Using Machine Learning to Discern Eruption in Noisy Environments: A Case Study Using CO₂-Driven Cold-Water Geyser in Chimayó, New Mexico

by Baichuan Yuan, Yen Joe Tan, Maruti K. Mudunuru, Omar E. Marcillo, Andrew A. Delorey, Peter M. Roberts, Jeremy D. Webster, Christine N. L. Gammans, Satish Karra, George D. Guthrie, and Paul A. Johnson

ABSTRACT

We present an approach based on machine learning (ML) to distinguish eruption and precursory signals of Chimayó geyser (New Mexico, U.S.A.) under noisy environmental conditions. This geyser can be considered a natural analog of CO₂ intrusion into shallow water aquifers. By studying this geyser, we can understand upwelling of CO₂-rich fluids from depth, which has relevance to leak monitoring in a CO₂ sequestration project. ML methods such as random forests (RFs) are known to be robust multiclass classifiers and perform well under unfavorable, noisy conditions. However, the extent of the RF method’s accuracy is poorly understood for this CO₂-driven geysering application. The current study aims to quantify the performance of RF classifiers to discern the geyser state. Toward this goal, we first present the data collected from the seismometer that is installed near the Chimayó geyser. The seismic signals collected at this site contain different types of noises such as daily temperature variations, animal movement near the geyser, and human activity. First, we filter the signals from these noises by combining the Butterworth high-pass (BH) filter and an autoregressive (AR) method in a multilevel fashion. We show that by combining these filtering techniques in a hierarchical fashion leads to a reduction in noise in the seismic data without removing the precursors and eruption event signals. We then use RF on the filtered data to classify the state of geyser into three classes: remnant noise, precursor, and eruption states. RF classifier is constructed based on the comprehensive features extracted using the Tsfresh Python package. We show that the classification accuracy using RF on the filtered data is greater than 90%. Denoising seismic signals from both daily trends and human activity enhances RF classifier performance by 7%. These aspects make the proposed ML framework attractive for event discrimination and signal enhancement under noisy conditions, with strong potential for application to monitor leaks in CO₂ sequestration.

INTRODUCTION

Growing interest in collecting field-scale data for quantifying CO₂ leakage has drawn attention to CO₂-driven cold-water geysers (Keating et al., 2011). Because of the high velocity of CO₂-rich fluid discharge, these geysers can be thought of as a natural analog to CO₂ leakage in a carbon sequestration project. CO₂-driven cold-water geysers are similar to thermally driven geysers (such as the ones in Yellowstone) because they have a conduit to release CO₂-rich fluids. Although thermally driven geysers (Manga and Brodsky, 2006; Hurwitz and Manga, 2017) have naturally existing conduits, in the case of CO₂-driven cold-water geysers, the conduits are man-made (typically a wellbore). There are various field sites of CO₂-driven cold-water geysers across the world (Glennon and Pfaff, 2005). Examples of some cold-water geysers in the U.S.A. include the crystal geyser in Utah, the Tenmile geyser in Utah, and the Chimayó geyser in New Mexico. The focus of this article is the Chimayó geyser in New Mexico. Our aim is to provide information on the state of this geyser by analyzing noisy seismic signals. To achieve this goal, we need a framework to differentiate signals that are far in time from an eruption event from those signals that are closer to the event under noisy conditions. An approach to develop such a framework is provided through machine learning (ML, Mudunuru, Chillara, et al., 2017; Rouet-Leduc et al., 2017; Holtzman et al., 2018;...
ML is a widely used tool for extracting features (Kanter and Veeramachaneni, 2015; Christ et al., 2018) and classifying seismic signals. In addition to earthquake detection and prediction (Rouet-Leduc et al., 2017, 2018), ML is shown to be successful in various other subsurface applications (Mudunuru, Karra, Harp, et al., 2017; Mudunuru, Karra, Makedonska, et al., 2017; Marone, 2018; Mudunuru et al., 2018; Vesselinov et al., 2018; Hunter et al., 2019). Within the context of ML, there are many available techniques that are able to predict the evolution of time-series and distinguish seismic signals, within a certain tolerance (De Silva and Leong, 2015; Konar and Bhattacharya, 2017). The presence of noise in data is a major concern in classification approaches as the performance of ML classifiers deteriorates (Li et al., 2010; Mnih and Hinton, 2012; López et al., 2013; Sáez et al., 2016). Interpretability of ML models is hindered by poor classification performance, which can have a negative effect on event discrimination capability. To overcome this drawback, we first filter the seismic signal multiple times using high-pass filter and an autoregressive (AR) model to remove or reduce the effect of different types of noises (temperature effect, human activity, animal activity, etc.). Second, we perform comprehensive feature extraction and downselect features to construct the random forests (RFs) model. Third, we classify whether a seismic signal is remnant noise or precursor or eruption signal. The proposed methodology is compared against classical time-series methods such as dynamic time warping (DTW, Xi et al., 2006; Konar and Bhattacharya, 2017) on filtered data and ML methods with partial filtering. We show that better accuracy can be achieved through a combination of both filtered data and a robust classification algorithm that is less sensitive to noise. For ML classification, we use RFs, which is a model-free ML method that is less influenced by noise (Zhu and Wu, 2004; Sáez et al., 2016). A major advantage of model-free ML methods is that they conform to the intrinsic data characteristics with fewer assumptions and without the use of any a priori models. Model-free methods (such as RF) construct nonparametric models using ensembles of multiple base learners without simplifying the underlying problem (Breiman, 2001; Genuer et al., 2008).

Main Contributions and Outline of the Article
The main contribution of this study is to develop an ML-based framework to classify seismic signals under noisy environments and apply it to the Chimayó geyser eruptions. An advantage of the proposed method is that even under high anthropogenic noise, we can differentiate the seismic signals that are far away from eruption time from the signals that are closer to the eruption time with an accuracy greater than 90%. Moreover, the computational time taken by the proposed ML model to classify a given seismic signal as an eruption or a noneruption event is $O(10^{-4})$ s on a laptop. This makes our ML methodology ideal for usage in eruption event discrimination (such as detecting precursors and differentiating it from background noise) or forecasting geyser eruptions in real time under noisy conditions. The article is organized as follows: The Geyser Data and ML Methodology section provides a detailed description of our proposed ML method for distinguishing eruption events from noneruption events. First, we describe the geyser data, the location of sensors, and relevant background on data collection. Second, we present a method to preprocess the seismic signals to remove temperature effects and background noise. Third, we present a feature engineering approach to comprehensively extract various features from the filtered seismic signals. Feature selection is based on scalable statistical hypothesis testing, which is used as a basis to downselect features for ML analysis. Feature importance is performed on the downselected features using the RFs method. Fourth, we construct ML classifiers based on RFs for discriminating events on the filtered and partially filtered data. The Results section discusses the results of the proposed ML method and DTW in classifying signals. The accuracy of the RF classifiers on filtered and partially filtered data is also provided in this section. Finally, conclusions are drawn in the Concluding Remarks section.

GEYSER DATA AND ML METHODOLOGY
Site Description and Data Resources
The Chimayó geyser located near Chimayó, New Mexico, is one of the few CO₂-driven cold-water geysers in the United States. It is a man-made geyser and was formed when locals drilled a well for water usage. There is a shallow aquifer that is beneath the geyser. The aquifer is within a sub-basin of the Rio Grande rift, where there is a regional development of CO₂ gas. Experiments performed on this geyser showed elevated CO₂ conditions at the site (Keating et al., 2011). Moreover, many of the drinking water wells in the neighborhood of this geyser have high levels of dissolved CO₂. The eruption cycle of the Chimayó geyser is a simple two part process. The system begins with a very long recharge period. Then, there is a possible bubbling activity or minor eruption before major eruption activity (Watson et al., 2014). The entire minor and major eruption duration averages approximately 5 min. The eruption dynamics of the geyser is mainly gas dominated due to a small cross-sectional area of the wellbore (Watson et al., 2014).

Figure 1 shows the map of the Chimayó geyser and the location of seismic station RGEYB. This seismic station is located approximately 3 m from the well. The sampling frequency of RGEYB is 200 Hz. We analyze 18 days of continuous data collected from May to November 2017. Eruption images are captured using a motion sensor installed at this station, which are used to label the seismic data. The motion sensor records the geysering event time and captures the image when CO₂-rich fluids are ejected out of the wellbore. The major challenge in analyzing this seismic dataset is that the anthropogenic noise is higher than the seismic signals corresponding to geyser’s activity (e.g., see Fig. 2c). The time between eruptions is 17–23 hrs whereas the background noise is around the clock (generally higher during the day).
region. An AR model divides the signal into two additive components, a predictable signal and a prediction error signal. Let the time-series data obtained from RGEYB seismometer be denoted as $X_t$, in which $t = 1, 2, \ldots, n$. Let $AR(p)$ be an AR model of order $p$ that estimates $X_t$ using the lagged variable $X_{t-i}$, $i = 1, 2, \ldots, p$, as: $X_t = \epsilon + \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t$, in which $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the parameters of the model, $\epsilon$ is the constant term, and $\epsilon_t$ is white noise. We use Akaike information criterion (AIC, Seabold and Perktold, 2010) to select the value of $p$. Using statsmodels Python package (Seabold and Perktold, 2010), we construct the $AR(p)$ model on the first 5% of the BH filtered data (within the training set). Then, we compute a one-step prediction for the rest of the seismic time-series data. $AR(p)$ model development does not involve testing data. $AR(p)$ model predictions form the basis to calculate prediction error filter (PEF). The PEF of the $AR(p)$ model is defined as the BH-filtered seismic time series minus the $AR(p)$ model prediction.

**Feature Engineering for Geyser State Classification**

After preprocessing the raw seismic signals, we apply a RF classifier on the filtered data to distinguish eruption and noneruption signals. ML methodology to classify geyser state starts with feature extraction on a sliding time window to extract a feature vector. Then, the feature vector is imported into an RF classifier, which provides feature importance and also acts as a decision-making tool that is capable of categorizing signals into different classes. In each time window, there are lots of data points due to the high-sampling rate (200 Hz). In our case, window length assumed is 2 min, which corresponds to 24,000 data points. Simply using the time series in the sliding window as the feature vector will lead to the curse of dimensionality. Hence, we compress the time-series data in a given window into a low-dimensional space (which is achieved by extracting features in that window).

Feature extraction is performed by using Tsfresh Python package (Christ et al., 2018). The Tsfresh algorithm characterizes time series with comprehensive and well-established features. Each feature vector is evaluated individually and independently and on its importance for predicting the target class label. For feature selection, we use a series of scalable statistical hypothesis tests available from Tsfresh Python package (Christ et al., 2018). The number of features selected depends on the false discovery rate, which is the $p$-value. The $p$-value selected for our hypothesis testing is equal to 0.05. A small $p$-value ($\leq 0.05$) indicates strong evidence against the null hypothesis that the feature is not relevant and should not

**Preprocessing Seismic Signals**

The undesirable seismic signals (noise) are removed in two stages. In the first stage of the filtering process, the Butterworth high-pass (BH) filter (Smith, 2013) is used to remove the daily trend. This filter removes the signals below a certain frequency using corners, thereby allowing higher frequencies to pass through (Robinson and Treitel, 2000). The second stage of filtering uses an AR model to enhance precursor and eruption signals. The reason to use an AR model for denoising is that an AR model is able to learn the filter directly from the seismic data (Claerbout, 1964; Soubaras, 1994; Naghizadeh and Sacchi, 2011). As the precursor in seismic signal has a lower amplitude compared to the eruption signal, a filtering method without prior information would be ideal for our analysis. We find that a classical time-series forecasting method like an AR model could be used to denoise the seismic signals (Soubaras, 1994; Lesage, 2008; Naghizadeh and Sacchi, 2011; Kang and Harlim, 2012).

In the AR modeling, each time-series data point is regressed on its neighborhood values, called the prediction error filter (PEF). The PEF of the AR model is defined as the BH-filtered seismic time series minus the AR model prediction.
be added. Meaning that, we reject the null hypothesis and the feature should be kept. A large p-value (>0.05) indicates weak evidence against the null hypothesis, so we fail to reject the null hypothesis. More details on the features extracted for ML analysis are provided in Appendix.

ML for Geyser State Classification

RF is a popular ensemble ML method that uses many trees to classify unknown signal data by averaging the prediction of each tree (Breiman, 2001). Many variable unbiased decision trees are learned by RF. RF methods, when trained, maintained, and reinforced properly have great potential in solving real-world problems (Hammer et al., 2012; Dammeier et al., 2016; Hibert et al., 2017; Maggi et al., 2017; Provost et al., 2017; Rouet-Leduc et al., 2017; Rouet-Leduc et al., 2018). They can efficiently process noisy signals through randomization, which is achieved by constructing decision trees in the ensemble using bootstrap samples (Breiman, 1996). RF selects a random subset of predictor variables at each node to grow the decision tree. It is most likely that individual decision trees avoid noise contributing input records (Genuer et al., 2008). This is because decision trees that model noise tend to overfit the training data. As a result, the noise modeling decision trees have high variance and low bias (Genuer et al., 2008). However, RF reduces the variance through majority voting rule, resulting in an ML model that has a low bias and low variance (Breiman, 2001). That is, the overall contribution of noise modeling decision trees on label or class prediction is minimal (Breiman, 2001; Genuer et al., 2008). Hence, RF classification is robust even under noisy environments. Given the above, it is evident that RF methods could be well suited to deal with noisy seismic signals. It should be noted that the variance in RF model (in some sense) captures or describes the generalizability of the RF model. This variance should not be confused with noise in the seismic signals.

For classification, we label the seismic signal based on the ground-truth images. This results in three label classes. Class 1 corresponds to signals that are far away in time from a major eruption event. That is, any data point in the seismic signal that falls beyond 3 min before the eruption event is categorized as class 1, which can be due to geyser recharge or remnant noise. Class 2 describes the signals that are within 1–3 min before the eruption. This label corresponds to precursors to a major eruption event. Class 3 represents the signals that fall within the major eruption event, which is 2 min. In this state, the geyser is degassing and releasing CO2-rich fluids out of the wellbore rapidly. RF classifier constructed on the training dataset is then applied to distinguish these three classes on the test dataset. Out of 18 days, the first 70% constitute the training data. The remaining 30% are the test data. For each day, we have one sample of each class. This corresponds to 13 samples of each class for training RF classifier. For testing RF classification, we have five samples for class 1, five samples for class 2, and six samples for class 3. The size of each sample is 2 min (which is equal to the window length). This time window corresponds to 24,000 data points, which is the dimensionality of a single training/test sample. The Feature Engineering for Geyser State Classification section summarized the features extraction and feature downselection performed in each sample. Algorithm 1 summarizes the proposed ML workflow to denoise the seismic signals, extract seismic time-series features, and classify the state of Chimayó geyser using RF classifier.

RESULTS

In this section, we present preprocessing, feature importance, and RF-based classification results of the proposed ML methodology. Figures 2c and 3a show examples of the recorded signals (vertical component) at RGEYB on 7 July 2017 and 11 July 2017, respectively. The seismic sensor RGEYB provides three components. Seismic components 1, 2, and 3 correspond to vertical motion of the ground, north–south direction, and east–west direction. 18 days of data collected at RGEYB from 1 May 2017 to 1 November 2017 were analyzed. From these figures, it is clear that the amplitude of the eruption (vertical component) is relatively small compared to the anthropogenic noise. Quantitatively, the amplitude around the major eruption event is approximately less than 1% of the peak amplitude of the noise signal. As there are roads, crops, and animal activity around the well, the effect of human noise on the recorded
seismic signals is significant. We perform ML workflow analysis on seismic component 1, which is the vertical motion of the ground. Based on our data analysis, the other two seismic components also showed similar characteristics, and are not shown here for the sake of conciseness. Figure 3b shows an example of the daily trend, which is calculated by averaging the seismic amplitude over all the data points sampled in a given second. The peak in average seismic amplitude is located around 6 a.m. local time in Chimayó, New Mexico, which is the coolest hour of the day. Figure 3c shows the filtered plot after applying BH filter with a corner frequency of 0.1. From this figure, we can see that the temperature effects are removed from the seismic data after high-pass filtering.

After removing the temperature effect, we train and apply AR(p) model to reduce the anthropogenic noise from univariate time-series data. As described in the Preprocessing Seismic Signals section, the first 5% of the 18 days (within the training data) is used to construct the AR(p) model. To construct the AR(p) model, we used statsmodels Python package (Seabold and Perktold, 2010). Other inputs or parameters for AR(p) model are as follows: Ordinary Least Squares (OLS) solver estimates the AR(p) model coefficients with convergence tolerance of $10^{-8}$. AIC selected an optimal lag length $p = 66$ (0.33 s). Figure 3d,e shows the AR(p) model prediction of background noise. We perform one-step prediction of the AR(p) model on the rest of the data. It should be noted that the entire 18 days of seismic data is analyzed. For illustration purposes, we show only certain representative chunks of time series that are mean shifted, stacked, and concatenated together. From these figures, it is clear that the AR(p) model predicts the background noise relatively well when compared with the eruption signals with an $R^2$-score of 0.86 on the remaining data. As a result, PEF subtracts the predictions of AR(p) model from the BH-filtered seismic signal to obtain enhanced seismic data.

Figure 4 shows examples of enhanced signals obtained from the AR(p) filtering process. These results show that

Algorithm 1. Overview of proposed ML methodology to classify geyser state from noisy seismic signals.

1. **Input**: Seismic time-series $X_t$ and eruption event times.
2. **Preprocessing**:
   - Apply BH filter with corner/cut-off frequency of 0.1 on $X_t$ using Obspy Python package. This results in a detrended signal or partially filtered signal $X_t^h$.
   - Fit AR($p$) model on the first 5% of the partially filtered training data using statsmodels Python package.
   - AR($p$) model development does not involve testing data.
   - Make one-step AR($p$) model prediction $\hat{X}_t^h$ on the rest of the data.
   - Calculate PEF signals, which is $X_t^{\text{PEF}} = X_t^h - \hat{X}_t^h$.
3. **Training Random Forests classifiers**:
   - Filtered seismic time-series are divided into training data and test data.
     - $X_t^{\text{PEF}}$ is divided into $X_t^{\text{PEF,train}}$ and $X_t^{\text{PEF,test}}$.
   - Training dataset for RF-classifier: $X_t^{\text{PEF,train}}$.
   - Testing dataset for RF-classifier: $X_t^{\text{PEF,test}}$.
   - Filtered seismic time-series are labeled into three different classes:
     - **Class 1**: System is far away from eruption.
     - **Class 2**: System is close to eruption. This state corresponds to possible precursor in seismic signal.
     - **Class 3**: System is in the stage of eruption.
   - Feature extraction and feature selection on $X_t^{\text{PEF,train}}$ using Tsfresh Python package.
   - RF-classifier construction on down-selected features. ML analysis performed using scikit-learn Python package.
4. **Output**: Classification results on test dataset.
   - Down-selected feature generation on $X_t^{\text{PEF,test}}$ using Tsfresh Python package.
   - RF-classifier accuracy for BH filtered + AR($p$) model filtered seismic test dataset, which is $X_t^{\text{PEF,test}}$.
   - Geyser state is in either **Class 1** or **Class 2** or **Class 3**.
The confusion matrix tells us that the RF classifier is making many correct predictions, which is true in our case. The RF classifier also provides information on the class prediction probabilities for the test data (which is shown in Fig. 4). That is, the predicted state of the geyser is the one with the highest mean probability estimate across the trees in the RF. The predicted geyser state of an input test sample is computed as the mean-predicted class probabilities of the trees in the RF. From Figure 5, RF classifier is clearly able to distinguish between eruption and noneruption signal (which can be either remnant noise or precursor signal) with a probability greater than 90% for most cases. However, the RF method class prediction capability between precursors and noise signal is achieved with a reduced probability (which is around 70%). This is because the amplitude of the remnant noise signal is approximately in the order of the precursory signal. This may cause the RF classifier to misclassify a precursor signal as noise and vice versa.

Table 2 provides a summary of the feature importance (top 10 features) based on RF classifier. Attributes related to fast Fourier transform (FFT) coefficients and aggregated linear trend account for 20% of the feature importance. These features provide information on the energy content of the seismic signal and its variation over different chunk lengths. Chunk length is defined as the size of the chunks to aggregate the signal. It specifies how many seismic time-series values are in each chunk. Data aggregation is any process in which chunks of signals are gathered and expressed in a summary form to extract useful information (such as energy content, its variation, minimum possible energy, etc.) through statistical analysis. This process and the related aggregated features reveal information on how the signal and its energy content changes from remnant noise to precursor to eruption. To calculate aggregated linear trend, first we take the filtered seismic signal and construct a lower sampled version of it by applying the aggregation function (minimum/maximum/variation/mean/median) on consecutive signal chunks. Second, we calculate a linear least-squares regression for values of the filtered signal that were aggregated over chunks against the sequence from 0 up to the number of chunks minus one. Third, we extract attributes. These include \( p \)-value, \( r \)-value, intercept, slope, and standard error of the estimated gradient. Other important features, in addition to aggregated linear trend attributes, include partial autocorrelation lag, FFT coefficients, index mass quantile, time-reversal asymmetry statistic, FFT aggregated, and change quantiles (see the Feature...
et al. amplitude and its burstiness (Goh and Barabási, 2008; Karsai predictions for different stages. The variation in the seismic event classification. It also marks random forest (RF) classifier (b) Example eruption event B: enhanced signals and Figure 4. ▲ Figure 4. Enhanced signals after noise filtering and geyser state classification. Two different geyser eruption examples and noise filtered seismic signals near eruption event are shown. (a) Example eruption event A: enhanced signals and event classification. (b) Example eruption event B: enhanced signals and event classification. It also marks random forest (RF) classifier predictions for different stages. The variation in the seismic amplitude and its burstiness (Goh and Barabási, 2008; Karsai et al., 2012) can be linked to the amount of CO2-rich fluid being released or ejected from the geyser. Higher amplitude may correspond to large amounts of CO2-rich fluid being released. Gradual increase in seismic amplitude can be related to slow release of CO2-rich fluids from the wellbore. From these eruption event examples, it is clear that the seismic amplitude of precursors and eruption signals are considerably different. Eruption event B has higher seismic amplitude than eruption event A. Moreover, eruption signal amplitude in event B has higher burstiness than event A, which may indicate that a lot of CO2 gas as well as CO2-rich fluids are ejected from the geyser in a short span of time. The color version of this figure is available only in the electronic edition.

Table 3 provides a summary of classification accuracy of different ML methods to discern geyser state from noisy data. Three different methods are investigated, which are DTW, RF without PEF, and RF with PEF. To construct RF model, we used scikit-learn Python package (Pedregosa et al., 2011). 13 days (approximately 70%) of the data is used for training and remaining 5 days (approximately 30%) is used for testing. k-nearest neighbors with DTW (Xi et al., 2006) is applied for geyser state classification. The value of k is chosen to be equal to 1 according to the meta analysis in Mitsa (2010). The number of trees in the RF ML model is equal to 100 and bootstrapping is used when building trees. Gini criterion (Breiman, 2001) is used to measure the quality of split for the tree nodes. A minimum number of samples required to be at a leaf node is equal to 1 and all the training samples are given equal weight. Classification accuracy is based on precision, recall, and $F_1$-score over three classes. The precision of the RF classifier for a given geyser state, say class 1, is defined as the proportion of the test samples that RF model predicted to belong to class 1 that actually belong to class 1. The recall for class 1 is the proportion of samples that belong to class 1 that are correctly predicted to belong to class 1 by the RF model. The $F_1$-score provides the weighted average of the precision and recall. Classification accuracy reaches the best and worse when $F_1$-score is equal to 1 and 0.

Table 4 provides the precision, recall, and $F_1$-score averaged over three classes. From these tables, it is evident that DTW with 1-nearest neighbor performs poorly. The most surprising aspect of classification results is that RF without PEF also performs decently. Meaning that RF prediction on the partially filtered seismic data after removing temperature effects has an average $F_1$-score of 0.87. However, RF classification is enhanced significantly after removing the anthropogenic noise (which is by applying AR($p$) model on the temperature filtered seismic data). From these tables, it is clear that the RF classifier with PEF can distinguish the geyser states better (>90%) than RF without PEF. Moreover, the results (see Fig. 3) from the proposed ML approach implies that the time series near the eruption (class 2) are substantially different from the time-series data long before the eruption (class 1). The precursors can be bubbling activity, minor eruption or CO2-rich fluid slowing oozing out of the wellbore or small CO2-bubbles coalescing to form larger bubbles, which was not previously identified. Therefore, the inference from this classification study reinforces the fact that precursory signals (Rouet-Leduc et al., 2017, 2018) exist in the seismic data before the major eruption event takes place. Tables 5 and 6 provide the influence of the size of the training set on the final results of the RF classification. For 70% training and 30% testing, we get a classification accuracy to be around 94%. However, the classification accuracy decreases greatly if we move to 30% training and testing on the remaining 70% data. For this case, the average classification accuracy is around 67%. For 50% training and 50% testing, the average classification accuracy is around 74%.
Figure 5. RF classification probabilities. The examples of class probabilities prediction for testing day (a) one, (b) two, (c) three, and (d) four for remnant noise, precursors, and eruption signals using RF classifier for all the test cases are shown. RF classification probabilities are shown by the bar plots. The predicted class probabilities show that the noise and precursor stages are relatively hard to differentiate compared to the eruption stage. Moreover, this figure illustrates that our task is more than a trivial extension of eruption detection. For most cases, we successfully classified the noise and precursory signals. However, severe weather (such as the New Mexican monsoon) can impact our classification result. For instance, RF classifier provides close predicted class probabilities on test day four for noise and precursor stages. The color version of this figure is available only in the electronic edition.

From these tables, it is clear that classification accuracy decreases greatly if less amount of data is used for training RF classifier. We also performed statistical analysis of RF classifier for different random initializations. Analysis is performed to justify the accuracy of preprocessing step based on AR(p) model. For 10 different random initializations, $F_1$-score of RF classifier with preprocessor based on AR(p) model has a mean of 0.89 and standard deviation of 0.025. Without AR(p) model pre-processing, the $F_1$-score of RF classifier has a mean of 0.76 and standard deviation of 0.047. Based on the mean and standard deviation, we can see that the preprocessing step makes a significant difference on RF classifier prediction. In addition, we also performed a test of significance to show that the PEF improves the RF classifier’s predictions. The hypothesis testing is based on Welch’s $t$-test in which the null hypothesis is that the RF classifier with and without the PEF has the same mean accuracy. The alternate hypothesis is that the RF classifier with the PEF has a higher mean accuracy than the RF classifier without the PEF. Welch’s $t$-test is a two-sample location test which is used to test the hypothesis that the two populations have equal means. The significance level for a given hypothesis test is a value for which a $p$-value less than or equal to 0.05 is considered statistically significant. Using Welch’s $t$-test for the means of two independent test samples of scores with different population variance, we obtain $t$-statistic value 7.5786 and the corresponding $p$-value which is equal to $3.056 \times 10^{-6}$. This rejects the null hypothesis that these two test samples have identical average (expected) values. This means that the improvement of the RF classifier predictions based on the PEF is statistically significant.

**Limitations of Current ML Analysis**

As discussed in previous sections, we rely on RF, a tree-based ensemble learning method for the small number of samples and large number of features problem (Díaz-Uriarte and De Andres, 2006; Genuer et al., 2008). It is natural to ask how far this ability can be taken? There is no clear answer to this question. Some real-world problems such as Adenocarcinoma, Colon, Leukemia, and Prostate have only limited number of samples/data points ($<100$) and thousands of features (see Genuer et al., 2008, their table 2 and Appendix, and Díaz-Uriarte and De Andres, 2006). Even in such high-dimensional problems, the RF method has been successful in classifying small size test samples into different classes.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Normalized Confusion Matrix for Random Forest (RF) Classifier Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td>True Label</td>
<td>Class 1 (noise)</td>
</tr>
<tr>
<td>Class 1 (noise)</td>
<td>1.00</td>
</tr>
<tr>
<td>Class 2 (precursor)</td>
<td>0.20</td>
</tr>
<tr>
<td>Class 3 (eruption)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Information on quality of output of the RF classifier on the test dataset. Normalized by the class support size (number of elements in each class) to give quantitative interpretation on which class is being predicted incorrectly. The number of elements in class 1, class 2, and class 3 are equal to 5, 5, and 6, respectively.
To ensure that our RF classifier estimates are reasonable, we performed ML experiments on the unseen noise data. For this analysis, we selected a 2 min interval from 10 to 30 min before eruption for geyser state classification. A total of 35 test samples were analyzed of which 17 test samples 10 min before eruption and 18 test samples 30 min before eruption. This corresponds to approximately two test samples per day for 18 days of data. These test samples were not used in the training process. The same features as discussed in the Preprocessing Seismic Signals section were extracted in this 2 min time interval for testing the performance of the RF classifier. The results of the RF classification with the PEF are as follows: 

1,1,1,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1

for 17 test samples taken 10 min before eruption and 

2,1,1,1,2,2,1,1,1,1,1,2,1,1,1,2,1,1

for 18 test samples taken 30 min before eruption.

Table 2
Summary of the RF Classification Feature Importances (The Higher, The More Important The Feature)

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Importance</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum aggregate linear trend attribute:</td>
<td>0.0164</td>
<td>Minimum signal amplitude changes across smaller chunks</td>
</tr>
<tr>
<td>Intercept, chunk length = 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. aggregate linear trend attribute: Slope,</td>
<td>0.0165</td>
<td>Variation in energy for bigger chunks</td>
</tr>
<tr>
<td>chunk length = 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum aggregate linear trend attribute:</td>
<td>0.0171</td>
<td>Minimum energy changes in smaller chunks</td>
</tr>
<tr>
<td>Slope, chunk length = 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index mass quantile: 40% of the mass of signal</td>
<td>0.0180</td>
<td>Energy concentration or distribution of signal</td>
</tr>
<tr>
<td>Var. aggregate linear trend attribute: Intercept,</td>
<td>0.0182</td>
<td>Signal variation across bigger chunks</td>
</tr>
<tr>
<td>chunk length = 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum aggregate linear trend attribute:</td>
<td>0.0190</td>
<td>Minimum signal amplitude changes across medium chunks</td>
</tr>
<tr>
<td>Intercept, chunk length = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum aggregate linear trend attribute:</td>
<td>0.0199</td>
<td>Minimum energy changes in medium chunks</td>
</tr>
<tr>
<td>Slope, chunk length = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFT coefficient attribute: Angle</td>
<td>0.0211</td>
<td>Angle corresponding to FFT of seismic signal</td>
</tr>
<tr>
<td>Minimum aggregate linear trend attribute:</td>
<td>0.0234</td>
<td>Minimum energy changes in bigger chunks</td>
</tr>
<tr>
<td>Slope, chunk length = 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. aggregate linear trend attribute: Intercept,</td>
<td>0.0234</td>
<td>Signal variation across smaller chunks</td>
</tr>
<tr>
<td>chunk length = 5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These top 10 features, which account for 20% of the feature importance. FFT, fast Fourier transform, Var., variance.

Table 3
Summary of Classification Accuracy of Different Machine Learning (ML) Methods

<table>
<thead>
<tr>
<th>System/Geyser State</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-Score</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>0.45</td>
<td>1.00</td>
<td>0.62</td>
<td>0.71</td>
<td>1.00</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.50</td>
<td>0.40</td>
<td>0.44</td>
<td>1.00</td>
<td>0.60</td>
<td>0.75</td>
<td>1.00</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td>Class 3</td>
<td>1.00</td>
<td>0.17</td>
<td>0.29</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Analysis is performed using 70% of training data (13 days) and 30% of the test data (remaining 5 days). DTW, dynamic time warping; PEF, prediction error filter.

Table 4
Summary of Average Classification Accuracy of Different ML Methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average Precision</th>
<th>Average Recall</th>
<th>Average F₁-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW</td>
<td>0.67</td>
<td>0.50</td>
<td>0.44</td>
</tr>
<tr>
<td>RF without PEF</td>
<td>0.91</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>RF with PEF</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Analysis is performed using 70% of training data (13 days) and 30% data for testing (remaining 5 days).
taken 30 min before eruption. The results of the RF classifier without the PEF on the unseen noisy dataset are: [1,2,1,1,1,2,1,1,2,1,1,1,1,1,1,1,1,1,1] for 17 test samples taken 10 min before eruption and [2,2,1,1,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1,1] for 18 test samples taken 30 min before eruption. The class label 1 corresponds to noise (class 1) and 2 corresponds to precursor (class 2). To summarize, we have an accuracy of 80% (28 samples out of 35 samples) in classifying noise, which is a reasonable generalization of the RF classifier with the PEF on unseen test data. For the RF classifier without the PEF, we have an accuracy of 74% in classifying noise on the unseen test data. In the original test dataset, the recall for noise for the RF classifier with the PEF was 100% and the only errors were precursors (20%) that were incorrectly predicted to be noise. For this extended noise test set, 20% are incorrectly classified as precursors. This 20% error rate suggests that the original test set was not large enough to capture the 80% recall for class 1 that is observed in this larger test dataset. We acknowledge that an 80% recall for class 1 is lower than the value reported in Table 3. This 80% recall value is probably closer to the true overall class 1 recall for the RF classifier with the PEF.

To conclude, we also acknowledge the limitations of the small sample size. As the size of the dataset is small, we performed bootstrapping analysis to estimate the variance of the RF classifier with and without the PEF. For this bootstrapping analysis, first we selected 39 training samples by sampling from the training set with replacement. The RF classifier with and without PEF is trained on this new resampled training dataset. Then, we selected 16 test samples by sampling from the test dataset with replacement, and used this new test dataset to calculate the trained RF classifiers’ accuracy. We repeated the above process (which is training and testing the RF classifiers) 1000 times to estimate the mean and standard deviation of the RF classifiers’ accuracy with and without the PEF. Mean and standard deviation of the accuracy of the RF classifiers are equal to 0.8 and 0.13 (with the PEF) and 0.76 and 0.12 (without the PEF), respectively. Then, Welch’s $t$-test is performed. The obtained $t$-statistic and $p$-value are equal to 5.4639 and 5.24 x $10^{-8}$. This $p$-value implies that we reject the null hypothesis that the RF classifiers with and without the PEF have the same mean accuracy. The above statistical analysis shows that the RF classifier results are robust with respect to the particular training and test datasets sampled from the population and the improvement resulting from PEF is statistically significant. This statistical significance test demonstrates that the small training and test datasets seems to be adequate to justify the conclusions of the current ML study. Our future work involves collecting more data to overcome this small sample size problem.

**CONCLUDING REMARKS**

In this article, we presented an ML methodology to classify Chimayó geyser state using noisy seismic signals. Our method is based on RFs combined with BH filter and AR($p$) model. First, we provided site description, sensor information, and data sources of the Chimayó geyser. Sensor data analysis and classification of geyser state is focused on the seismic station RGEYB, which is closest to the geyser location. We also described the types of noises present in the seismic data during the data collection process. Second, to achieve enhanced signals, we proposed an approach to preprocess the seismic signals to remove different types of noises. In the first stage of the filtering process, BH filter is used to remove the temperature effects. This filtering method allows higher frequencies to pass through, thereby removing the effects of temperature on the seismic signals. Then, in the second stage of the filtering process, AR($p$) model is used to remove/reduce the background noise. AR($p$) model is trained on this noise using 5% of the seismic data. Then, it is used for predicting background noise on the remaining 95% of the data. The corresponding prediction accuracy of AR($p$) model is greater than 85.

### Table 5

Summary of Classification Accuracy Based on RF with PEF with 100 Trees

<table>
<thead>
<tr>
<th>System/Geyser State</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>0.57</td>
<td>1.00</td>
<td>0.73</td>
</tr>
<tr>
<td>Class 2</td>
<td>1.00</td>
<td>0.33</td>
<td>0.50</td>
</tr>
<tr>
<td>Class 3</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Analysis is performed using 50% of training data (9 days) and 50% data for testing (remaining 9 days).

### Table 6

Summary of Classification Accuracy Based on RF with PEF with 100 Trees

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average Precision</th>
<th>Average Recall</th>
<th>Average $F_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>0.62</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.62</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.80</td>
<td>1.00</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Analysis is performed using 30% of training data (5 days) and 70% data for testing (remaining 13 days).
showed that AR(p) model largely reduces the variance of the anthropogenic noise while still keeping the precursors and eruption event signals. Once these noises (temperature, animal, and human activities) are removed, we use RF to classify the filtered signals. Time-series features are extracted and downselected using statistical relevance tests for RF classification. 13 days of data are used for training and five days of data are used for testing the RF model. The average classification accuracy of RF model on the unseen filtered data is greater than 90%. Moreover, the event discrimination ML methodology reinforces that precursors in seismic signals exist in the data before the major eruption event takes place. This demonstrates the capability of the proposed ML approach to distinguish eruption and precursory signals from background noise. The proposed ML approach is general and is not limited to geysers state classification. It can readily be applied to datasets to detect anomalies or disruptions or noises in the data, thereby amplifying the signals that may constitute a physical phenomenon not detected in past. Finally, the proposed approach streamlines the development of forecasting methods to extract useful and real-time actionable information in noisy environments (Kong et al., 2018; Zhang et al., 2018).

DATA AND RESOURCES

The seismic datasets collected as a part of this project can be obtained from the corresponding author, Maruti K. Mudunuru (maruti@lanl.gov).

ACKNOWLEDGMENTS

The authors thank the support of the Los Alamos National Laboratory (LANL) Directed Research and Development Directed Research Award 20170004DR. The authors would like to extend their thanks to Jim Roberts of Chimayó, New Mexico, for the use of his property to perform the measurements used in this article. B. Y. and Y. J. T. like to thank the support of 2018 LANL Applied Machine Learning Summer School Fellowship. M. K. M. gratefully acknowledges the support of LANL Chick-Keller Postdoctoral Fellowship through Center for Space and Earth Sciences (CSES) and University of California (UC)/LANL Entrepreneurial Postdoctoral Fellowship through Richard P. Feynman Center for Innovation. M. K. M., B. Y., and Y. J. T. thank Youzuo Lin for many useful discussions. M. K. M. also thanks Ting Chen, Molly Cerniecek, and Don Hickmott for their inputs and feedback during the course of the project. The authors also want to acknowledge the comments provided by Karianne Bergen and the two anonymous reviewers that substantially improved the article.

REFERENCES


APPENDIX

The comprehensive features extracted for machine learning (ML) analysis include (Christ et al., 2018):

- Scalar-valued features such as absolute energy, absolute sum of changes, approximate entropy, augmented Dickey–Fuller test, autocorrelation, binned entropy, power measure for time-series nonlinearity, change quantiles, count above mean, first location of maximum and minimum, large standard deviation, last location of minimum and maximum, longest strike above and below mean, maximum Langevin fixed point, maximum, mean, mean absolute change, mean change, mean second derivative, median, minimum, number of continuous wavelet transform (CWT) peaks, quantile, range count, sample entropy, set property, skewness, standard deviation, sum of recuring data points, sum of recuring values, time-reversal asymmetry statistic, value count, kurtosis, and variance.

- Vector-valued features such as aggregated autocorrelation, aggregated linear trends, autoregressive (AR) coefficients, continuous wavelet transform coefficients, energy ratio by chunks, coefficients of fast Fourier transform (FFT), coefficients of AutoRegressive Integrated Moving Average (ARIMA), coefficients of CWT, Friedrich coefficients, index mass quantile, linear trend, partial autocorrelation, power spectral density using Welch’s method, and symmetry looking.

The feature extraction procedure based on Tsfresh algorithm (Christ et al., 2018) provides over 700 time-series features for our case. Feature selection is performed using scalable statistical hypothesis testing. This testing procedure (Christ et al., 2018) reduces the dimension of our time-series data from 700 features to around 100 features. Then, feature importance is performed using random forests (RFs) on the downselected features. Feature importance provides a score that indicates usefulness of each feature in the construction of the decision trees within the RF model. It is calculated for a single decision tree based on the amount that each feature split point improves the
RF model performance measure, weighted by the number of observations the node is responsible for. The performance measure is based on Gini impurity or information gain/entropy (Breiman, 2001). Then, the feature importances are averaged across all of the decision trees within the RF model. The more a feature is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each feature in the time-series dataset, allowing extracted features to be ranked and compared against each other.

A summary of top 10 features in the downselected features, their feature importances, and feature descriptions are provided in Table 2. The Results section describes the details. Other important features in the downselected features include:

- Partial autocorrelation lag provides the value of the partial autocorrelation function at a given lag. The lag $k$ partial autocorrelation of a time-series $\{x_t, t = 1...T\}$ equals the partial correlation of $x_t$ and $x_{t-k}$, adjusted for the intermediate variables $\{x_{t-1}, ..., x_{t-k+1}\}$. It is defined as
  \[ \alpha_k = \frac{\text{Cov}(x_{t-k}, x_{t})}{\sqrt{\text{Var}(x_{t-k})\text{Var}(x_{t})}} \]
  with $1 \leq k \leq n-1$. The resulting FFT coefficients will be complex. The feature outputs include the real part of FFT coefficients, the imaginary part of FFT coefficients, absolute FFT coefficients, and absolute index mass quantile.

- Index mass quantile provides the relative index $i$ in which $q\%$ of the mass of the time-series $x_t$ lie left of $i$. For example, for $q = 50\%$, this feature calculator will return the mass center of the time series.

- Time-reversal asymmetry statistic provides the value of the function
  \[ \frac{1}{n-2\text{lag}} \sum_{i=0}^{n-2\text{lag}} x_{i+\text{lag}}^2 - x_{i+\text{lag}} x_{i+\text{lag}+1} - x_{i+\text{lag}} x_{i+\text{lag}+1}^2, \]
  which is $E[L^2(X)^2 \times L(X) - L(X) \times X^2]$, in which $E$ is the mean and $L$ is the lag operator. The feature outputs the value of the above function for different lags.

- FFT aggregated provides the spectral centroid (mean), variance, skew, and kurtosis of the absolute Fourier transform spectrum. The feature outputs the values of absolute Fourier transform spectrum for the above described inputs.

- Change quantiles provide information on the quantiles change. First, the feature calculator fixes a corridor given by the quantiles $q_i$ and $q_k$ of the distribution of time-series $x_t$. Then, it calculates the average and absolute value of consecutive changes of the time-series $x_t$ inside this corridor, which is the feature output.

More detailed description of the above features can be found in Christ et al. (2018). RF model uses the downselected features instead of each data point in the time window.

Baidhuan Yuan  
Department of Mathematics  
University of California Los Angeles  
520 Portola Plaza  
Los Angeles, California 90095 U.S.A.

Yen Joe Tan  
Department of Earth and Environmental Sciences  
Columbia University  
303F Oceanography, 61 Route 9W - PO Box 1000  
Palisades, New York 10964-8000 U.S.A.

Maruti K. Mudanuru  
Satish Karra  
Computational Earth Science Group  
Earth and Environmental Sciences Division  
Los Alamos National Laboratory  
TA-03, Building 0261, Mailstop T003  
Los Alamos, New Mexico 87545 U.S.A.  
maruti@lanl.gov

Omar E. Marciello  
Andrew A. Delorey  
George D. Guthrie  
Paul A. Johnson  
Geophysics Group  
Earth and Environmental Sciences Division  
Los Alamos National Laboratory  
TA-03, Building 0215, Mailstop D446  
Los Alamos, New Mexico 87545 U.S.A.

Peter M. Roberts  
Geophysics Group  
Earth and Environmental Sciences Division  
Los Alamos National Laboratory  
TA-03, Building 0215, Mailstop D452  
Los Alamos, New Mexico 87545 U.S.A.

Jeremy D. Webster  
Christine N. L. Gammans  
Geophysics Group  
Earth and Environmental Sciences Division  
Los Alamos National Laboratory  
TA-03, Building 0216, Mailstop F665  
Los Alamos, New Mexico 87545 U.S.A.

Published Online 13 February 2019