Automatic Waveform Quality Control for Surface Waves Using Machine Learning

Chengping Chai, Jonas Kintner, Kenneth M. Cleveland, Jingyi Luo,
Monica Maceira, Charles J. Ammon

Description of the Supplemental Material

The supporting information includes a figure (Figure S1) summarizing characteristics of the surface-waveform dataset DA, a map (Figure S2) of seismic event and station locations for dataset DC, a figure (Figure S3) showing the distribution of original quality labels, a figure (Figure S4) showing the spatial distribution of quality labels for two earthquakes, a figure (Figure S5) showing the time windows used to compute statistical features, two examples of hyperparameter tunning (Figure S6), a comparison (Figure S7) of ROC curves, a feature importance plot for random forest (Figure S8), and a table (Table S1, uploaded separately) listing all the seismic networks used by this study.

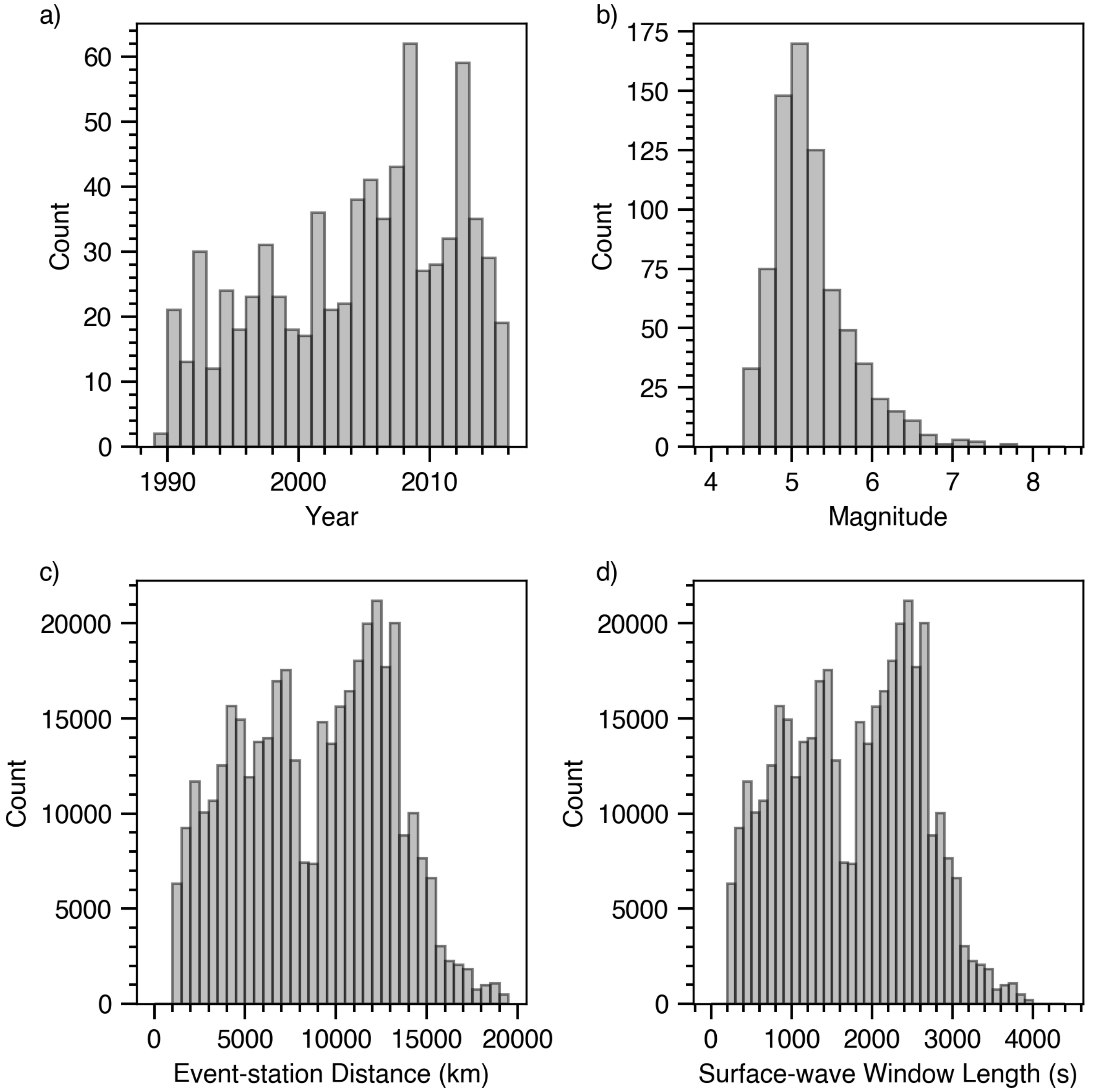

Figure S1. Histograms characterizing the properties of the training dataset DA: (a) origin year of earthquakes; (b) magnitude of earthquakes; (c) the distance between each earthquake and observing seismic station; and (d) the length of surface-wave window defined by a group velocity range from 5.0 to 2.5 km/s. The variable duration of the signals is one of the unusual aspects of this classification problem.

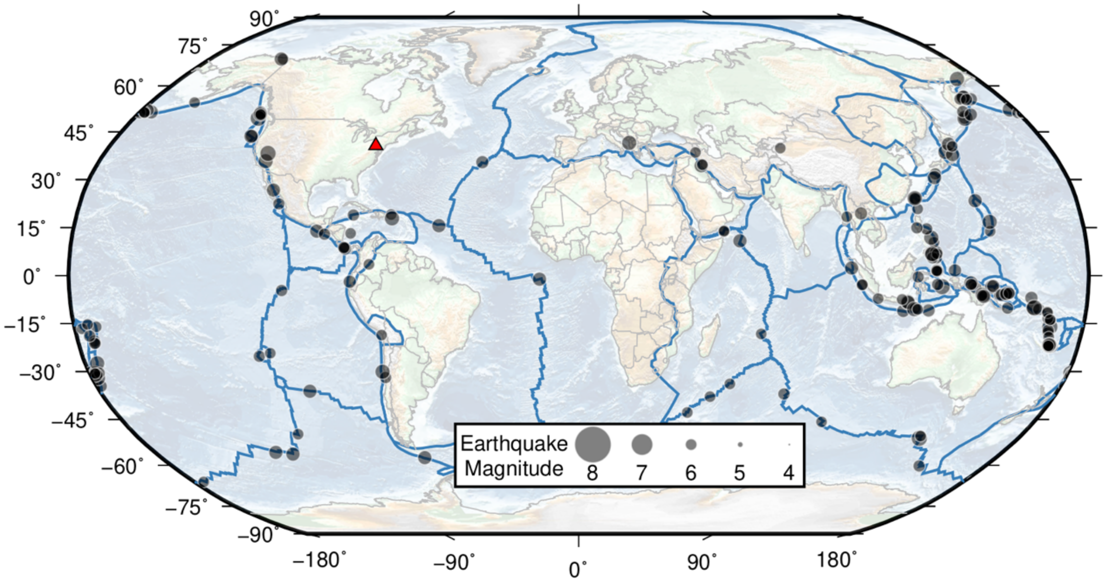

Figure S2. A map of seismic events (gray circles) and the location of seismic station SSPA (red triangle) that were used in the dataset DC.

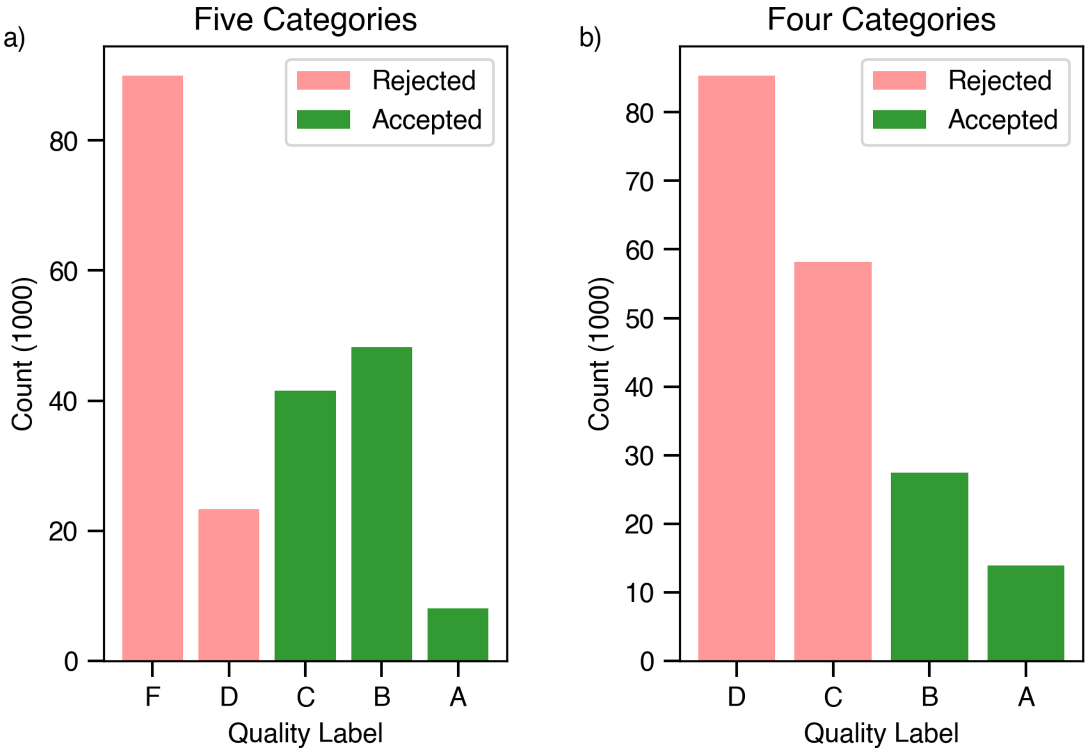

Figure S3. Distributions of original quality labels in dataset DA for (a) five categories and (b) four categories.

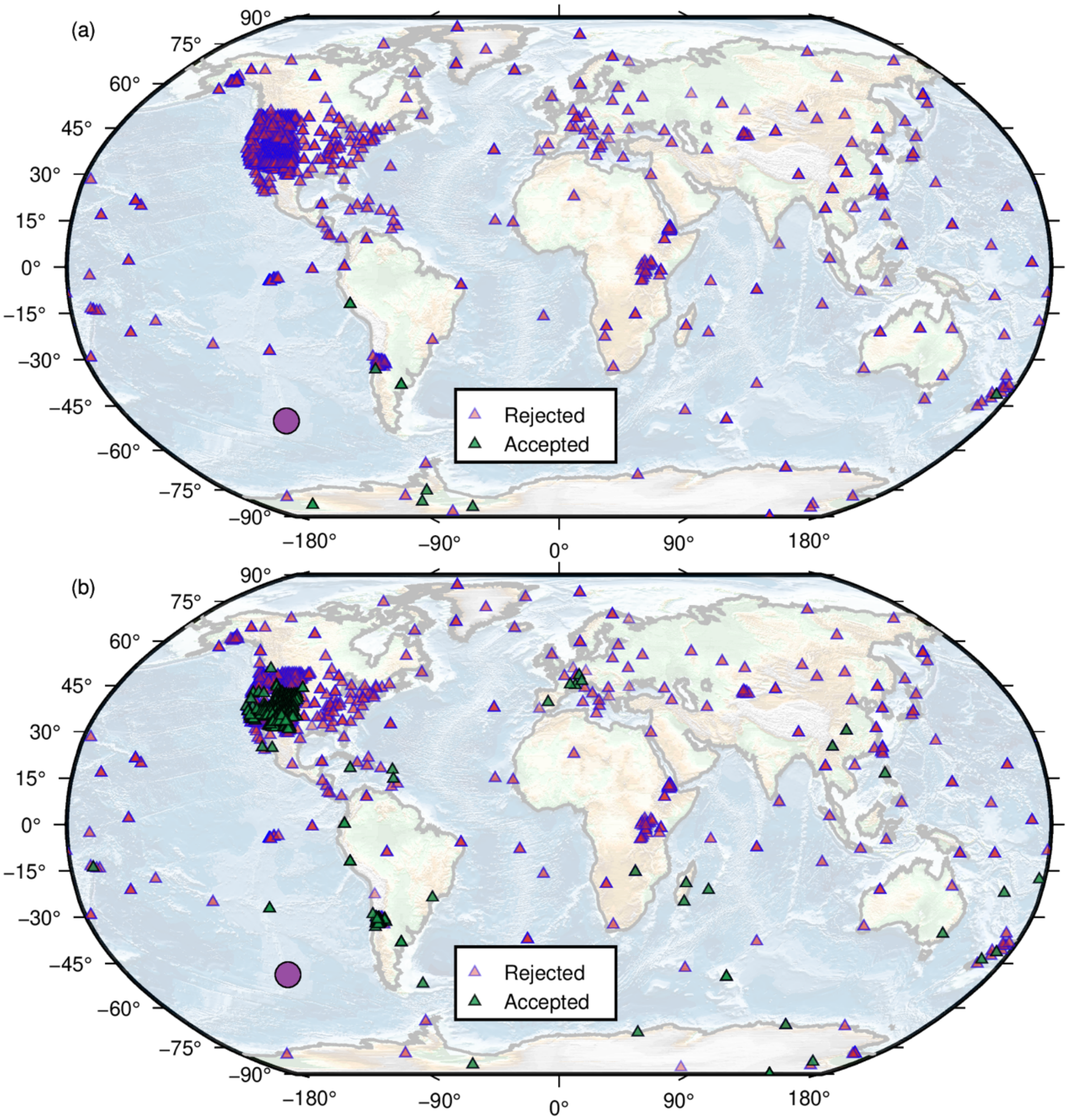

Figure S4. Spatial distributions of quality labels (triangles) for two sample earthquakes (circles) in dataset DA. The event in (a) occurred on 2018/06/12T16:53:34 UTC with a magnitude of 5. The event in (b) occurred on 2018/09/13T15:45:26 UTC with a magnitude of 5.2.

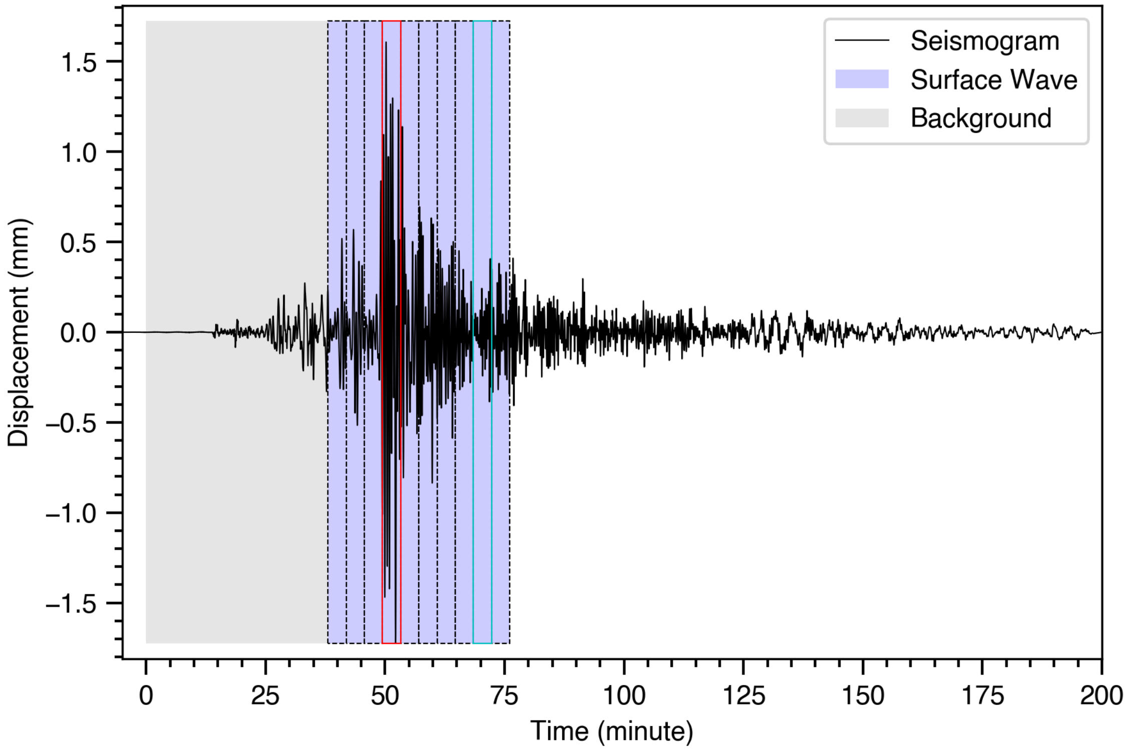

Figure S5. An example surface-wave seismogram with the time windows used for feature engineering illustrated. The dash boxes represent the ten evenly divided time windows. The red box indicates the time window with the maximum absolute energy. The blue box represents the time window with the minimum absolute energy.

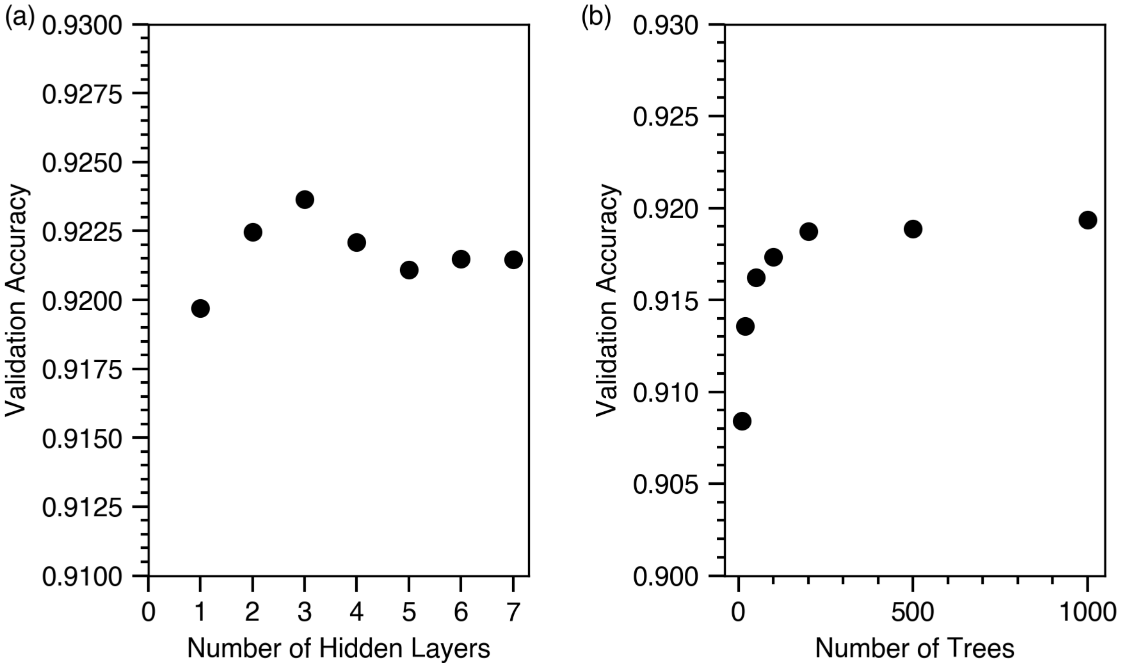

Figure S6. Two examples showing how we select (a) the number of hidden layers for artificial neural networks and (b) the number of trees for the random forest algorithm. Three hidden layers was selected. The RF model contains 200 trees.

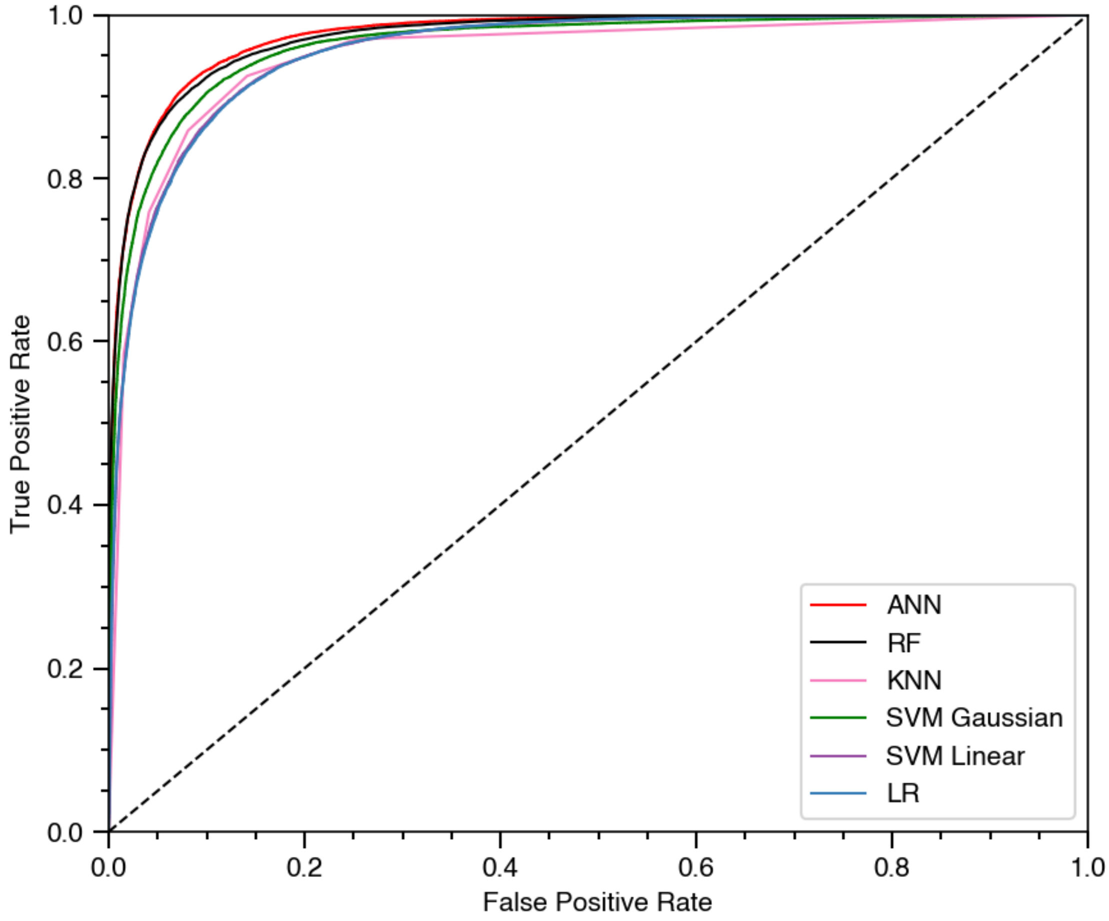

Figure S7. A comparison of the Receiver Operating Characteristic (ROC) Curves for the examined machine learning algorithms constructed using the test set of dataset DA. LR stands for logistic regression. SVM means support vector machine, KNN represents K-nearest neighbors, RF is in short for random forests, ANN represents artificial neural networks.

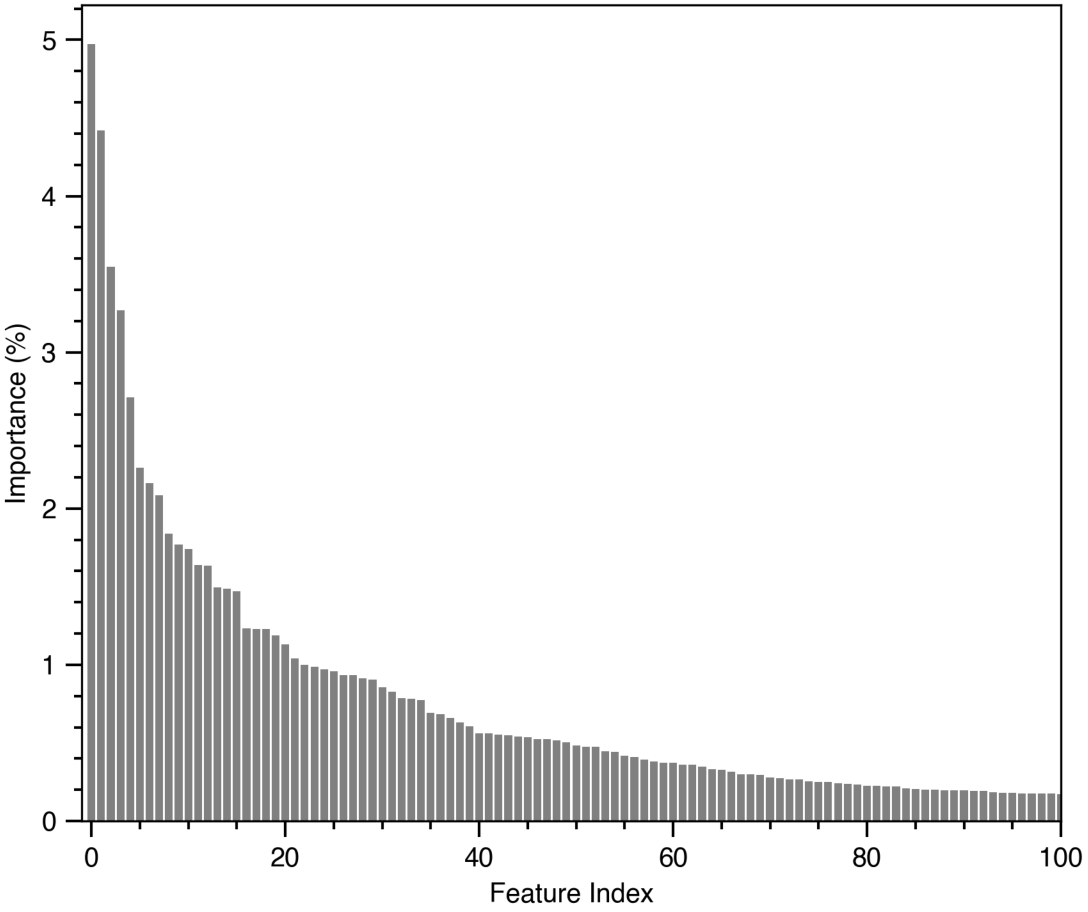

Figure S8. The relative importance of features for the random forest algorithm. The top three features are standard derivation ratio, maximum amplitude ratio, and minimum amplitude ratio between the surface wave and background time windows.

Table S1. A list of seismic networks used.