Supporting Information for

Automatic Waveform Quality Control for Surface Waves Using Machine Learning

Chengping Chai[1], Jonas Kintner[2], Kenneth M. Cleveland[2], Jingyi Luo[3], Monica Maceira[1], Charles J. Ammon[4]

1. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
2. Los Alamos National Laboratory, Los Alamos, New Mexico, USA
3. School of Data Science, University of Virginia, Charlottesville, Virginia, USA
4. Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania, USA

Contents of this file

Additional Supporting Information (Files uploaded separately)

Introduction

The supporting information includes two paragraphs (Text S1) that explain the performance metrics used to compare the different ML algorithms: the F1 score, the Receiver Operating Characteristic (ROC) Curve, and area under the ROC curve (AUC). Also included is 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, plots of example waveforms (Figure S4) that were accepted and rejected by a human analyst, a figure (Figure S5) showing the spatial distribution of quality labels, a diagram (Figure S6) summarizing the two stages of our workflow, a comparison (Figure S7) of ROC curves, a comparison (Figure S8) of confusion matrices, two figures with additional waveform examples (Figure S9 and S10) that were assigned different quality labels by a human analyst and the ANN model, quality control results for group velocity measurements (Figure S11), and a table (Table S1, uploaded separately) listing all the seismic networks used by this study.

Text S1.

Assessing the performance of a classification scheme is typically approached using several metrics of algorithm performance. The metrics are defined in terms of the positive and negative success and failure rates of the classifier when applied to a set of observations independent of the ML training procedure. True positive means that both the predicted label (from the ANN model) and the true label (from a human analyst) are positive (in our case, the waveform is accepted for analysis). False positive means that the predicted label is positive, but the true label is negative (rejected). False negative means that the predicted label is negative, but the true label is positive. True negative means both the predicted label and the true label are negative.

An F1 score can be computed by counting the number of samples in each of these four categories and computing

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{True\ Positive}{True\ Positive + 0.5 \times (False\ Positive + False\ Negative)}$$

The F1 value ranges from 0 (worst performance, no true labels) to 1 (best performance, no false labels). An F1 value of 0.9 corresponds to about 2 false negatives or false positives (combined) for every 9 true positives; an F1 value of 0.95 corresponds to about 10 false negatives or false positives (combined) for every 95 true positives. Machine learning models can provide probabilities associated with each label (accepted or rejected in our case) and a probability threshold can be used to translate the probabilities to labels. For each candidate threshold, we can compute true positive and false positive rates. A ROC curve is a plot of the true positive rate versus the false positive rate for a set of thresholds. The area between the ROC curve and the horizontal axis (the false-positive rate) is called the area-under-the-curve (AUC) score. A machine learning model is usually considered better with a higher AUC score.

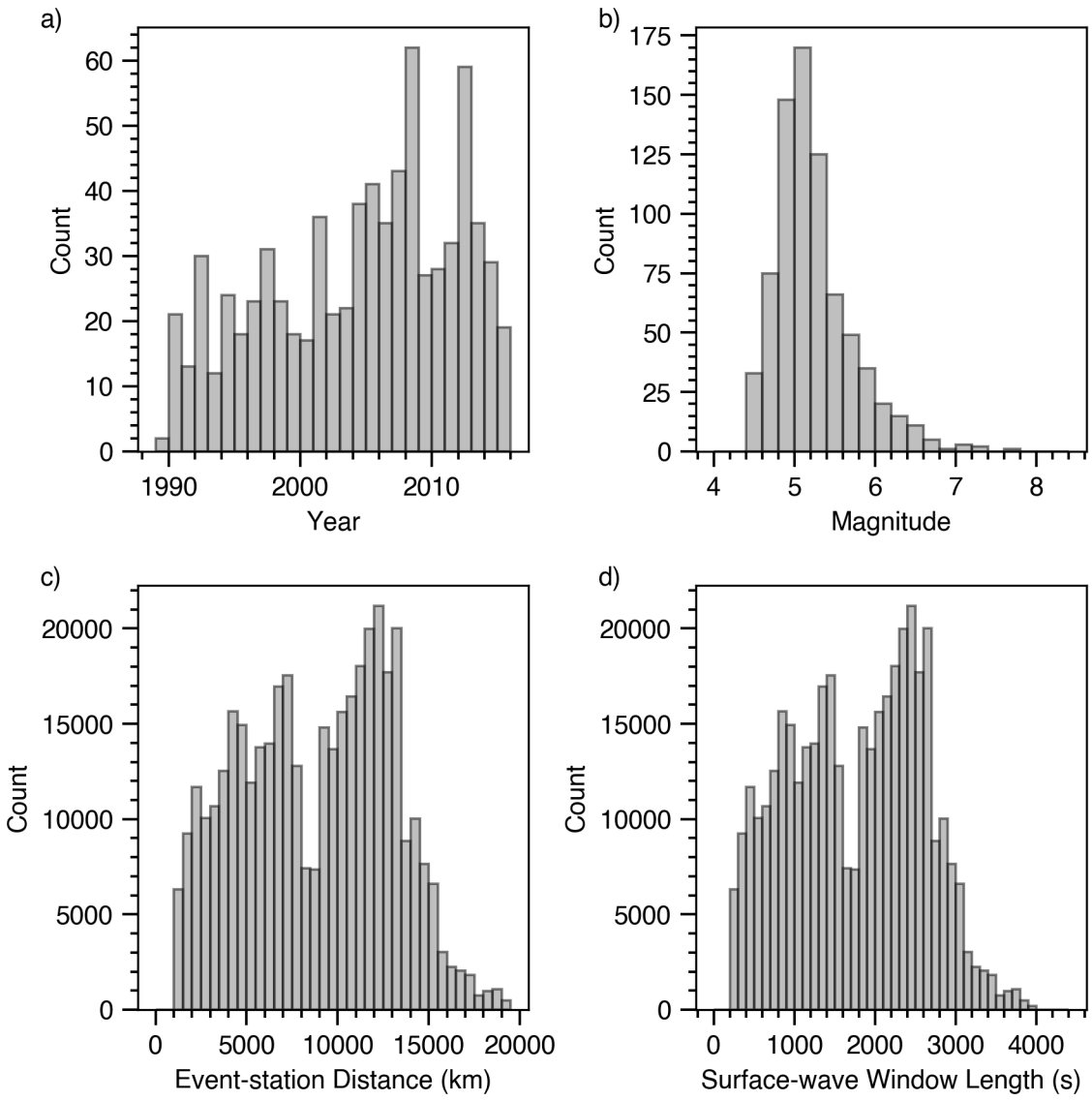

Figure S2. Histograms characterizing the properties of the training dataset DA: (a) origin year of earthquakes; (b) magnitude of earthquakes; (c) the distance between each earthquake and observing seismic station; and (d) the length of surface-wave window defined by a group velocity range from 5.0 to 2.5 km/s. The variable duration of the signals is one of the unusual aspects of this classification problem.

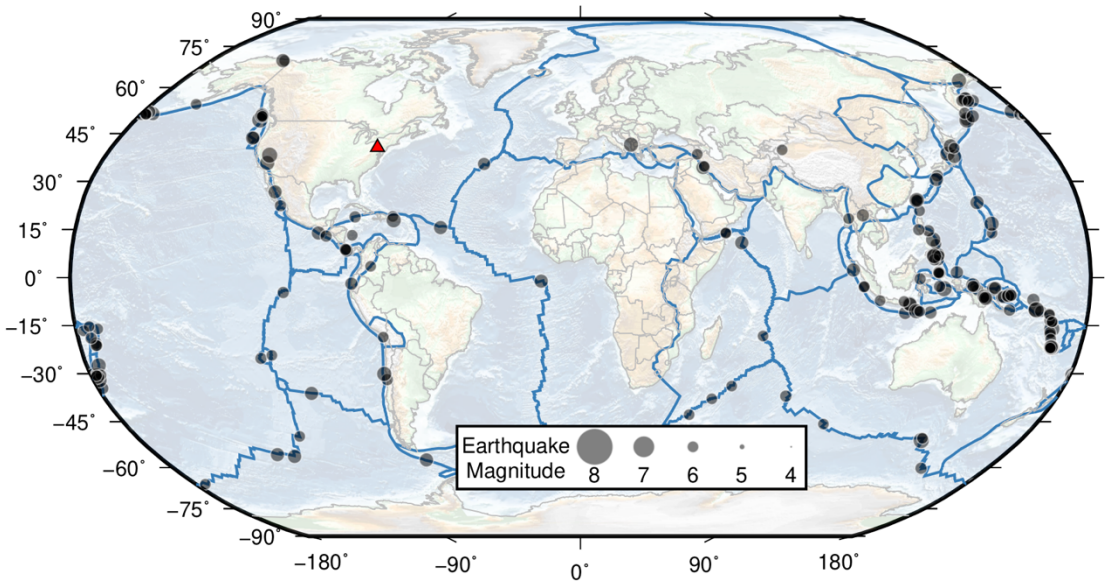

Figure S2. A map of seismic events (gray circles) and the location of seismic station SSPA (red triangle) that were used in the dataset DC.

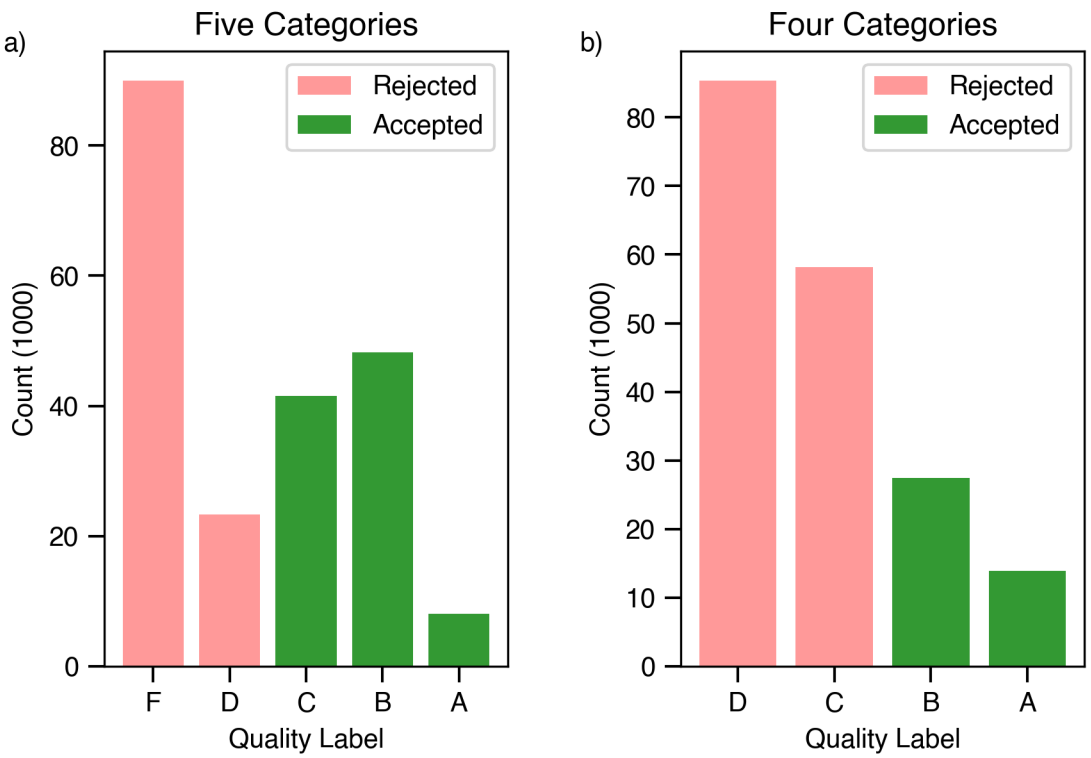

Figure S3. Distributions of original quality labels in dataset DA for (a) five categories and (b) four categories.

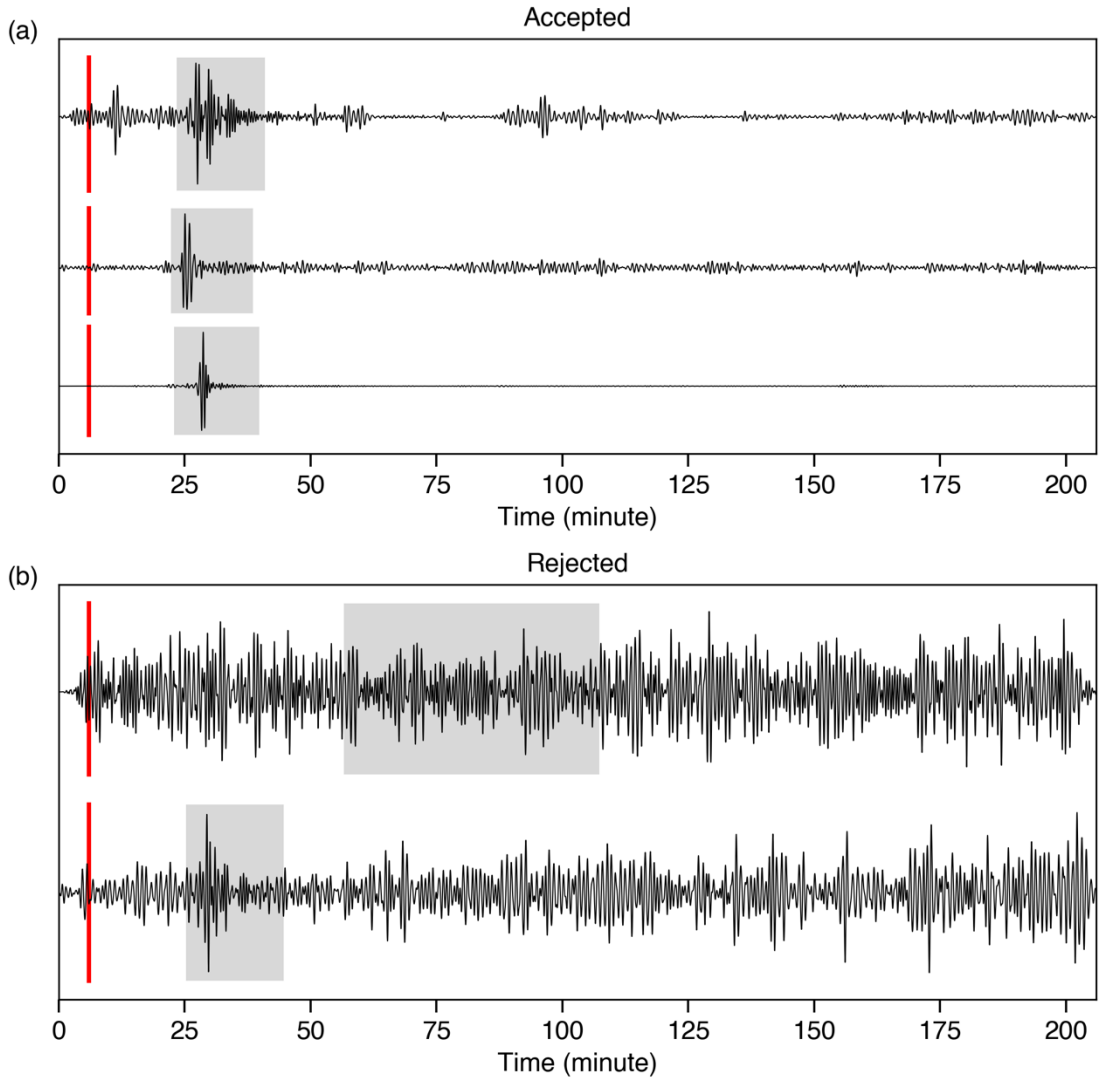

Figure S4. Example displacement waveforms in dataset DA that were (a) accepted and (b) rejected by a human analyst. The red vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity of 2.5 km/s and a maximum of 5 km/s.

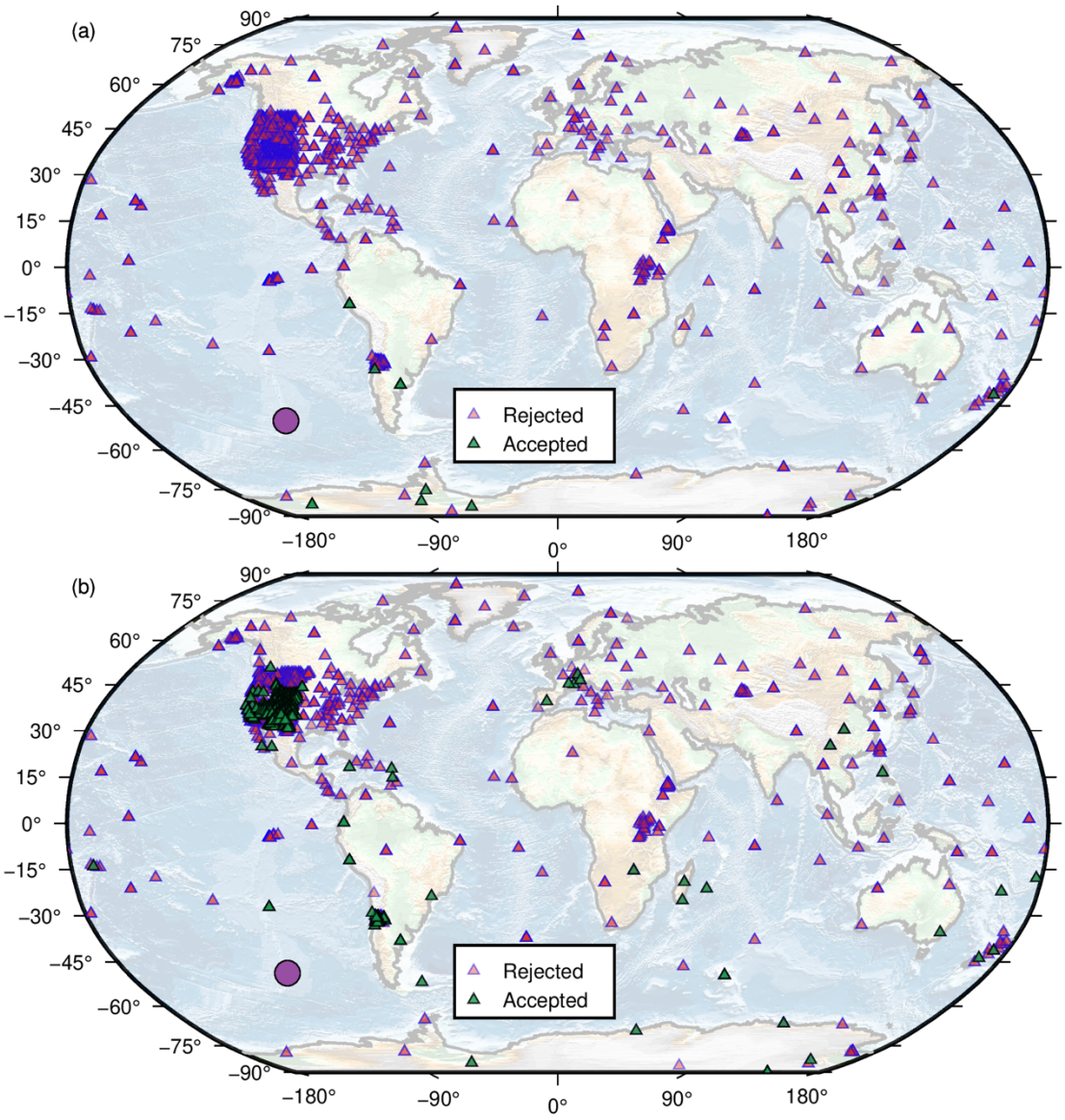

Figure S5. Spatial distributions of quality labels (triangles) for two sample earthquakes (circles) in dataset DA. The event in (a) occurred on 2018/06/12T16:53:34 UTC with a magnitude of 5. The event in (b) occurred on 2018/09/13T15:45:26 UTC with a magnitude of 5.2.

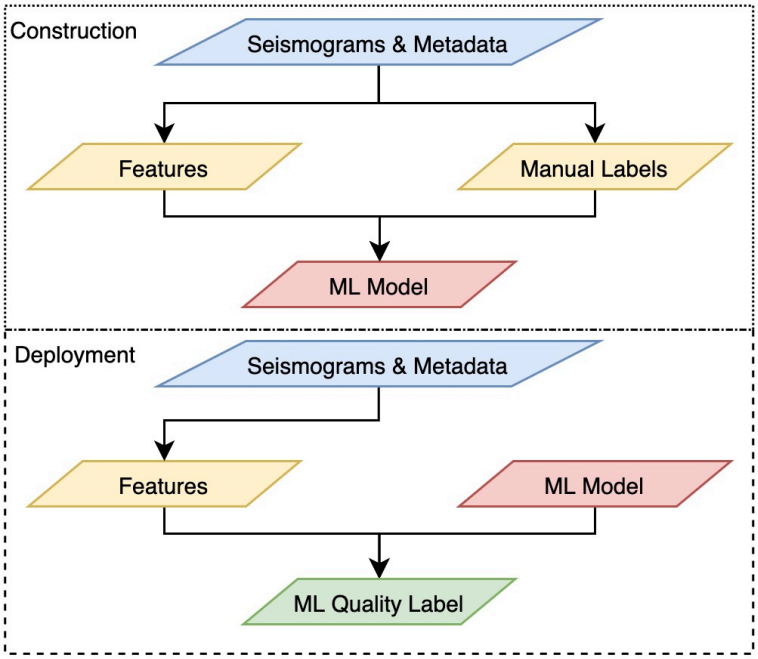

Figure S6. A flowchart illustrating the major steps of the (top) model construction and (bottom) model deployment stages. ML represents machine learning.

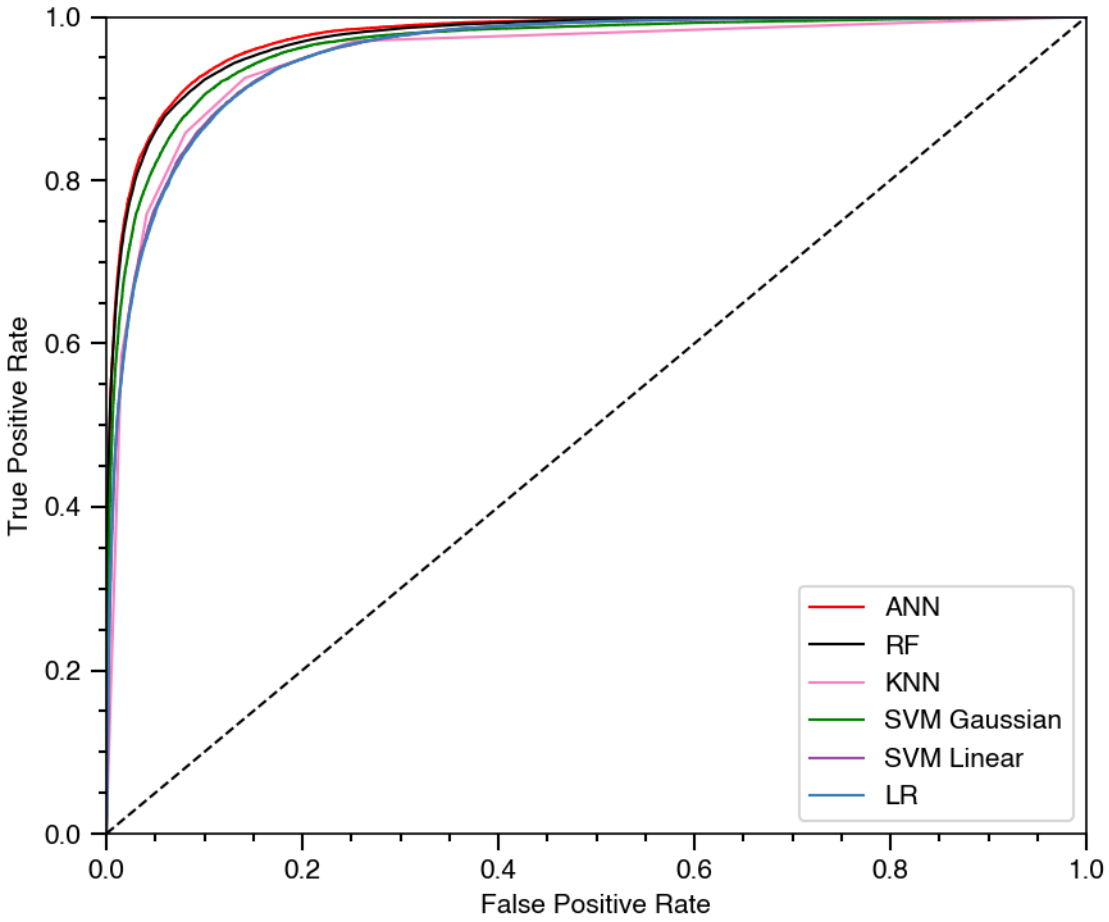

Figure S7. A comparison of the Receiver Operating Characteristic (ROC) Curves for the examined machine learning algorithms constructed using the test set of dataset DA. LR stands for logistic regression. SVM means support vector machine, KNN represents K-nearest neighbors, RF is in short for random forests,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

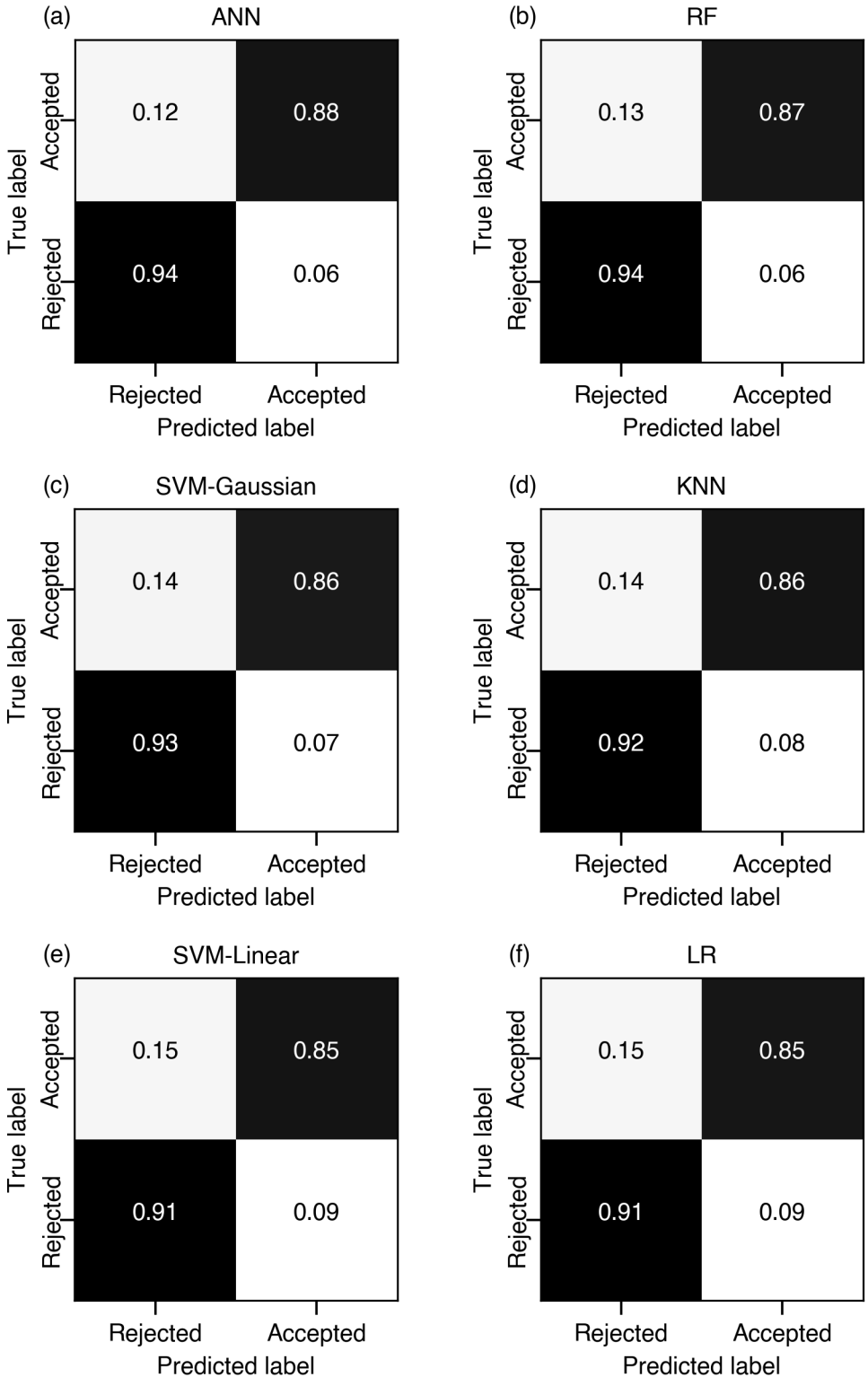

Figure S8. A comparison of confusion matrices for different machine learning algorithms using the test set of dataset DA.

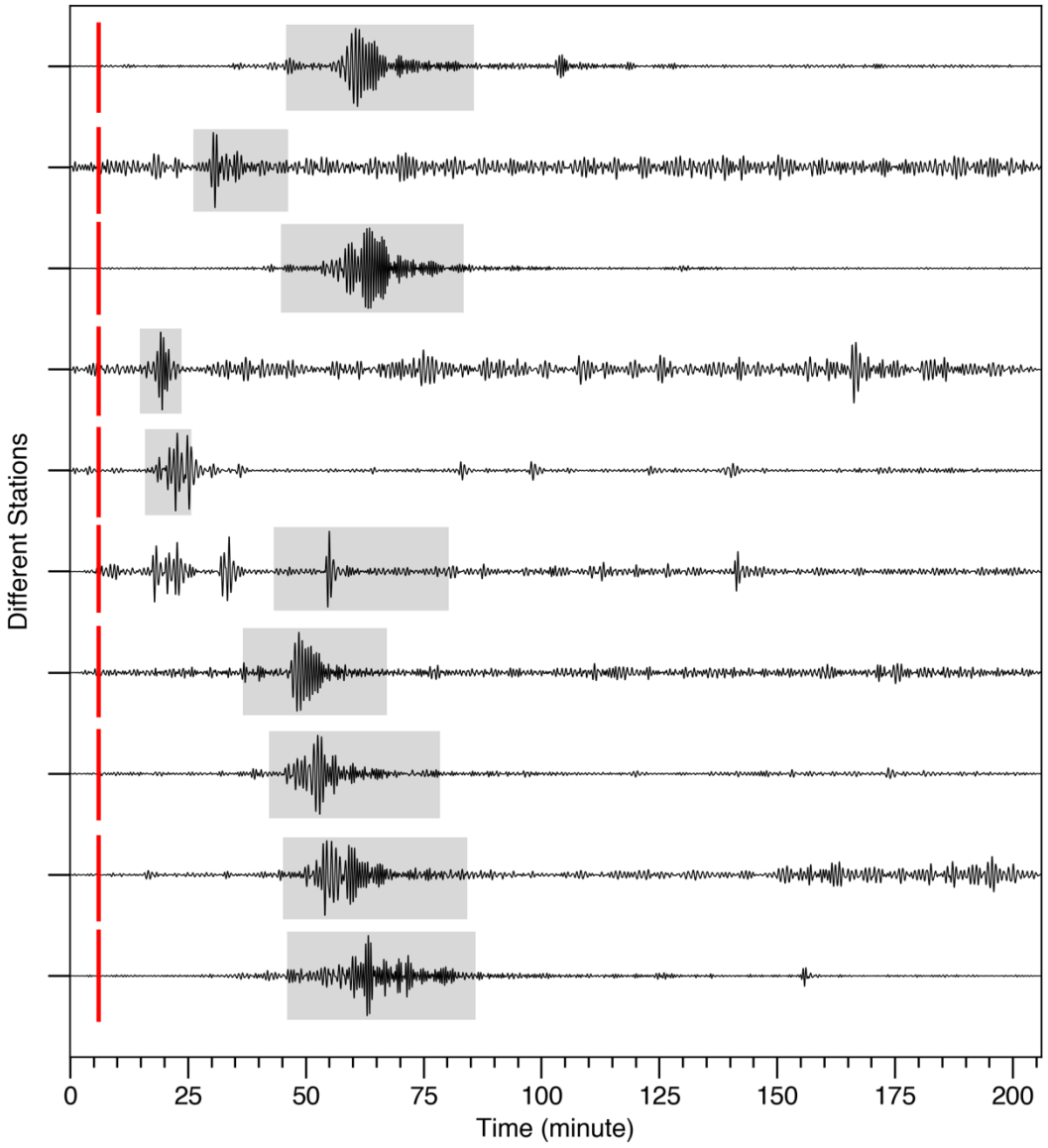

Figure S9. Waveform examples from the test set of dataset DA that were rejected by a human analyst but accepted by the ANN model. The vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity of 2.5 km/s and a maximum of 5 km/s. Most of these misclassifications are likely the result of analyst fatigue. The fifth waveform from the bottom shows enough complexity outside the surface wave window to raise suspicion of the signal. A total of 2861 (6%) seismograms out of 51474 human-rejected waveforms were accepted by the ANN model.

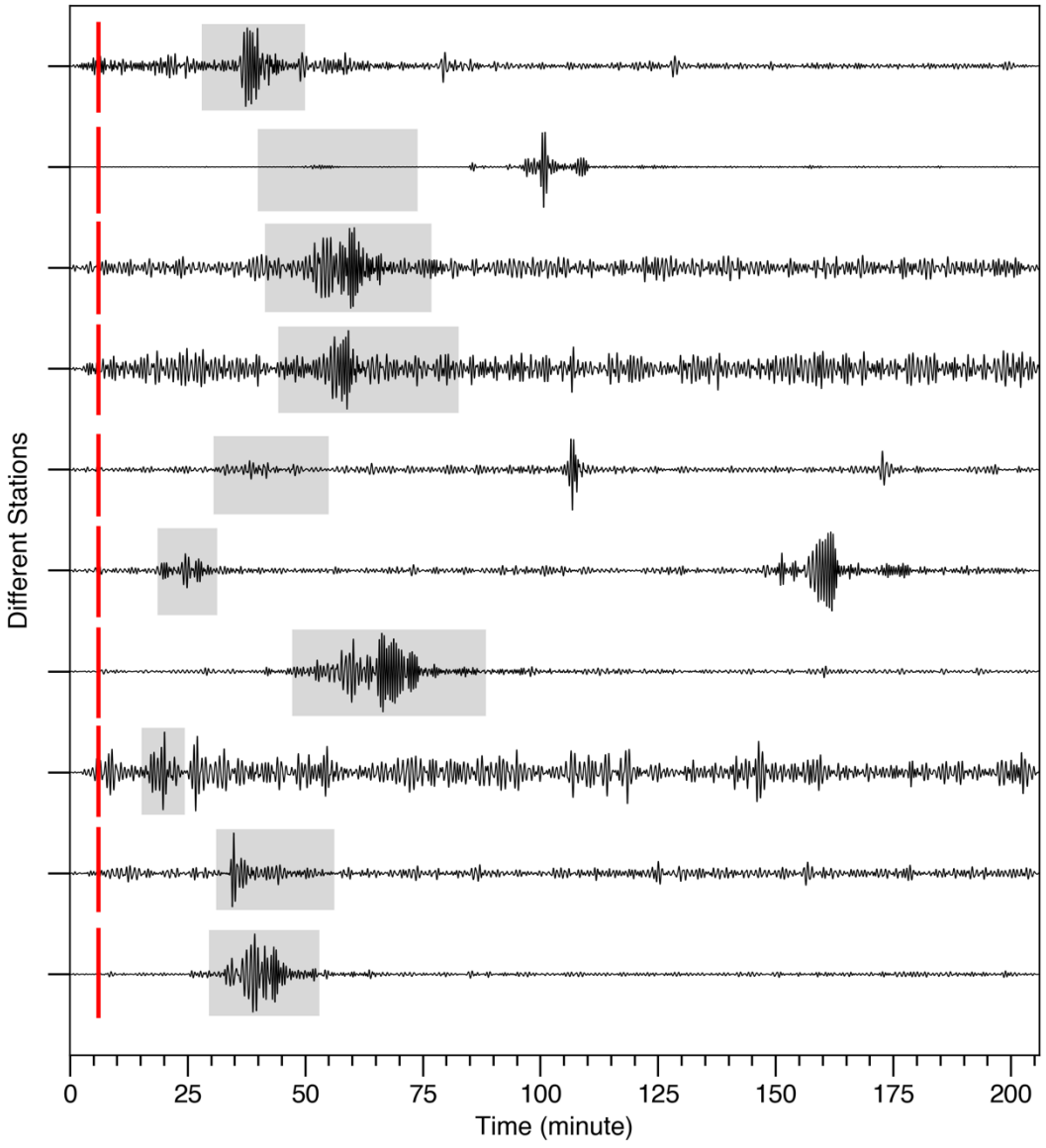

Figure S10. Waveform examples from the test set of dataset DA that were accepted by a human analyst but rejected by the ANN model. The vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity of 2.5 km/s and a maximum of 5 km/s. A total of 3368 seismograms (12%) out of 27731 human-accepted waveforms were rejected by the ANN model.

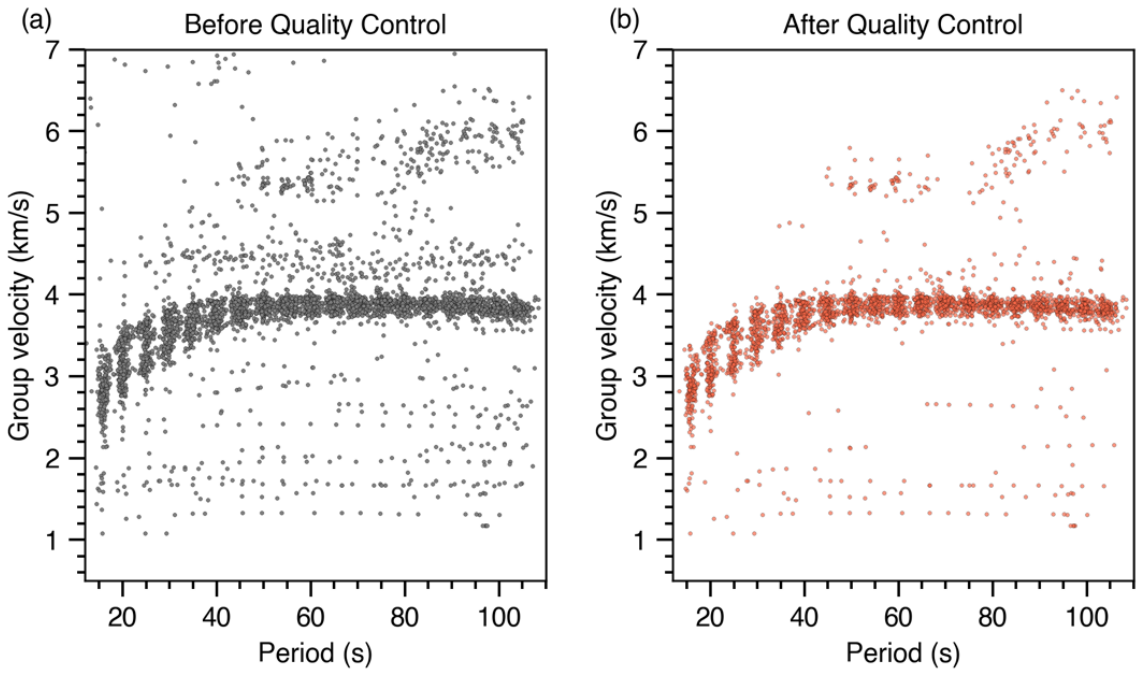

Figure S11. Automatic group velocity measurements (a) before and (b) after using the ANN model for quality control. Automated group velocities are estimated using a simple multiple filter analysis and automated identification of the time of the maximum in a Gaussian-filtered surface waveform. An unrealistic automated group velocity estimate is likely a result of surface-waveform with low signal-to-noise such that the maximum is not associated with the surface wave.

Table S1. A list of seismic networks used.