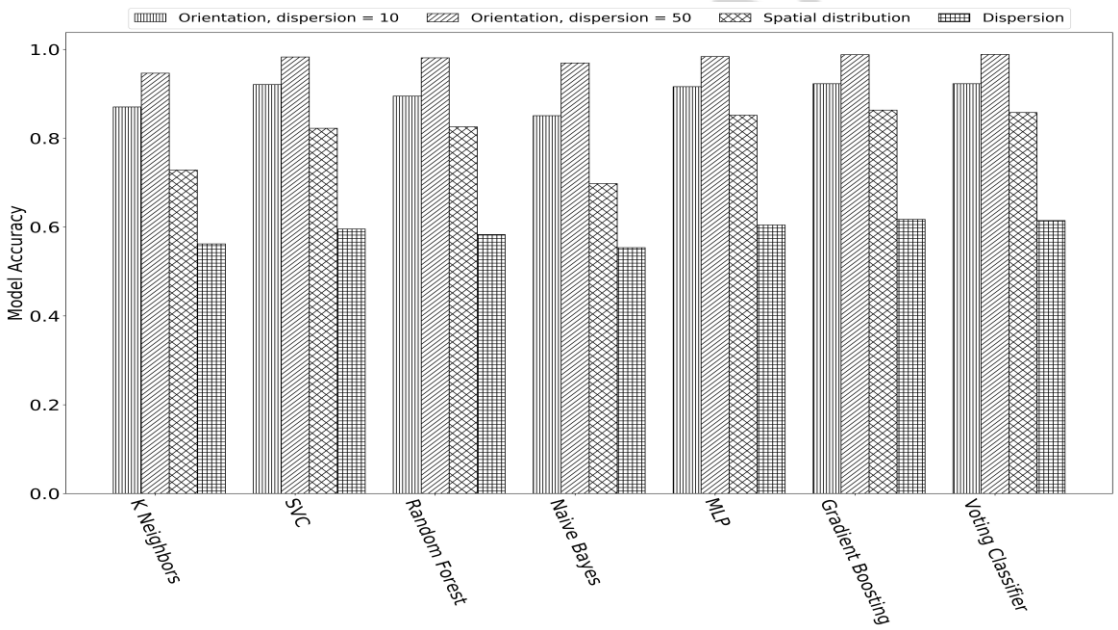

Figure 5. Open Mechanical Discontinuities Classification performance for P wave

4.2. Comparison of accuracy - P + S wave

It is obvious that the accuracy of the P wave model for the classification of dispersion is not satisfied. Especially the classification of dispersion, the accuracy only reaches 0.6. We add shear waves (S wave) inside to capture more information. The simulation model will record the P and S wave travel time at 28 sensors for each case. The dataset contains thousands of samples, 56 features and several targets depend on the task purpose. It has been proved that the overall accuracy of 7

classifiers has been risen in varying degrees. The following three sections will further discuss how the accuracy changed case by case.

4.2.1. Classification of the Dominant Crack Orientation

The model sets are the same with the upper section, four orientations are our target labeled as 0, 1, 2, and 3. For dispersion equal to 10, we have 20,000 samples. 6000 of them generated the testing data to evaluate the classification performance. The average accuracy increased from 0.89 to 0.95 which means the S wave helps the classifier to characterize the crack states. On the other hand, when dispersion set at 50, the accuracy also rises 0.01 and finally achieve 0.99. the classification of those two cases are more reliable when adding S sonic front wave travel time.

4.2.2. Classification of the Crack Spatial Distribution

Turning now to the experiment evidence for spatial distribution classification. Processing and the dataset used by is comparable to that used in the upper sections. As can be seen from the comparison of figure 5 and 6, this study did not show any significant increase in Naive Bayes classifier. However, the overall accuracy of the 7 classifiers was found increased from 0.81 to 0.85 after adding the features of S wave.

4.2.3. Classification of the Crack Dispersion with Dominated Orientation

Then this system of classification was developed for the purpose of distinguishing the crack dispersion. What is striking about the figures is that this case has a significant improve.

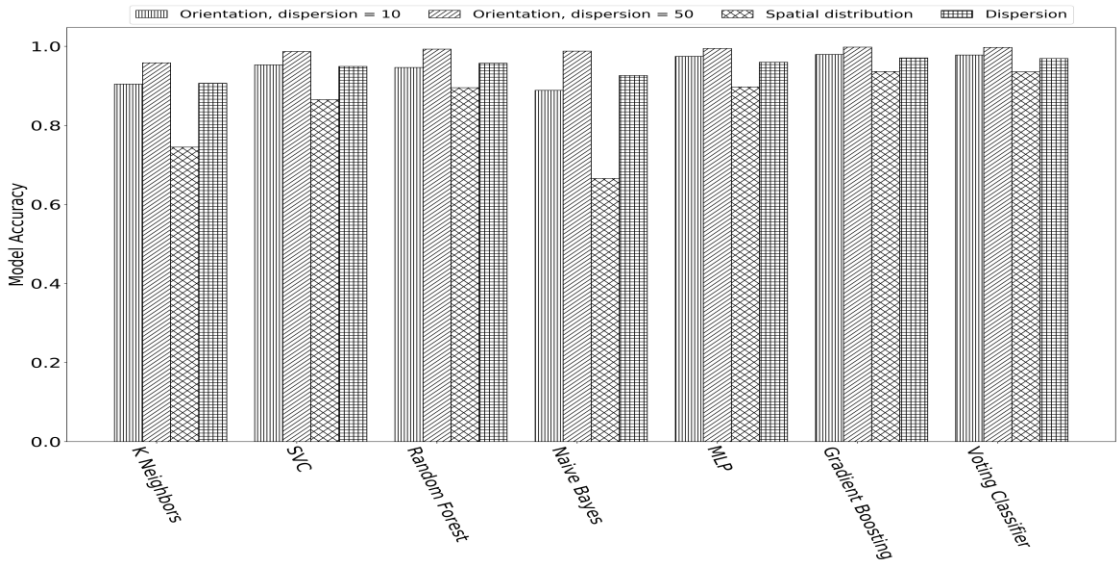

Figure 6. Open Mechanical Discontinuities Classification performance for P + S wave

Overall, comparison of the two results reveals more than 50% growth. The final accuracy can reach 0.95.

4.3. Comparison of accuracy - P + S wave + Pressure Diffusion

The previous cases were designed to determine the effect of mechanical waves. Furthermore, we added pressure diffusion as another set of features. The feature number expended to 84 features. Similarity, the sensor located around the material recorded the front time of the sonic waves and pressure diffusion. This task expected to be the best performance in our study. Then more evaluation metrics are used to analysis the classification results such as precision, recall and F-1 score. The section below describes the cases designed in this task.

4.3.1. Classification of the Dominant Crack Orientation

As mentioned, the four dominant orientations are 0°, 45°, 90° and 135° which labeled as 0, 1, 2, 3. The large dataset has 20,000 rows and 84 columns stand for the features from three different waves. The classifiers accuracy is 0.95 and 0.99 for dispersion equal to 10 and 50, respectively. The table below illustrates precision, recall and F1 score for both low and high dispersion are good enough. precision, recall and F-1 score can even reach 1 at higher dispersion case. It means the performance of this task are perfect for determine the dominant orientation.

4.3.2. Classification of the Crack Spatial Distribution

This case is the classification to identify the crack spatial distribution such as random, Linear, Single Gaussian and Double Gaussian. No significant differences were found after the pressure diffusion added. The overall accuracy is remaining around 0.85. Precision, recall and F-1 score are 0.92 indicates a good model.

4.3.3. Classification of the Crack Dispersion with Dominated Orientation

The last case in this chapter is the dispersion classification. The average accuracy of 7 classifiers increased to 0.96. The results in this chapter indicate that the classifier performance for open crack characterization in the material. The next chapter, therefore, moves on to discuss the difference between open and embedded closed mechanical discontinuities.

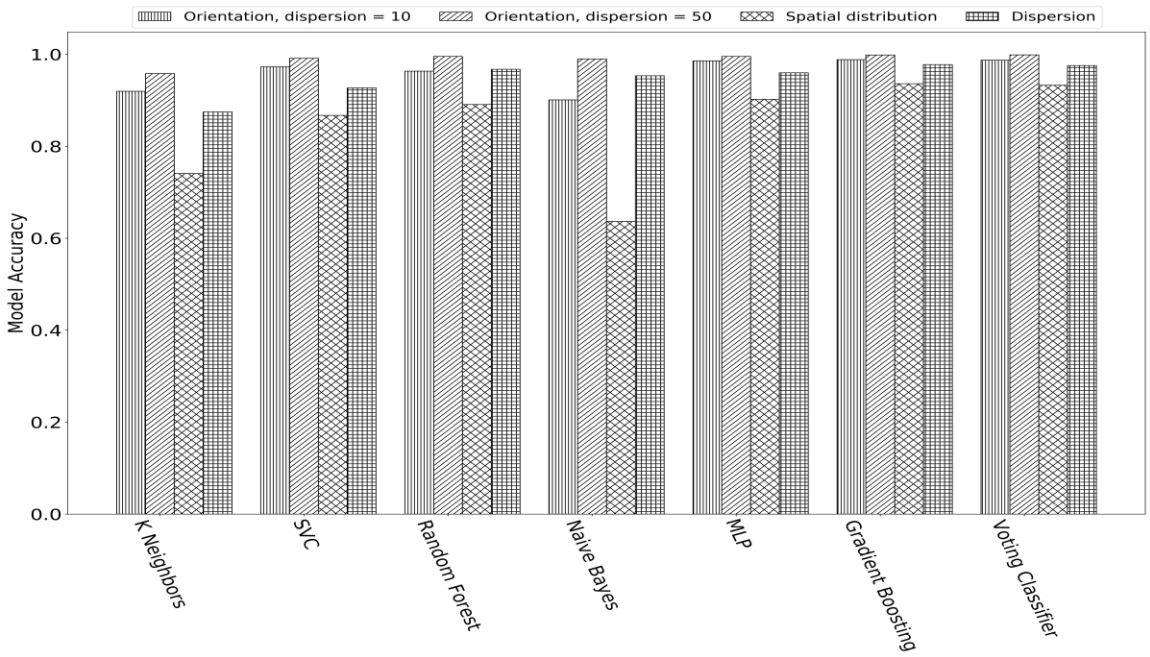

Figure 7. Open Mechanical Discontinuities Classification performance for P + S wave and Pressure Diffusion

Open Crack	Precision	Recall	F-1 Score
Orientation-case1	0.99	0.99	0.99
Orientation-case2	1	1	1
Dispersion	0.98	0.98	0.98
Distribution	0.92	0.92	0.92

Table 1. Evaluation metrics for P + S wave and Pressure Diffusion

5. Data-Driven Classification of Materials with Embedded Closed Mechanical Discontinuities

5.1. Comparison of accuracy - P wave

As we mentioned at the beginning of our study, the matrix is set as sandstone with 20% porosity. And the embedded closed mechanical discontinuities are fully filled by limestone. The first task is capturing the information of mechanical P wave from 28 sensors around the material. Then 4 cases are used to test the classifiers.

5.1.1. Classification of the Dominant Crack Orientation

This section will have two cases: 1). Four dominant crack orientations with dispersion equals 10. 2). Four dominant crack orientation with dispersion is 50. The higher dispersion case gives better accuracy than lower case. The accuracy of case 1 is 0.995 and case 2 already reach 0.999. The finding of this section is the same with open crack: higher dispersion model is more reliable.

5.1.2. Classification of the Crack Spatial Distribution

Then, we check the accuracy of the cemented crack model for spatial distribution classification. The overall accuracy is around 0.87 for 7 classifiers when using compressional wave travel time as dataset. The model capacity may further ameliorate in the latter parts after have the shear wave and pressure diffusion.

5.1.3. Classification of the Crack Dispersion with Dominated Orientation

The dispersion classification is the worst case compared to the other cases. The accuracy is around 0.6. SVC and Voting classifiers have the highest accuracy are about 0.61. Most classifiers do not perform very well on materials with different crack dispersions which is same as the open crack. That's the reason why we introduce shear wave to our model. The next section will discuss the findings with compressional and shear waves.

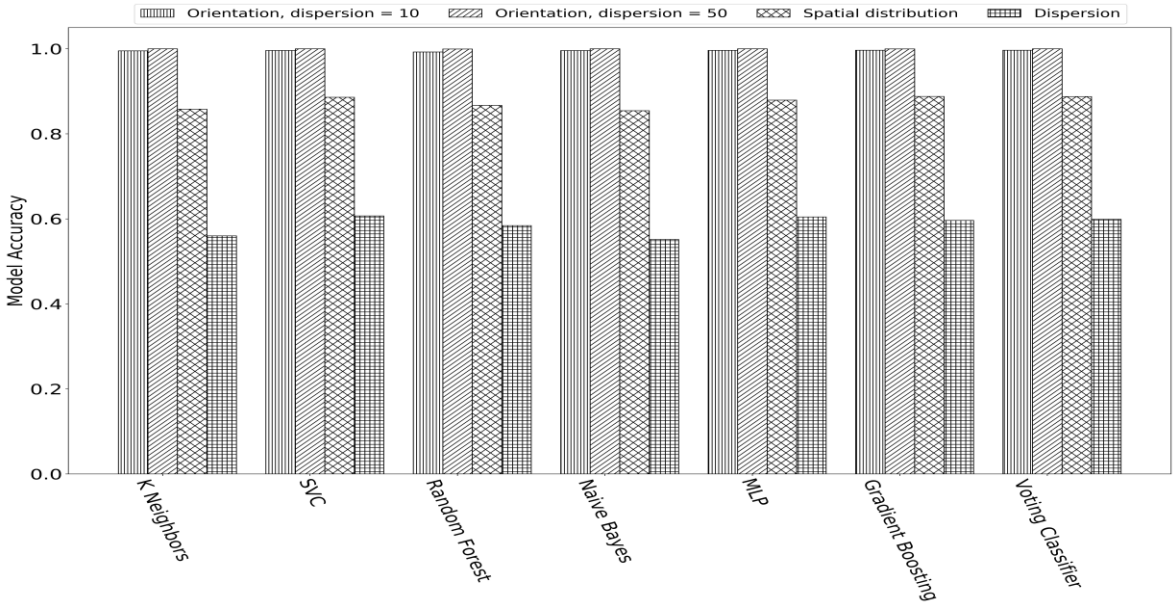

Figure 8. Embedded Closed Mechanical Discontinuities Classification performance for P wave

5.2. Comparison of accuracy - P + S wave

This section set out with the aim of assessing the importance of shear waves in closed mechanical discontinuities. The classification of crack dispersion has the greatest improvement among the four cases. However, the accuracy of crack orientation classification remains similar with the upper section. In summary, characterization of crack dispersion requires higher information content captured by the sensors as compared to the characterization of crack orientation in the crack cluster.

5.2.1. Classification of the Dominant Crack Orientation

The best classifier for classify the dominant crack orientation is voting classifier for both lower and higher dispersion. The overall accuracy for both cases can reach 0.90. Moreover, the higher dispersion case can reach 1.00. No obvious changes are made in the classification of orientation.

5.2.2. Classification of the Crack Spatial Distribution

The spatial distribution case has been proved can be increased by adding feature numbers. The average of 7 classifiers changes from 0.87 to 0.89. However, the Naïve Bayes classifier is decreased to 0.72. Beyond that, the other classifiers are around 0.9. The best accuracy is about 0.95 from gradient boosting classifier.

5.2.3. Classification of the Crack Dispersion with Dominated Orientation

The dispersion case with only compressional wave travel time has poorly performance. Figure 8 reveals that there has been a steep rise with more sensor information. Numerically, the accuracy dramatically increases from 0.60 to 0.94. This combination of findings provides some support for the significant meaning of adding the shear wave.

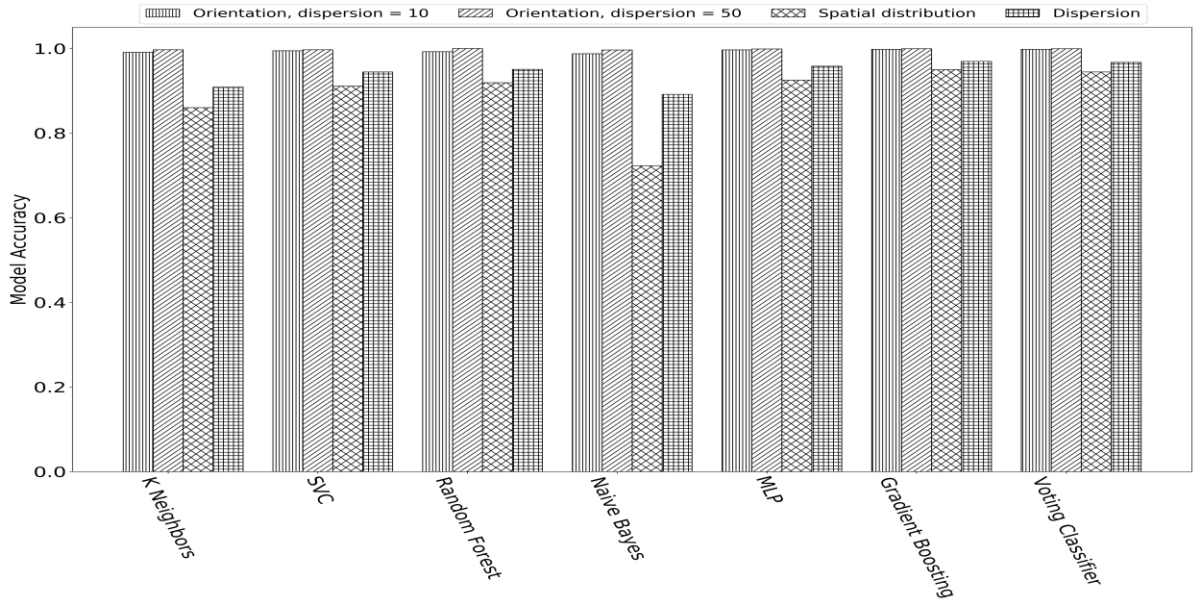

Figure 8. Embedded Closed Mechanical Discontinuities Classification performance for P + S wave

5.3. Comparison of accuracy - P + S wave + Pressure Diffusion

5.3.1. Classification of the Dominant Crack Orientation

A more comprehensive study would consider how the model perform after add pressure diffusion. The data-driven classifiers were trained and tested on simulated dataset generated for simple sonic wave propagation and pressure diffusion through simple crack-bearing materials. According to Figure 9, both lower and higher dispersion cases for orientation classifier can achieve 1.00 when using 84 features in the model. The results, as shown in Table 1, indicate that precision, recall and F-1 score are also each 1.00.

5.3.2. Classification of the Crack Spatial Distribution

The experimental procedure is the same as in previous cases. It is investigated that spatial distribution classification gets the benefit from the pressure diffusion. Overall, the model accuracy increased to 0.90. The worst model is from Naïve Bayes around 0.67. The other models have performed well with an accuracy in the range of 0.87 to 0.95. Further statistical tests revealed precision, recall and F-1 score is around 0.93.

5.3.3. Classification of the Crack Dispersion with Dominated Orientation

The classification of dispersion has been significant improved by shear waves. Then the pressure diffusion also helped to further optimize the data-driven model. The final classifiers accuracy is about 0.96. Gradient boosting and voting classifiers have the best performance gives accuracy around 0.98.

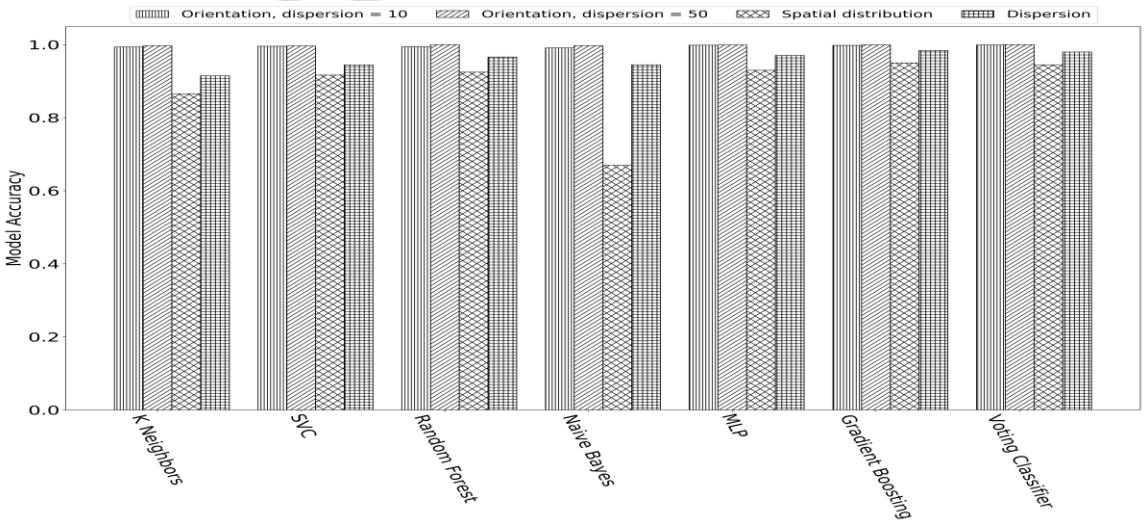

Figure 9. Embedded closed mechanical discontinuities classification performance for P + S wave and pressure diffusion

Cemented Crack	Precision	Recall	F-1 Score
Orientation-case1	1	1	1
Orientation-case2	1	1	1
Dispersion	0.97	0.97	0.97
Distribution	0.93	0.93	0.93

Table 2. Evaluation metrics for P + S wave and pressure diffusion

In general, gradient boosting and voting classifiers are the top two among the 7 classifiers from our study. In the following section, we will compare the difference of open and closed mechanical discontinuities for these two models.

6. Comparison of Data-Driven Classification of Materials with Embedded Closed vs. Open Mechanical Discontinuities

In this section, in order to simplify and clear the comparison of two types of discontinuities, we named our cases from case1 to case4. Case1 and case2 represent the classification of orientation with lower and higher dispersion, respectively. Case3 stands for the dispersion classification with vertical cracks. Finally, case4 is used to classify the spatial distribution. Each case will have three experiments: 1) with compressional waves only, 2) with compressional and shear waves, 3) combination of P, S waves and pressure diffusion. Then, we move on to discuss our findings after comparing.

6.1. Gradient Boosting Classifier

The horizontal bar plot shows the difference between closed and open discontinuities accuracy. There are several important differences between closed and open discontinuities. Positive value means cemented crack has better performance compared to open crack. Instead, the negative one such as case3 means that open crack has well performance for dispersion classification when using P and S waves. The most important clinically relevant finding was that add more features to the data-driven model will help to reduce the difference between open and closed crack.

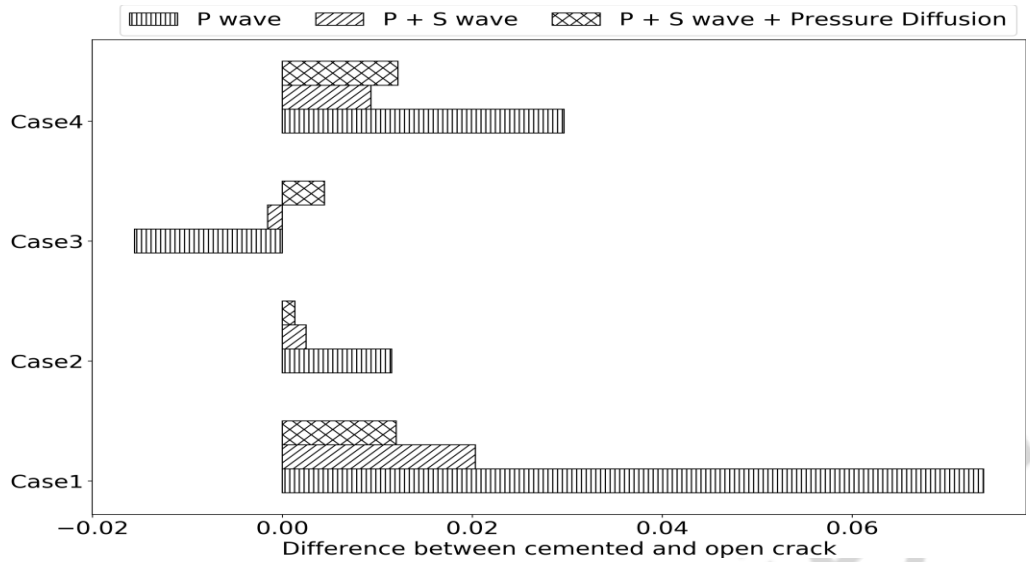

Figure 10. Comparison of open and closed mechanical discontinuities-Gradient Boosting

6.2. Voting Classifier

The voting classifier is an ensemble method that combines predictions of other 6 classifiers described above based on a certain rule. In our study, voting classifier performs the best for purposes of characterization. This finding in Figure 11 broadly supports the conclusion from gradient boosting classifier.

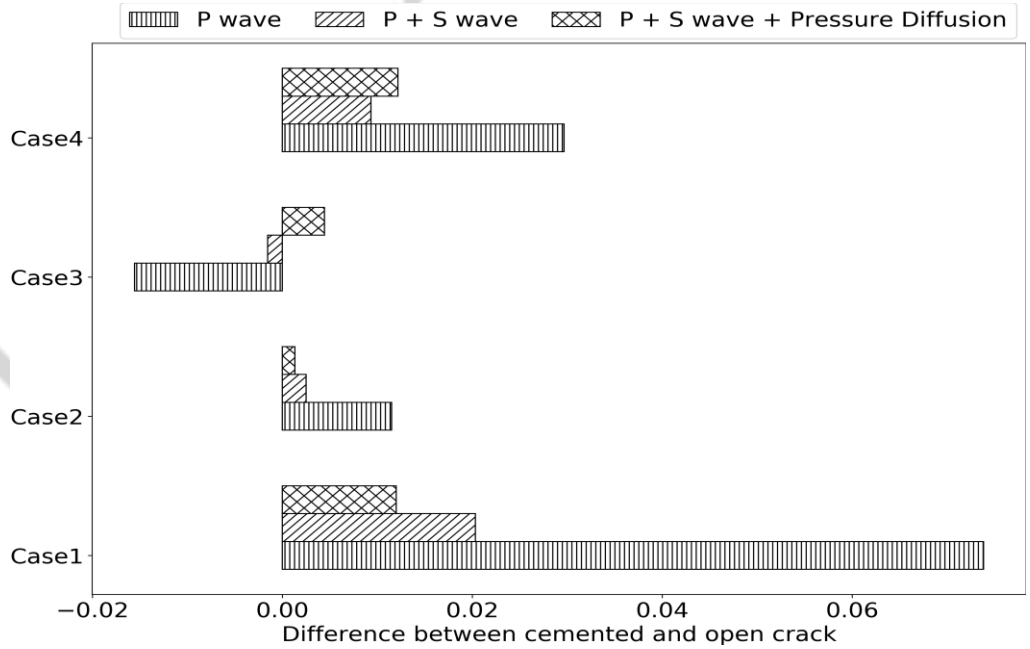

Figure 11. Comparison of open and closed mechanical discontinuities-Voting Classifier

7. Classification model optimization

7.1. Feature selection – Univariate filter methods

As the number of features increases by adding wave information, the model becomes more complex. Dimensionality reduction is the process of reducing the number of feature sets. In general, avoiding overfitting is a major motivation for performing feature reduction. Remove redundant features also helps to save computation time. Meanwhile, less data means that algorithms will train faster. Dimensionality reduction could be done by many feature selection methods. The univariate filter method is a common feature selection tool by using statistical tests to evaluate the features. Univariate filter methods evaluate each feature individually without considering feature interactions and providing scores to each feature. Chi-square is statistic way to show the relationship between categorical variables. Mutual information is a measure of dependency between random varies relies on the computation of the feature probability distribution. In this test, we set threshold for both chi-square and mutual information score to select features. The threshold we selected is 0.2 means only use the features that both chi-square and mutual information score higher than 0.2. The reduction features create the new training and testing dataset with new feature numbers. The classification models on the training dataset will test on testing dataset to evaluate the generalization of model performance. Consider that the cases for distinguish crack dominate orientation accuracy have already reach 1.00, we only apply this technique on the classification of dispersion and spatial distribution to reduce the overfitting and improve model performance. The number of features selected is changing case by case. However, from the evaluation metrics like model accuracy and F-1 score, the model performance reduced by dimensionality reduction. That means the model is not overfitting from the dataset.

7.2. Sensor addition

As we discussed, all the features from the model are not deductible. Features could be strongly relevant, relevant, weakly relevant or irrelevant (Bottou, 2010). Even if some features are irrelevant, having too many is better than missing those that are important. To the end, we can try to add sensor numbers around the material to improve the model. The sensor number increased from 28 to 40, each boundary will have 14 sensors instead of 10. The source is still one located in the middle of the left boundary, and other feature characterization remains same as the previous experiments. By doing this, the feature number for only compressional wave changed to 40, and

80 for P and S wave. Finally, we have 120 features after adding pressure diffusion. As a result, no clear benefit of additional features could be identified in this analysis.

8. Conclusions

In this study, we proposed classification algorithms to capture crack characterizations by measuring the sonic wave and pressure diffusion arrival times. FMM simulation is implemented to simulate the wave front travel time. The arrival times are used as input dataset to train and test 7 selected data-driven methods. The findings of this study have several important implications for future practice.

- Use both compressional and shear wave travel times can improve the data-driven model.
- Apply pressure diffusion with sonic waves can optimize the machine learning algorithms.
- Embedded closed discontinuities performs better than open discontinuities on the classification models.
- Voting classifier and gradient boosting classifier outperform other models in this study.
- This study shows that machine learning models exhibit best classification performance on classifying crack dominant orientations. The model accuracy and F-1 score can reach 1.00 when combine sonic waves and pressure diffusion.
- Neither reduce feature dimensionality nor add sensor numbers can improve the algorithms.

9. Recommends for future work

- Build a robust model to trace the development of dynamic crack in the material.
- Regression model can be used to in the similar way to analysis the fracture system.
- Use full set of travel time may improve the model performance.
- The effect of other crack parameter such as crack length or numbers can also be investigated.
- 3-D model could be created for the same purpose with better accuracy.

10. References

Zhang, L., & Einstein, H. H. (1998). Estimating the mean trace length of rock discontinuities.

Rock Mechanics and Rock Engineering, 31(4), 217-235.

Misra, S., & Li, H. (2020). U.S. Patent Application No. 16/529,462.

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.

Wu, Y., Misra, S., Sondergeld, C., Curtis, M., and Jernigen, J. 2019. Machine learning for locating organic matter and pores in scanning electron microscopy images of organic-rich shales. Fuel, 253, 662-676.