https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions

Near Real-Time Automated Classification of Seismic Signals of Slope Failures with Continuous Random Forests

Michaela Wenner[1,2], Clément Hibert[3], Lorenz Meier[4], and Fabian Walter[1]

[1]Laboratory of Hydraulics, Hydrology and Glaciology (VAW), ETH Zurich, Zurich, Switzerland
[2]Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland
[3]Université de Strasbourg, CNRS, EOST/IPGS UMR 7516, F-67000 Strasbourg, France
[4]Geopraevent Ltd., Zurich, Switzerland

Correspondence: Michaela Wenner (wenner@vaw.baug.ethz.ch)

Abstract. In mountainous areas, mass movements such as rockfalls, rock avalanches, and debris flows constitute a risk to property and human life. Seismology has evolved into a standard tool to study temporal and spatial variability of mass movements in recent years. Increasing data volumes and the demand for near real-time monitoring call for automated techniques to detect and classify seismic signals generated by such events. Ideally, a large-aperture seismic array recording a significant number of events is available for such applications. This is, however, rarely the case, as a result of cost and time constraints. For most sites, the number of previously recorded slope failures is low, which impedes a reliable application of classification algorithms. Here, we use supervised random forest to classify windowed seismic data on a continuous data stream of a small seismic array, that was installed as a post-event intervention measure after a major rock avalanche. The presented method aims to facilitate data evaluation for stakeholders to detect an increase in slope activity in a near real-time manner. We define three different classes: Noise, slope failures, and earthquakes. Due to the sparsity of slope failures, the training data set is highly imbalanced. We find that several standard techniques to handle such data sets do not increase prediction accuracy. However, a lowering of the prediction threshold for slope failures leads to a prediction accuracy of 80% for slope failures, 90% for earthquakes, and 99% for noise. The classifier is then used to classify 176 days of seismic recordings in 2019 containing four slope failure events. In total, the model classifies eight events as slope failures, of which three are actual slope failures. The other events are very local to regional earthquakes with relatively large magnitudes. One slope failure that has been reported by hikers is not classified as an event. This can be attributed to the small volume of the slope failure and thus low signal to noise ratio. We conclude that the method is suitable for continuous near real-time seismic monitoring.

1 Introduction

High mountain areas are particularly affected by a changing climate. Future projections predict a decrease in glaciated areas as well as the thawing of permafrost (Hock et al., 2019). This has implications on rock wall stability at high altitude and con-

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences
Discussions

sequently, on communities affected by such instabilities (e.g., Allen and Huggel, 2013; Phillips et al., 2017; Coe et al., 2018). The increasing threat to the mountain population, especially in densely populated areas, calls for new monitoring techniques at high temporal resolution and broad spatial coverage to improve predictability, early warning, increase of alarm time, and rapid post-event intervention. Prediction of rockfall events is, due to lack of data and knowledge on relevant processes and triggering mechanisms, still not possible. However, an increase in activity (pre-event acceleration and frequency of small events) is a possible precursor to larger events (Rosser et al., 2007). These events are the most dangerous for mountain communities, and the detection of precursor events is essential. Existing methods to monitor slope failures include point measurements (e.g., extensometers) and large scale monitoring such as terrestrial laser scanners, interferometric radar, and video image recognitions (e.g., Abellán et al., 2011). However, these techniques suffer from disadvantages like high operating costs, limited spacial coverage, and susceptibility to atmospheric conditions.

In the last decade, seismology has evolved into a method to monitor earth surface processes. Knowledge from wave propagation within the earth and generation mechanisms of seismic waves is transferred from its original study objectives, earthquakes, to the so-called field of environmental seismology (e.g., Burtin et al., 2008; Deparis et al., 2008; Helmstetter and Garambois, 2010; Gimbert et al., 2014; Hibert et al., 2014; Dietze et al., 2017; Larose et al., 2015; Allstadt et al., 2018; Lai et al., 2018). Seismic signals generated by mass movements are typically emergent with dominant frequencies of 5 – 10 Hz and few or no distinguishable seismic phases. Signal durations vary between seconds and several tens of seconds, depending on the type of slope failures and slope scales (e.g., Hibert et al., 2011; Vilajosana et al., 2008; Dietze et al., 2017). Rock avalanches usually show a cigar-shaped form similar to tremors observed in volcanic environments but can, depending on the event volume, show much larger signal amplitudes.

Seismometers can record mass movements up to hundreds of kilometers away from the event site (e.g., Allstadt, 2013; Ekström and Stark, 2013; Walter et al., 2020) and allow continuous monitoring of large areas with real-time data transmission. Additionally, the installation of seismometers is relatively low-cost and straightforward. On the other hand, these sensors are sensitive to various sources, i.e., earthquakes, anthropogenic noise, atmospheric signals, runoff, and slope instabilities. Consequently, to distinguish signals of slope failures from other generation mechanisms and to tackle the enormous data volumes of multiple continuously recording stations and large variety of signals, automated techniques to classify the signals are needed, that do not rely on an expert manually browsing through the data.

Hammer et al. (2013) and Dammeier et al. (2016) use a stochastic classifier, hidden Markov models (HMMs), to automatically detect and classify a variety of seismic sources, but focus on a regional scale with larger rockfall volumes detected tens to hundreds of kilometers away from the source. It has been shown, that HMMs successfully classify seismic signals on a continuous data stream. However, to minimize false detections and misclassifications, careful retraining and post-processing steps have been applied. Dammeier et al. (2016) compared the classification output with an earthquake catalog and argued that as an operational system, the on-duty operator could manually inspect the signal and decide if the event is an earthquake or a slope failure.

Another approach that is often used to detect seismic signals in classical seismology is the well established STA/LTA detection algorithm. However, classification of source mechanisms is only possible by looking at different frequency bands. For

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

signals with similar frequency content, amplitudes and signal durations, such as earthquakes and slope failures, a detection of only signals from one source mechanism with STA/LTA is impossible. Additionally, parameter selection is a tedious process that requires detailed knowledge of the data, and seismic signals of slope instabilities are characterized by an emergent onset, which makes detection difficult. For this reason, Helmstetter and Garambois (2010) adapted the algorithm for use in the frequency domain, which has been proven to detect seismic signals without impulsive onsets reliably (e.g., Hibert et al., 2017). Nevertheless, the detector does not allow a distinction between different generation mechanisms. Therefore, Hibert et al. (2017) and Provost et al. (2017) made use of a supervised machine learning algorithm, random forest, to automatically classify the detected local events. High accuracy (99 % and 93 %, respectively) emphasizes the capability of such algorithms to classify seismic signals. This two-step method requires however, an additional step of optimization by choosing the right parameters for the STA/LTA algorithm and the classifier. Additionally, STA/LTA algorithms fail to detect signals which emerge over a timescale larger than the long term average window. Yuan et al. (2019) use random forest to classify seismic signals in several minutes windows of seismic data. The study focuses on several days of data recorded near a geyser to detect pre-eruption seismicity, which is hidden in the noise.

In this paper, we use the Random Forest algorithm (Breiman, 2001), to perform the task of automatic signal classification on continuous data on a local scale, and throughout an extended time period with a large variety of noise signals. Previous local-scale approaches have used networks designed by experts and set-up as an array, ideal for monitoring such processes. However, due to cost and time constraints, this is not possible and not the case for most potential hazard sites. We show that by adjusting our methodology to work with a network of low-cost seismometers with a sub-optimal network configuration, the detection of slope failures is still possible without post-processing. The seismic network we use does neither allow for source location, nor particle motion analysis. Additionally, a small number of recorded slope failures and poor data quality make it difficult to train an accurate model.

The scope of this study is to describe a system which can, despite these difficulties, a) detect an increase in slope activity as an early warning for a plausible larger event in the near future and b) detect rock slope failures that possibly transition into hazardous debris flows early on, to enable down-stream communities to take action. The model is trained on data recorded in 2018 and tested on 2019 data to mimic real-time monitoring and assess its capability to be used as an alarm system for rock slope failures.

2 Study Site and Instrumentation

Pizzo Cengalo locates in Val Bondasca in the Swiss canton of Grisons (Fig. 1a) about 6 km south east of the down-slope village of Bondo at the Italian border. Pizzo Cengalo's slopes have been known to be unstable for several decades with several rock slope failures per year. After a large failure ($V \sim 1.5M\,m^3$) in 2011, systematic monitoring started in 2012 (Baer et al., 2017). In 2017, a large rock avalanche ($V > 3.5M\,m^3$) transitioned into a debris flow destroyed parts of the village of Bondo and killed eight hikers (Walter et al., 2020). An early warning was issued weeks prior to the catastrophic failure, as an acceleration of slope displacement was observed, as well as several smaller failure events prior to the rock avalanche. The large event in 2017

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

prompted an extension of the monitoring system, which included the installation of three seismometers (LERA1 – 3) close to the Bundasca river, the outlet of the catchment, some 3.5 km down-valley of Pizzo Cengalo and 2.5 km up-valley of Bondo.

The one component, short period seismometers (GeoSig 0.9 Hz) were installed along the channel with a mean inter-station distance of about 20 m. An array in such a configuration can for example be used to detect debris flows based on amplitude differences at the stations (Coviello et al., 2019). Since the installation of the seismometers, seismic signals and spectrograms have been made available to stakeholders in hourly time windows in the online portal of Geopraevent, with the goal to catch an increase in rockfall activity by visually evaluating the seismic data. Up until now, this required a daily visual inspection of the seismic signals by employees of the Canton of Grisons, who are not trained seismologists.

3 Methodology

To avoid the time consuming task of manually looking at the seismic signals, we introduce a method that automatically classifies consecutive time windows of seismic signals in a near real-time manner, adressing the challenge of a small number of training sample of the main class of interest. A schematic respresentation of the methodology is shown in Fig. 1b). In this study, we use random forest, a supervised ensemble machine learning algorithm (Breiman, 2001) to classifiy a running window on a continuous data stream on all stations of the network. Random forest is based on the majority vote of several weak decision trees, where each decision tree is built on a random subset of features and training data set. Decision trees consist of nodes, branches and final nodes. On each node, a split based on a threshold on a variable is performed, resulting in one or two branches. This process continues until a classification result is obtained in a final node, a so called leaf. The aggregated decision trees perform a majority vote, where the proportion of trees that voted for one class gives the probability for the class. The time window is then labeled according to the class with the most votes, i.e. the highest probability. We choose random forest, as it is a comprehensive machine learning algorithm that has shown to outperform other algorithms, like support vector machines and boosting ensembles, in a variety of cases (Fernández-Delgado et al., 2014) and already has been succesfully used to classify rock slope failures (Hibert et al., 2017; Maggi et al., 2017; Provost et al., 2017; Malfante et al., 2018; Hibert et al., 2019). Moreover, random forest gives an estimate of the feature importance by measuring the impurity, a measure of how many samples of how many different classes are in one node. The averaged impurity decrease from a feature over all decision trees then gives a ranking of the most discriminating. This allows a more detailed analysis of potential causes for misclassification. For a more detailed description of the algorithm see Breiman (2001). For the implementation of random forest we use "sci-kit learn", a python library for machine learning (Pedregosa et al., 2011).

3.1 Labeled Data Set

We focus on seismic data from the LERA array recorded in 2018 and 2019. To test how the classifier performs in a real-life application we train, validate and test the model on 2018 data, containing five slope failure events, and then use the model to classify 2019 data (four slope failure events). Seismograms and spectrograms of these events are shown in Appendix A1. The seismic signals show typical characteristics of slope-failure events, such as dominant frequencies between 5 - 10 Hz,

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences
Discussions

an emergent onset and a duration of several tens of seconds (e.g., Hibert et al., 2011). Here, we do not investigate source mechanisms and processes of seismogenic mass movements. The recorded signals are weak compared to other studies and thus not well suitable for such an endeavor. We focus solely on the detection and classification of the signals.

As random forest is a supervised machine learning algorithm, a data set of labeled seismic events has to be created to train it. This was done by manually looking at the seismic data recorded by the LERA network and close by stations of the Swiss Seismological Service (SED) (Stations XP.PICE1, CH.VDL, CH.FIESA). Additionally we use a list of observed slope failures made available by the Canton of Grisons and earthquake catalogs from SED and the European-Mediterranean Seismological Centre (EMSC). We decided to use three different classes: noise (NO), slope failures (SF) and earthquakes (EQ). The NO class contains samples of continuous noise as well as noise signals of anthropogenic and athmospheric origin. We use the SF class as an umbrella term for all types of mass movements that might occur (e.g., debris avalanches, rock falls). We consider this assumption to be valid, as seismic source mechanisms of granular flows are similar and generate signals with similar characteristics. We assume that discrepancies between EQ, NO, and SF class are more significant than discrepancies between granular flow generated signals. The EQ class contains a set of local, regional and teleseismic events. Extensive testing has shown, that the lumping of all earthquakes in one class does not negatively affect classifier performance. An example signal of each class is presented in Fig. 3. For the continuous noise we choose random times over the year and include samples from periods with strong precipitation (as an exception from 2019 data) which increased the energy of the background noise. For noise signals, earthquake signals and slope failure signals we manually picked start-time and end-time of the event when the signal exceeds the noise level. The number of events in each class is presented in Fig. 3d. Note that there are only five events that are related to slope failures in 2018, leaving us with a sparse training data set for this class. This issue is addressed further in section 3.3. We divide the catalog with labeled events in a train and test data set, with 70% of all events as training data and 30% as test data. This threshold was chosen in order to have three slope failure events in the training and validation data set and two events in the test data set for a meaningful assessment of the algorithm performance.

3.2 Data Stream Handling

To avoid the extra step of detecting events with a detection algorithm we classify a running window on the continuous data stream with an overlap of $2/3$ of the window length. The large overlap was chosen to avoid missing events on the window margins, but has not been tested for optimal performance. We transform the event catalog with start-times and end-times of all events into a catalog containing the times of all running windows that include an event. In order to increase the number of training samples, we make use of the network configuration at the study site. At the frequency band of interest (1-10 Hz), associated wavelengths are larger than the inter-station distance, resulting in very similar waveforms at all three stations. For earthquakes and slope failures, instead of using the same onset for the sliding windows on all stations, we choose a random onset with a maximum of $2/3$ of the sliding window before the event start-time. This way we catch different windows of the signal and increase the training data set by a factor of three without using the same window several times. For discrete noise signals, as they are often only recorded on one station, we choose sliding windows on the station on which the signal is

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

recorded, again with a random onset up to $2/3$ of the sliding window before the event start-time. For the continuous noise we choose a random station at each time step.

Our approach results in a training data set of 1423, 33 and 201 time windows containing noise, a slope failure signal and an earthquake signal, respectively. Following Provost et al. (2017), we then compute features of these sliding windows. As we do not use the entire waveform of the event, but only the parts that appear in the sliding window, we exclude features that are related to the entire waveform of the signal (e.g. duration and rise time). Additionally, the network configuration does not allow network features (signals are too similar between stations), nor does it allow for polarity features (only vertical component available). We are left with a total number of 55 features including waveform characteristics in the time and frequency domain (see Table A1). Before calculating the feautres, we apply a 4 corners butterworth bandpass filter (1 – 10 Hz). For the feature generation after Provost et al. (2017), we choose frequency bands of 1 – 3 Hz, 3 – 6 Hz, 5 – 7 Hz, 6 – 9 Hz and 8 – 10 Hz.

3.3 Imbalanced Data Set

The limited amount of data and more specifically the small number of SF events that happened in 2018 at Pizzo Cengalo lead to an imbalanced data set. As shown in Fig. 3d, the number of training events is highly disproportional. This imposes a problem for machine learning algorithms, as they generally try to optimize the score, i.e., the number of correctly labeled classes. In a highly imbalanced data set, the detectiong algorithm may be less sensitive to the minority class, as it does not drastically impair the score if it is labeled incorrectly. For our data set, with the events that we are most interested in being the minority class, it is important to address this problem. There are several possibilities to handle imbalanced data sets, either based on manipulation of the trainings data set or on changes within the algorithm. The simplest approaches are random undersampling (US) and naive oversampling (OS) of the training data. For random undersampling, only a random subset of training data of the majority class is chosen. This way, the data set becomes more balanced by reducing the samples in the majority classes. However, this might mean that important characteristics of the majority class are not captured. In contrast to undersampling, naive oversampling randomly multiplies samples in the minority class, increasing the risk of overfitting, the lack of generalization, within the minority class.

A more sophisticated way of increasing training samples in the minority class is synthetic minority over-sampling (SMOTE) (Chawla et al., 2002). SMOTE is based on the idea of creating new training samples in the minority class, by interpolating in the feature space between a sample and a random set of its k-nearest neighbors (Fig. 2c)). Therefore, a new sample is generated with similar features as already existing samples. This increses the sample size of the minortiy class but minimizes the problem of overfitting. On the algorithm level, random forest opens two possibilities for imbalanced data: setting a class weight on the minority class or undersample the training data for each single tree, a so called balanced random forest (BRF) (Lemaître et al., 2017). The presented results for BRF use both class weights and undersampling.

3.4 Training Process and Evaluation

Typically, the performance of a machine learning algorithm is evaluated using some kind of score, i.e. the number of correctly labeled samples. However, for an imbalanced data set, the overall score can be misleading, as it is only weakly sensitive to

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

misclassification of the minority class. To avoid this, we use two different metrics to evaluate model accuracy, the confusion matrix and receiver operating characteristic (ROC) curves.

A confusion matrix consists of the true label of the samples of each class as rows and the classifier predicted label as columns (Fig. 2a). For a perfect classifier all samples are located on the diagonal of the matrix. Using the confusion matrix, the classifier can be evaluated for each class separately. The ROC curve uses the true positive rate (TPR) and false positive rate (FPR) for different probability thresholds Fawcett (2006). TPR is defined as the number of true positives divided by the sum of true positives (TP) and false negatives (FN) (TPR = TP/(TP + FN)). FPR is defined as the number of false positives (FP) divided by the sum of false positives and true negatives (TN) (FPR = FP/(FP + TN)). Class prediction of random forest is based on the score of a class, i.e. its probability defined by the number of predictions out of all trees. By lowering the threshold for classification, i.e. the probability threshold for a class to be predicted, FPR and TPR increase as FN samples transition to TP and TN samples transition to FP. As an example, we consider an imaginary two class problem with a decision threshold of probability > 0.5 for the "positive" class (Fig. 2a) leading to a TP = 1, FP = 2 , TN = 3 and FN = 4 and resulting TPR = $1/5$ and FPR = $2/5$. When lowering the probability threshold for the "positive" class to be predicted to lets say 0.2, TP will increase, but so will FP, giving TP = 4, FP = 3, TN, = 2, and FN = 1. This results in larger values of TPR and FPR (TPR = $4/5$, FPR = $3/5$). When plotting FPR against TPR for each probability threshold one obtains the ROC curve with a monotonous increase. For a schematic drawing, see Fig. 2b). The best case scenario is a TPR of one and a FPR of zero (0,1) whereas a random classifier would result in a diagonal from (0,0) to (1,1). The area under the curve (AUC) can be used as a one value metric for model performance. The larger the area under the ROC curve, the better the model accuracy.

Both metrics, confusion matrix and ROC/AUC, can directly be transferred into a multiclass environment. For the confusion matrix, this results simply in several columns and rows. The ROC curve can be computed for each class separately, by bundling all other classes together. For a more representative measure, we use k-fold cross validation when computing the ROC curves (e.g., Stone, 1974). As our training data set only contains three events in the minority class, we use 3-fold cross validation, with random 2/3 of each class in the training data set and 1/3 used for validation. This way, we obtain three ROC curves and AUC values per class trained and tested on three different random subsets. We then take the mean TPR and FPR to plot the ROC curves. We computed the 95% confidence level using Student's t-distribution for small sample sizes (n=3).

4 Results

4.1 ROC Analysis

We computed ROC curves and AUC values for different window sizes and different under and oversampling techniques. Figure 4a shows the AUC values for the SF class plotted in a heatmap. Rows show different window sizes (10s - 60s) and columns different techniques. Associated standard deviations are shown in Fig. 4b. Darker colors mark a larger AUC value and a smaller standard deviation, respectively. Although almost all values lie within the confidence intervalls, overall it can be observed that the smallest window size and the largest window size give slightly better values with a small standard deviation. Additionally, random forest and balanced random forest (BRF) perform better than under and over sampling techniques.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

On a continuous data stream a 10 seconds window with 6 seconds of overlap does not leave enough time to compute features and classify the event in real-time. The 60 seconds window, on the other hand, results in a classification delay of one minute and we assume that chances are higher to miss smaller events which are masked by a large amount of noise in the 60 seconds window. As a compromise with equally large AUC value and small standard deviation we choose 40 second windows and classical random forest. The ROC curve of this configuration is shown in Fig. 4c. To make sure we catch most of the windows containing a slope failure signal, we set the target TPR to > 0.9. From the 3-fold cross validation ROC analysis, we obtain a mean probability threshold of 0.23 for a TPR > 0.9 in the SF class. As a next step, the optimal model parameters (i.e., number of decision trees, number of features chosen for each tree, maximum tree depth, ...) for the 40 seconds window size and random forest are chosen using the randomized cross validation search of scikit-learn (Pedregosa et al., 2011). We use the obtained classifier to classify the test data set from the 2018 data, containing two RF events, which was not used for the model set up and is therefore an unbiased evaluation of the classifier.

In contrast to the training data set, the test data set contains time windows of all three stations which begin (and end) at the same time for each station. This way, the model classifies the time window for each station separately and we then perform a majority vote over the three stations. We set the probability threshold for the SF class to 0.23 as obtained from the ROC curve. Consequently, for probabilities higher than 0.23 for the SF class, the window will be classified as slope failures, even if another class has a higher probability. In case of every seismic station classifying the same time window into a different class, the time window will be labeled as noise.

4.2 Classifier Performance on Test Data

We end up with a label for all 40 seconds time windows, which were labeled on each station using random forest and a majority vote over all stations gives the final label. The results of this setup are shown in Fig. 5. The normalized confusion matrix (Fig. 5a) shows a misclassification of 20% for slope failures. Additionally 10% of earthquakes are classified as noise. The misclassification rate of noise is however very small (1%). The most discriminating features are presented in Fig. 5b). The colors denote spectral and waveform features. Distinctive features are spectral gyration radius (gamma2), spectral centroid (gamma1), central frequency of the first quartile (Fquart1), variance of the normalized fast Fourier transform (FFT) (VarFFT), frequency at the maximum of the FFT (FmaxFFT), frequency at spectrum centroid (FCentroid), energy of the last 2/3 of the autocorrelation function (INT2), and the energy of the seismic signal in the frequency band of 1-3 Hz (ES[0]) (Provost et al., 2017). Figure 5c) shows the four most distinguishing features plotted against each other, with the diagonal showing the uni-variate distribution of the feature. Figure 5b) shows that the by far most discriminating features are characteristics in the frequency domain. This is consistent with the fact that the windowing eliminates information from the entire waveform, amplitudes of signals strongly depend on emitted seismic energy and source receiver distance and the commonly observed differences in frequency patterns of noise signals, continuous seismic noise and other events (see Fig. 3). Figure 5c) shows however, that there is a large overlap between the classes, even for the most discriminating features, which highlights the necessity of a large number of features to distinguish the event type.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

4.3 Classifier Implementation

As a next step, the model is used to classify data from 2019 imitating conditions of a near real-time classification. We first compute the signal features of 40 second time windows with an overlap of 26 seconds for each station and perform a classification. Next, a majority vote of the stations is performed and a label is assigned to the time window. We compare the results to an event catalog compiled from hiker reports, manual classification of seismic data, and image and radar interferogram correlations.

In 176 days in 2019 (Julian Day 94 to 270), 20 days have at least one window which was classified as slope failure. To exclude misclassified windows because they only contain a small portion of a signal, we set a minimum threshold of three consecutive SF classifications. With this threshold, we limit the number of slope failure detections to eight. Out of these eight, three are actual slope failures, two of which happened on April 4th and 26th and one on July 16. Seismic waveforms and associated classifications for July 16 are shown in Fig. 6, with a zoom on a rockfall event (first zoom) and a noise signal (second zoom). Four events that were classified as slope failures are earthquakes on May 22, July 29, August 8 and August 29. Two out of these earthquakes originate from the German lakeside of Lake Constance about 160 km north-west of Pizzo Cengalo, with a distance between the epicenters of about 3 km and magnitudes of 3.6 and 3.4 (EMSC catalog). The two other earthquakes are Magnitude 3.3 and 2.2 earthquakes with epicenters about 170 km south and 33 km west of Pizzo Cengalo, respectively (EMSC catalog). The last event that was classified as slope failure on August 13 is characterized by a duration of 10 s and is not listed in any earthquake catalog. Presumably this event is a small, very local earthquake. A slope failure that was observed by hikers on August 14 was classified as noise. Often events create dust clouds which are easily noticeable despite a small mass-movement volume. The misclassification likely results from a low signal to noise ratio (SNR), hence a probably relative small volume of the event, as all windows containing the event were classified as noise.

5 Discussion

5.1 Data volume and network set-up

The LERA network was set-up in the aftermath of a large rock avalanche event in 2017. Since then, slope activity has strongly decreased. Nevertheless, it is crucial to monitor the site. However, the decrease in activity implies automatic detection and classification is based on a small number of training events in the slope failure class and a comparably large number of events in other classes.

To address the problem of an imbalanced data set and resulting misleading scores, we test several techniques to handle such data sets and use receiver operating characteristic curves for performance assessment. The area under the curve for SF is largest for a generic random forest. Further assessment shows that SF's true positives are largest when using a technique to handle imbalanced data sets, but leading to a tremendous increase in events from EQ being classified as SF. Using a generic random forest, SF is underrepresented, ending up with zero true positives but also zero false positives. By lowering the probability threshold for SF, the true positive rate increases, whereas the false positive rate stays low. Therefore, for this dataset, we decided to ignore problems with imbalanced data sets and solve misclassifications by lowering the probability threshold. It

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

remains to be seen if this approach works best for other data sets, but in our case it gives the best results by maximizing the number of true positives in SF and minimizing the number of EQ classified as SF. Generally, the problem of an imbalanced data set can be tackled by increasing the amount of training data in the minority class. A classifier trained for an area that is more active or has been monitored during a longer period is expected to give better results with higher accuracy. Additionally, the small number of events in the slope failure class can lead to overfitting, i.e., an insufficient generalization of the model. Hence, small deviations in signal characteristics can lead to misclassification and undetected slope failures.

The LERA network has an unfavorable aperture to detect and classify seismic signals, as the stations are set-up in a line and only tens of meters apart. This prohibits the usage of network characteristics, as arrival time differences and amplitude ratios. Feature importance analysis shows that the classifier predominantly uses spectral features to distinguish between different classes. Provost et al. (2017) found that several waveform features, e.g., duration and the ratio between ascending and descending time of the signal, are powerful distinctive features of slope failures and earthquakes. These are however, characteristics of the entire waveform of an event. In our case, the use of constant window size with start and end regardless of event start and end sacrifices this information and the classifier therefore relies on spectral features. The spectral content of these earthquakes and slope failures at our site is highly similar, which complicates a correct classification (see Fig. 3). Using the continuous classifier, however, we eliminate high numbers of false detections (Dammeier et al., 2016), reduce parameter selection effort, and eventually create a more transparent detection and classification system.

Continuous random forest correctly classifies events in the test data with a high signal to noise ratio. The visually observed slope failure that was classified as noise by the model barely exceeded the background noise. For most waveform and spectral features, especially the most discriminating ones (feature importance analysis), values of time windows containing the slope failure signal do not differ from that of windows that contain only noise. This suggests that characteristic features of windows containing signals with low signal to noise ratio are dominated by noise and are therefore misclassified. Additionally, misclassifications of local to regional earthquakes as slope failures reduce the accuracy of the classifier. Figure 3 a) and b) shows an example signal of an earthquake and a slope failure and clearly illustrates the similarity of the two signals. A more extensive training data set could potentially resolve such confusion.

5.2 Classification of continuous data stream vs. classification of detected events

Several studies have shown, that classification algorithms accurately classify events detected with the STA/LTA approach (e.g., Hibert et al., 2017; Provost et al., 2017). To quantify to performance of the algorithm presented here, we benchmark the continuous approach against a two-step approach with an STA/LTA detection. For this, we train an RF classifier on the manually picked labeled data set from 2018 and run an STA/LTA algorithm over the 2019 data and classify the detected events. After extensive testing, we define the parameters that provide accurate detection for our data set as an STA window length of 1s and an LTA window length of 18s. The detector turns on when the STA/LTA ratio exceeds four and turns off when the STA/LTA ratio becomes lower than two. Additionally, we use a coincidence trigger, with a threshold of three, which means that the STA/LTA threshold needs to be exceeded at all three stations.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences Discussions

The continuous approach correctly classifies three slope failures (TP), misses one slope failure (FN), and classifies four earthquakes as slope failures (FP). The two-step approach of STA/LTA detection correctly classifies two slope failures (TP), misses two slope failures (FN), and classifies six earthquakes as slope failures (FP). For a simple comparison, we can use the critical success index (CSI = TP / (TP + FN + FP)) which ignores all non events (TN). For the continuous approach, we obtain a CSI of 0.375, whereas, for the STA/LTA approach, we obtain a CSI of 0.2. This indicates that our continuous approach performs slightly better. However, the small number of events prohibits a statement on robustness. Interestingly, there is a large overlap in the earthquakes being misclassified as slope failures between the two approaches.

Even though not a focus of this study, we note that STA/LTA detection algorithms tuned to detect short signals (several tens of seconds) miss events of long duration and gradual amplitude increase, such as debris flows, volcanic tremors, lahars, and glacier lake outburst floods. Coviello et al. (2019) show that with a window size of 10 s and 100 s for STA and LTA respectively, debris flows can be detected, excluding other events as earthquakes. However, this also excludes the detection of short slope-failure signals. The continuous approach is capable of detecting such events and is therefore applicable in multiple contexts and different sites. For example, intense precipitation raises the noise level by an increase in runoff and, consequently, seismogenic sediment transport (Tsai et al., 2012; Burtin et al., 2016). Similarly, snow cover and strong temperature fluctuations can affect the instrument itself and change the noise level. Preliminary implementation of a fourth class called runoff with two days of increased water discharge (measured with gauges) found two more days of peak discharge. Using the two step-method of STA/LTA, requires a second STA/LTA algorithm with its own parameters to detect these signals. Consequently, applying continuous random forest in different circumstances is potentially a low effort, as there is no need to fine-tune the detection algorithm.

6 Conclusions

We designed a classifier testing a continuous implementation of random forest for automated classification of slope failure seismic signals in near-real-time. The resulting detector could be used as a tool to alert stakeholders in case of slope failure activity and therefore help mitigate damages to property and human lives.

We show that a near-real-time classification of seismogenic slope failures is feasible. As data collection increases continuously, approaches to filter for rare occurrence events gain in importance. Our approach enables us to detect the occurrence of rare events of high interest in a large data set of more than a million windowed seismic signals. Nevertheless, misclassification is a challenge that an imbalanced training data set enhances.

An added value is gained from a time consumption point of view: Manually going over the seismic data is a tedious task, and especially for non-experts uncertainty and misclassification rates can be high. An automatic classifier can run in the background on a standard machine (in our case 2017, Intel Core i7) and alert stakeholders in case of an event classified as slope failure. A subsequent manual inspection is still advisable, but significantly less time-intensive then continuously monitoring the signals. Lastly, using a continuous implementation of random forest with the features we chose paves the way for the deployment of such a system with semi-supervised and unsupervised algorithms. Both semi-supervised and unsupervised frameworks exist

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions

for random forest. In contrast to supervised machine learning algorithms, such algorithms find clusters of similar patterns in a data set and, therefore, detect unseen patterns.

Code and data availability. Data supporting this research are available in Geopravent (2017) and are not accessible to the public or research community. To gain access contact Lorenz Meier. Computed feature files and code used in this study are available on https://github.com/michaelawenner/Automatic_classification_Bondo.

Appendix A: Slope Failure Signals and List of Computed Features

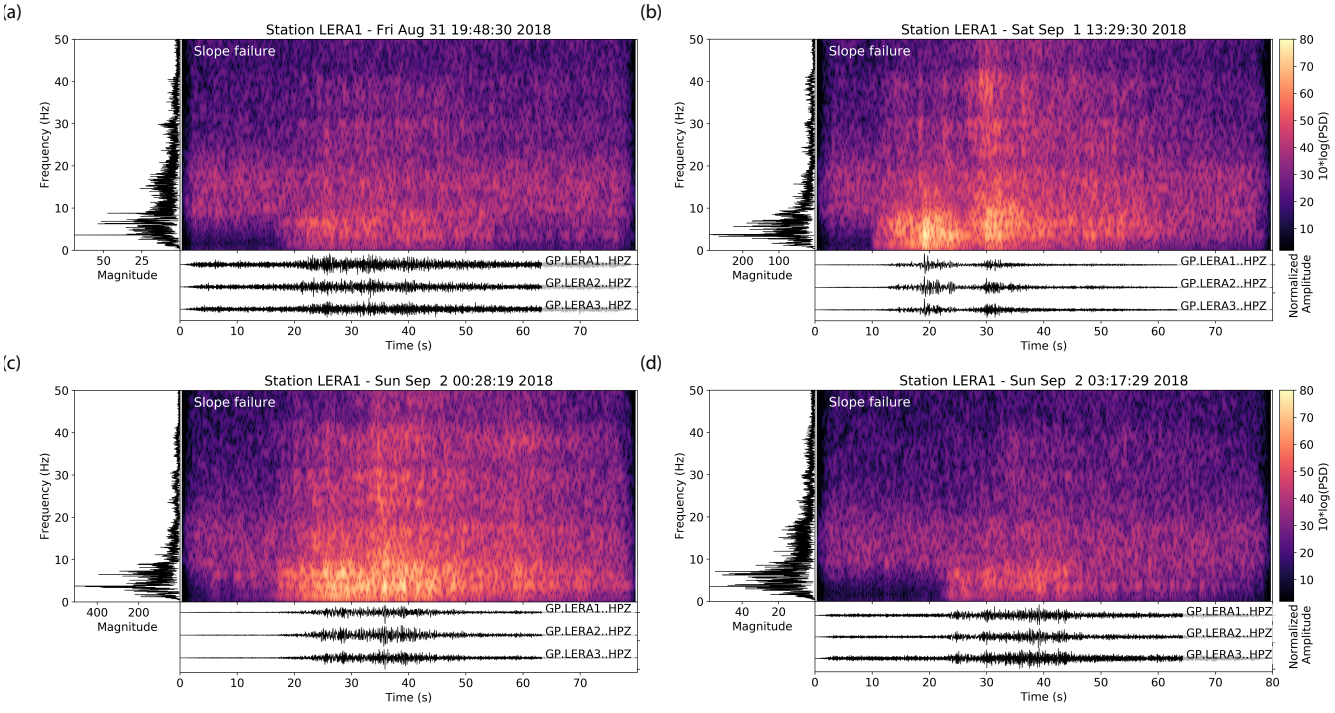

Figure A1. Seismic signals, spectrograms and spectra of four additional slope failure events in 2018 used for training.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020


Natural Hazards and Earth System Sciences
Discussions

Author contributions. LM with Geopraevent installed the three seismometers and assured data transmission. CH provided the feature computation code. MW processed and analyzed the data with the help of CH and FW. MW prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The project is funded by WSL's strategic initiative Climate Change Impacts on Alpine Mass Movements (CCAMM). FW's salary was funded by the Swiss National Science Foundation (GlaHMSeis Project PP00P2_157551 and PP00P2_183719). The Canton of Grisons partly payed MW's salary. We thank Geopraevent Ltd., who installed and maintain the seismic stations, provided the data, an online data portal and implemented the algorithm to run in real-time. We are thankful for constructive feedback from Velio Coviello and an anonymous reviewer on an earlier version of the manuscript. We also want to acknowledge the Obspy developer team Beyreuther et al. (2010); Krischer et al. (2015) for providing an easy to use framework for seismic data handling.

Table A1. Table with all features, slightly adjusted from Provost et al. (2017)

	Waveform Features:	
1	Ratio of the mean over the maximum of the envelope	–
2	Ratio of the median over the maximum of the envelope	–
3	Kurtosis of the raw signal (peakness of the signal)	$\frac{m_4}{\sigma^4}$, with m_4: fourth moment, σ: standard deviation
4	Kurtosis of the envelope	see 3
5	Skewness of the raw signal	$\frac{m_3}{\sigma^3}$, with m_3: third moment
6	Skewness of the envelope	see 5
7	Number of peaks in the autocorrelation function	–
8	Energy in the first third part of the autocorrelation function	$\int^{T/3} \mathrm{C}(\tau)d\tau$, with T: signal duration, C: autocorrelation function
9	Energy in the remaining part of the autocorrelation function	see 8
10	Ratio of 8 and 9	–
11 – 15	Energy of the signal filtered in 1 — 3 Hz, 3 — 6 Hz, 5 — 7 Hz, 6 — 9 Hz and 8 — 10 Hz	$\int_0^T y_f(t)dt$, with y_f: filtered signal in the frequency range [f1-f2]
16 – 20	Kurtosis of the signal in 1 — 3 Hz, 3 — 6 Hz, 5 — 7 Hz, 6 — 9 Hz and 8 — 10 Hz frequency range	see 3
21	Maximum of the envelope	–
	Spectral Features:	
22	Mean of the DFT	DFT: discrete Fourier transform
23	Max of the DFT	–
24	Frequency at the maximum	–
25	Frequency of spectrum centroid	–
26	Central frequency of the 1st quartile	–
27	Central frequency of the 2nd quartile	–
28	Median of the normalized DFT	–
29	Variance of the normalized DFT	–
30	Number of peaks (> 0.75 DFTmax)	$\mathrm{DFT_max}$: maximum of the DFT
31	Mean value for the peaks	–
32 -- 35	Energy in [0,$\frac{1}{4}$]Nyf, [$\frac{1}{4}$,$\frac{1}{2}$]Nyf, [$\frac{1}{2}$,$\frac{3}{4}$]Nyf, [$\frac{3}{4}$,1]Nyf	$\int_{f_1}^{f_2} \mathrm{DFT}(f)df$ with f_1, f_2: the considered frequency range
36	Spectral centroid	$\gamma_1 = \frac{m_2}{m_1}$, with m_1 and m_2 being the first and second moment
37	Gyration radius	$\gamma_3 = \sqrt{\frac{m_3}{m_2}}$, with m_3 being the third moment
38	Spectral centroid width	$\sqrt{\gamma_1^2 - \gamma_2^2}$

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

	Spectrogram Features:	
39	Kurtosis of the maximum of all discrete Fourier transforms (DFTs) Kurtosis as a function of time t	Kurtosis $\left[\max_{t=0,..,T}(\mathrm{SPEC(t,f)})\right]$ with SPEC(t,f): spectrogram
40	Kurtosis of the maximum of all DFTs as a function of time t	see 39
41	Mean ratio between the maximum and the mean of all DFTs	mean $\left(\frac{\max(\mathrm{SPEC})}{\mathrm{mean(SPEC)}}\right)$
42	Mean ratio between the maximum and the median of all DFTs	see 41
43	Number of peaks in the curve showing the temporal evolution of the DFTs maximum	–
44	Number of peaks in the curve showing the temporal evolution of the DFTs mean	–
45	Number of peaks in the curve showing the temporal evolution of the DFTs median	–
46	Ratio between 43 and 44	–
47	Ratio between 43 and 45	–
48	Number of peaks in the curve of the temporal evolution of the DFTs central frequency	–
49	Number of peaks in the curve of the temporal evolution of the DFTs maximum frequency	–
50	Ratio between 48 and 49	–
51	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and mean frequency	–
52	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and median frequency	–
53	Mean distance between the 1st quartile and the median of all DFTs as a function of time	–
54	Mean distance between the 3rd quartile and the median of all DFTs as a function of time	–
55	Mean distance between the 3rd quartile and the 1st quartile of all DFTs as a function of time	–

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

References

Abellán, A., Vilaplana, J. M., Calvet, J., García-Sellés, D., and Asensio, E.: Rockfall monitoring by Terrestrial Laser Scanning - Case study of the basaltic rock face at Castellfollit de la Roca (Catalonia, Spain), Natural Hazards and Earth System Science, https://doi.org/10.5194/nhess-11-829-2011, 2011.

Allen, S. and Huggel, C.: Extremely warm temperatures as a potential cause of recent high mountain rockfall, Global and Planetary Change, https://doi.org/10.1016/j.gloplacha.2013.04.007, 2013.

Allstadt, K.: Extracting source characteristics and dynamics of the August 2010 Mount Meager landslide from broadband seismograms, Journal of Geophysical Research: Earth Surface, 118, 1472–1490, https://doi.org/10.1002/jgrf.20110, 2013.

Allstadt, K. E., Matoza, R. S., Lockhart, A. B., Moran, S. C., Caplan-Auerbach, J., Haney, M. M., Thelen, W. A., and Malone, S. D.: Seismic and acoustic signatures of surficial mass movements at volcanoes, https://doi.org/10.1016/j.jvolgeores.2018.09.007, 2018.

Baer, P., Huggel, C., McArdell, B. W., and Frank, F.: Changing debris flow activity after sudden sediment input: a case study from the Swiss Alps, Geology Today, https://doi.org/10.1111/gto.12211, 2017.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J.: ObsPy: A python toolbox for seismology, Seismological Research Letters, https://doi.org/10.1785/gssrl.81.3.530, 2010.

Breiman, L.: Random forests, Machine Learning, https://doi.org/10.1023/A:1010933404324, 2001.

Burtin, A., Bollinger, L., Vergne, J., Cattin, R., and Nábělek, J. L.: Spectral analysis of seismic noise induced by rivers: A new tool to monitor spatiotemporal changes in stream hydrodynamics, Journal of Geophysical Research: Solid Earth, https://doi.org/10.1029/2007JB005034, 2008.

Burtin, A., Hovius, N., and Turowski, J. M.: Seismic monitoring of torrential and fluvial processes, https://doi.org/10.5194/esurf-4-285-2016, 2016.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, https://doi.org/10.1613/jair.953, 2002.

Coe, J. A., Bessette-Kirton, E. K., and Geertsema, M.: Increasing rock-avalanche size and mobility in Glacier Bay National Park and Preserve, Alaska detected from 1984 to 2016 Landsat imagery, Landslides, https://doi.org/10.1007/s10346-017-0879-7, 2018.

Coviello, V., Arattano, M., Comiti, F., Macconi, P., and Marchi, L.: Seismic Characterization of Debris Flows: Insights into Energy Radiation and Implications for Warning, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1029/2018JF004683, 2019.

Dammeier, F., Moore, J. R., Hammer, C., Haslinger, F., and Loew, S.: Automatic detection of alpine rockslides in continuous seismic data using hidden Markov models, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1002/2015JF003647, 2016.

Deparis, J., Jongmans, D., Cotton, F., Baillet, L., Thouvenot, F., and Hantz, D.: Analysis of rock-fall and rock-fall avalanche seismograms in the French Alps, https://doi.org/10.1785/0120070082, 2008.

Dietze, M., Turowski, J. M., Cook, K. L., and Hovius, N.: Spatiotemporal patterns, triggers and anatomies of seismically detected rockfalls, Earth Surface Dynamics, https://doi.org/10.5194/esurf-5-757-2017, 2017.

Ekström, G. and Stark, C. P.: Simple scaling of catastrophic landslide dynamics, Science, 339, 1416–1419, https://doi.org/10.1126/science.1232887, 2013.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems?, Journal of Machine Learning Research, https://doi.org/10.1117/1.JRS.11.015020, 2014.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Geopravent: Bondo, Val Bondasca, Seismic, https://data.geopraevent.ch/index.php, updated hourly, 2017.

Gimbert, F., Tsai, V. C., and Lamb, M. P.: A physicalmodel for seismic noise generation by turbulent flow in rivers, Journal of Geophysical Research F: Earth Surface, https://doi.org/10.1002/2014JF003201, 2014.

Hammer, C., Ohrnberger, M., and Fäh, D.: Classifying seismic waveforms from scratch: A case study in the alpine environment, Geophysical Journal International, https://doi.org/10.1093/gji/ggs036, 2013.

Helmstetter, A. and Garambois, S.: Seismic monitoring of Schilienne rockslide (French Alps): Analysis of seismic signals and their correlation with rainfalls, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1029/2009JF001532, 2010.

Hibert, C., Mangeney, A., Grandjean, G., and Shapiro, N. M.: Slope instabilities in Dolomieu crater, Réunion Island: From seismic signals to rockfall characteristics, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1029/2011JF002038, 2011.

Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N. M., Satriano, C., Maggi, A., Boissier, P., Ferrazzini, V., and Crawford, W.: Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1002/2013JF002970, 2014.

Hibert, C., Provost, F., Malet, J. P., Maggi, A., Stumpf, A., and Ferrazzini, V.: Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm, Journal of Volcanology and Geothermal Research, https://doi.org/10.1016/j.jvolgeores.2017.04.015, 2017.

Hibert, C., Michéa, D., Provost, F., Malet, J. P., and Geertsema, M.: Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska, Geophysical Journal International, https://doi.org/10.1093/gji/ggz354, 2019.

Hock, R., Rasul, G., Adler, C., Cáceres, B., Gruber, S., Hirabayashi, Y., Jackson, M., Kääb, A., Kang, S., Kutuzov, S., Milner, A., Molau, U., Morin, S., Orlove, B., and Steltzer, H. I.: Chapter 2: High Mountain Areas, in: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate, 2019.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J.: ObsPy: A bridge for seismology into the scientific Python ecosystem, Computational Science and Discovery, https://doi.org/10.1088/1749-4699/8/1/014003, 2015.

Lai, V. H., Tsai, V. C., Lamb, M. P., Ulizio, T. P., and Beer, A. R.: The Seismic Signature of Debris Flows: Flow Mechanics and Early Warning at Montecito, California, Geophysical Research Letters, https://doi.org/10.1029/2018GL077683, 2018.

Larose, E., Carrière, S., Voisin, C., Bottelin, P., Baillet, L., Guéguen, P., Walter, F., Jongmans, D., Guillier, B., Garambois, S., Gimbert, F., and Massey, C.: Environmental seismology: What can we learn on earth surface processes with ambient noise?, https://doi.org/10.1016/j.jappgeo.2015.02.001, 2015.

Lemaître, G., Nogueira, F., and Aridas, C. K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research, 2017.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., and Amemoutou, A.: Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano, Seismological Research Letters, https://doi.org/10.1785/0220160189, 2017.

Malfante, M., Dalla Mura, M., Metaxian, J. P., Mars, J. I., Macedo, O., and Inza, A.: Machine Learning for Volcano-Seismic Signals: Challenges and Perspectives, IEEE Signal Processing Magazine, https://doi.org/10.1109/MSP.2017.2779166, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 2011.

Phillips, M., Wolter, A., Lüthi, R., Amann, F., Kenner, R., and Bühler, Y.: Rock slope failure in a recently deglaciated permafrost rock wall at Piz Kesch (Eastern Swiss Alps), February 2014, Earth Surface Processes and Landforms, https://doi.org/10.1002/esp.3992, 2017.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Provost, F., Hibert, C., and Malet, J. P.: Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier, Geophysical Research Letters, https://doi.org/10.1002/2016GL070709, 2017.

Rosser, N., Lim, M., Petley, D., Dunning, S., and Allison, R.: Patterns of precursory rockfall prior to slope failure, Journal of Geophysical Research: Earth Surface, https://doi.org/10.1029/2006JF000642, 2007.

Stone, M.: Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B (Methodological), https://doi.org/10.1111/j.2517-6161.1974.tb00994.x, 1974.

Tsai, V. C., Minchew, B., Lamb, M. P., and Ampuero, J. P.: A physical model for seismic noise generation from sediment transport in rivers, Geophysical Research Letters, https://doi.org/10.1029/2011GL050255, 2012.

Vilajosana, I., Suriñach, E., Abellán, A., Khazaradze, G., Garcia, D., and Llosa, J.: Rockfall induced seismic signals: Case study in Montserrat, Catalonia, Natural Hazards and Earth System Science, https://doi.org/10.5194/nhess-8-805-2008, 2008.

Walter, F., Amann, F., Kos, A., Kenner, R., Phillips, M., de Preux, A., Huss, M., Tognacca, C., Clinton, J., Diehl, T., and Bonanomi, Y.: Direct observations of a three million cubic meter rock-slope collapse with almost immediate initiation of ensuing debris flows, Geomorphology, https://doi.org/10.1016/j.geomorph.2019.106933, 2020.

Yuan, B., Tan, Y. J., Mudunuru, M. K., Marcillo, O. E., Delorey, A. A., Roberts, P. M., Webster, J. D., Gammans, C. N. L., Karra, S., Guthrie, G. D., and Others: Using machine learning to discern eruption in noisy environments: A case study using CO 2-driven cold-water geyser in Chimayó, New Mexico, Seismological Research Letters, 90, 591–603, https://doi.org/10.1785/0220180306, 2019.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

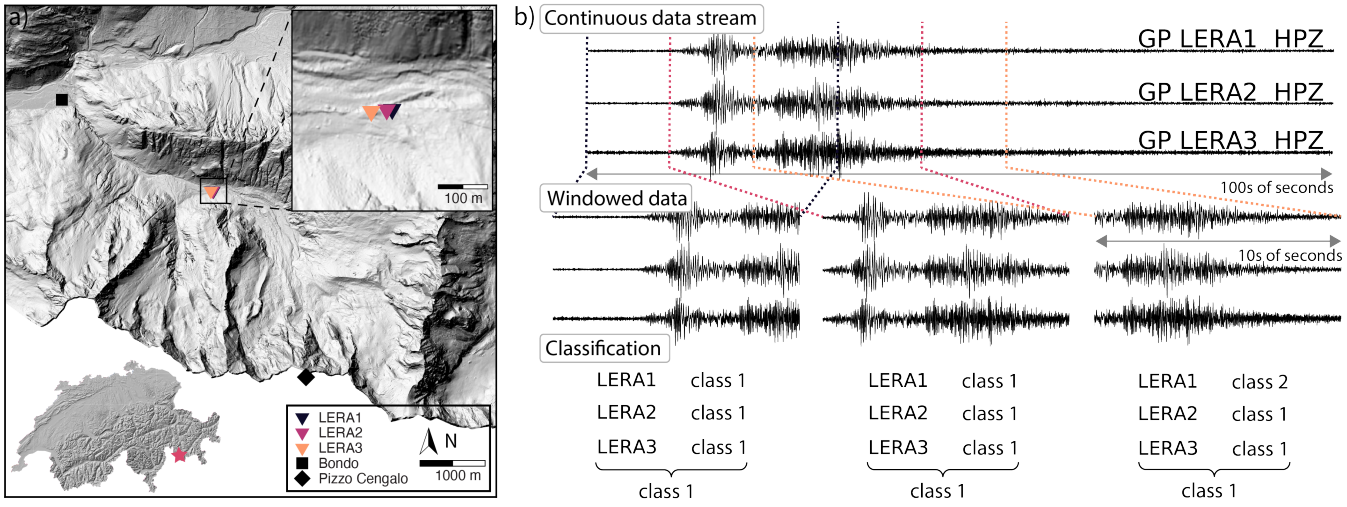

Figure 1. (a) Overview of the Bondasca valley. Outline of Switzerland in the lower left corner, with the star marking the location of the Bondasca valley. Location of seismic stations (LERA1 – 3) depicted as colored triangles. Zoom-in on the seismic stations in the upper right corner. (b) Scheme of the classification with continuous data stream, windowed data, classification per station and label for window.

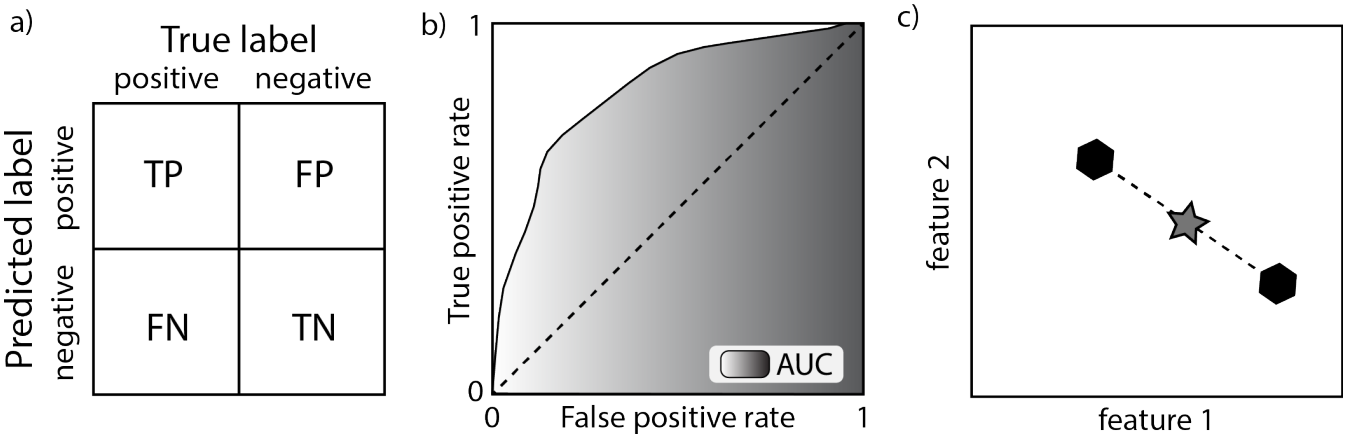

Figure 2. (a) Confusion matrix for a two class problem (positive, negative), with true labels as columns and predicted labels as rows. True positives (TP) and true negatives (TN) on the diagonal and false negatives (FN) and false positives (FP) on the off diagonal elements. (b) ROC curve with true positive rate (TPR) on the y axis and false positive rate (FPR) on the x axis. Shaded area (AUC) as measure for model accuracy. (c) Schematic representation of SMOTE in a two feature space. Hexagons mark training samples with two features that exist in a minority class. The grey star marks the new sample created with SMOTE. Generally, SMOTE considers the k nearest neighbors of a sample (not shown in this scheme).

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

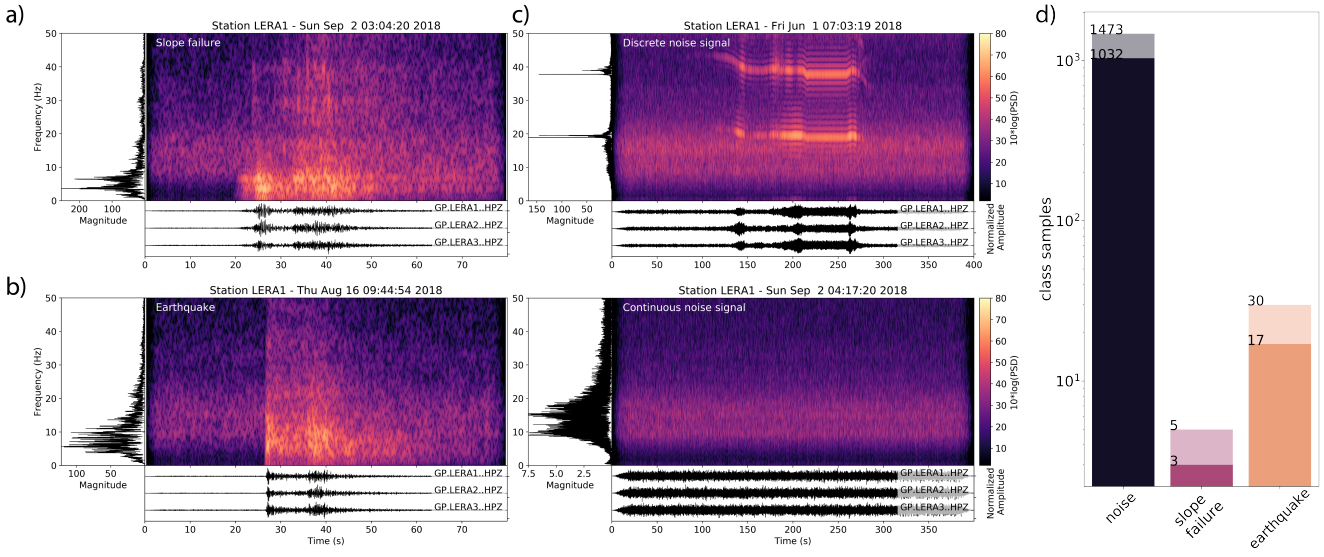

Figure 3. Spectrogram, waveforms of stations LERA1-3 and spectrum of (a) a slope failure, (b) an earthquake and (c) noise signal (top panel) and continuous noise (bottom panel). The slope failure signal shows an emergent onset with several energy bursts and dominant frequencies around 7 Hz. The earthquake signals shows a similar waveform apart from a sharp onset with dominant frequencies slightly higher than the slope failure signals. The displayed discrete noise signal shows a helicopter passing by the station. It is characterized by its discrete frequency band and its overtone. We observe Doppler gliding due to a moving source. Continuous noise shows low but constant spectral power with dominant frequencies between 10 and 25 Hz. (d) Event count per class. Bars show total number of events with the transparent area being the amount of test events and the solid area being the amount of training events. Numbers denote the count of all events and the count on training events.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

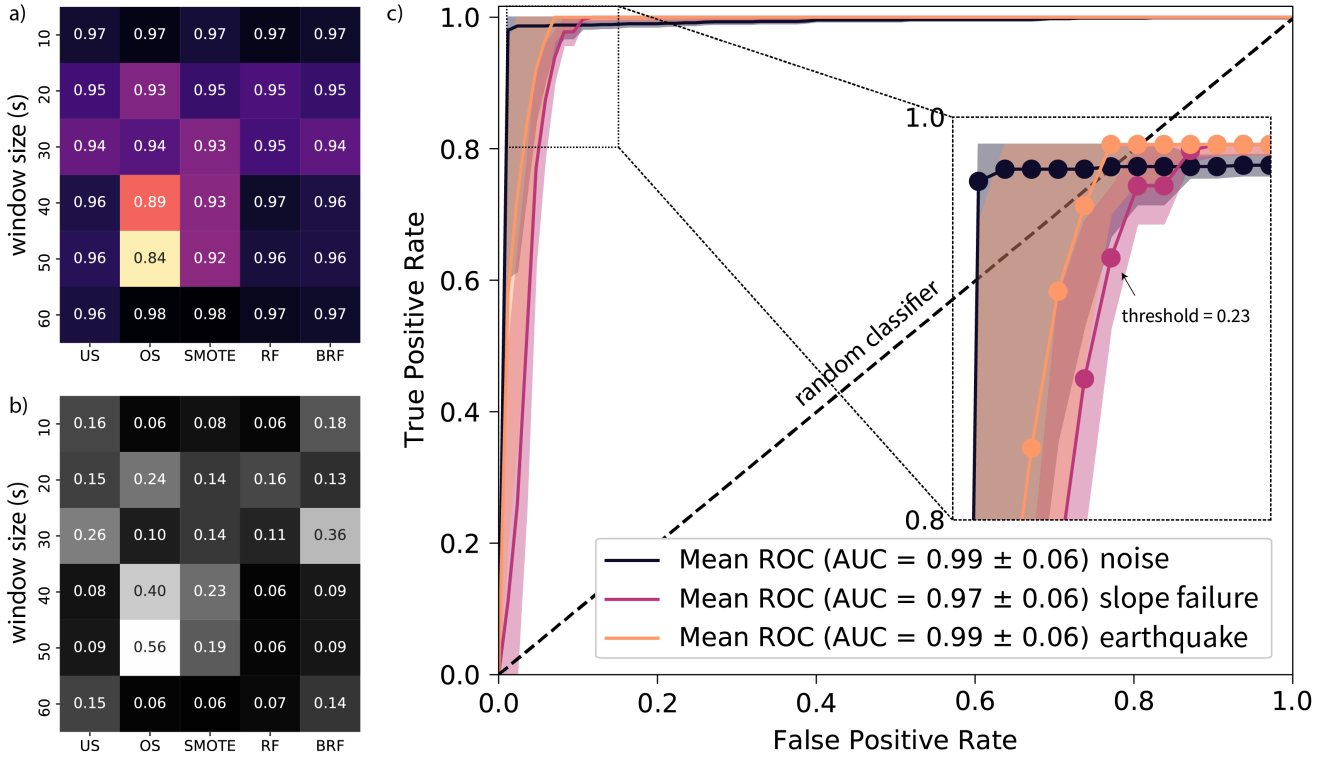

Figure 4. (a) Heatmap of AUC values for SF class for different window sizes and different methods to handle imbalanced data sets, namely undersampling (US), oversampling (OS), synthetic minority oversampling (SMOTE), random forest with original data set (RF), balanced random forest (BRF). The darkest colors denote the largest AUC values. (b) Heatmap of 95% confidence interval of AUC values. Darker colors denote smaller standard deviation (c) ROC curve for 40 seconds time window and random forest. 3-fold cross validation with mean as solid line and 95% confidence interval. Zoom in to corner with circles denoting TPR and FPR values for different thresholds. Probability threshold for TPR rate of SF class > 0.9 is depicted.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

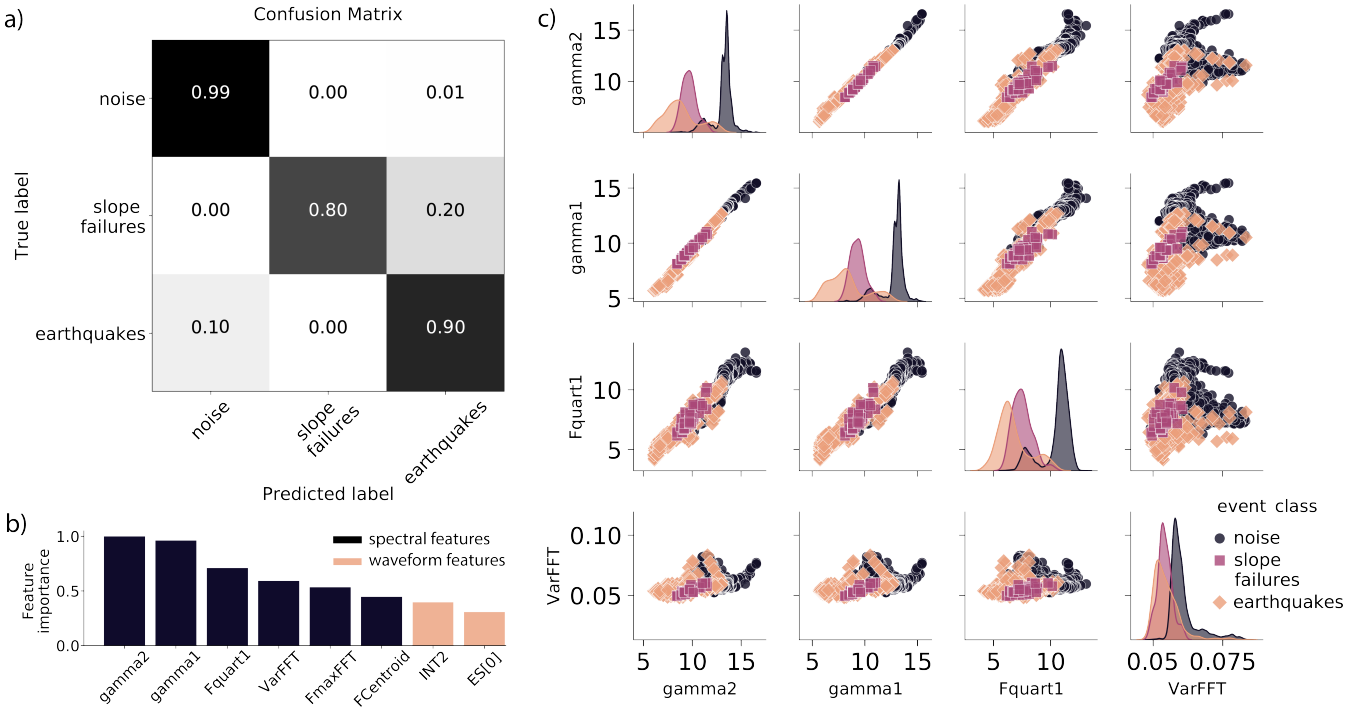

Figure 5. (a) Normalized confusion matrix of final model test. The darker the colors, the higher the values. For an ideal classifier, all samples would be located on the diagonal. (b) Eight most distinct features normalized to one. Dark columns mark spectral features (characteristics of signal in frequency domain), light columns mark waveform features (characteristics of signal in time domain). Labels: spectral gyration radius (gamma2), spectral centroid (gamma1), central frequency of the first quartile (Fquart1), variance of the normalized FFT (VarFFT), frequency at the maximum of the FFT (FmaxFFT), frequency at spectrum centroid (FCentroid), energy of the last 2/3 of the autocorrelation function (INT2), and the energy of the seismic signal in the frequency band of 1-3 Hz (ES[0]) Provost et al. (2017). (c) Pairplot of four most distinct features. Per cell two features plotted against each other, except for diagonal. Diagonal shows univariate distribution of the feature. Colors mark different event classes.

https://doi.org/10.5194/nhess-2020-200
Preprint. Discussion started: 8 July 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

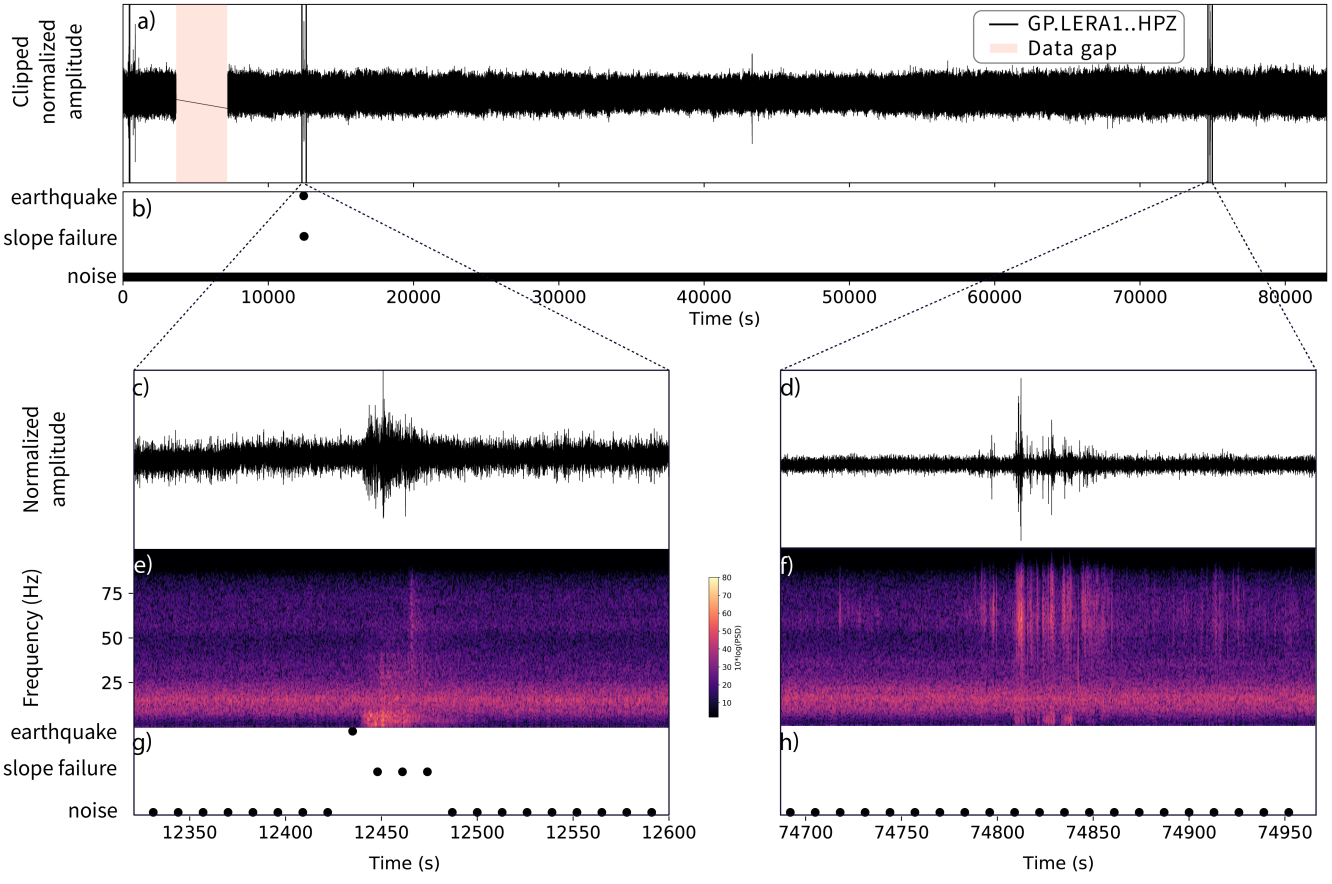

Figure 6. Classifier tested on one day in 2019 (July 16). (a) Waveform of one day, with data gap (orange area) (b) label of each 40 seconds time window. (c) Waveform and (e) spectrogram with spectral power of a slope failure and (g) associated classifications. (d) Noise event with (f) spectrogram and (h) associated classifications.