Posted on 30 Nov 2022 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.1002/essoar.10507941/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Automatic Waveform Quality Control for Surface Waves Using Machine Learning

Chengping Chai[1,1], Jonas A. Kintner[2,2], Kenneth M. Cleveland[2,2], Jingyi Luo[3,3], Monica Maceira[4,4], and Charles Ammon[5,5]

[1]Oak Ridge National Laboratory (DOE)
[2]Los Alamos National Laboratory
[3]University of Virginia
[4]Oak Ridge National Laboratory (DOE)
[5]Pennsylvania State University

November 30, 2022

Abstract

Surface-wave seismograms are widely used by researchers to study Earth's interior and earthquakes. Reliable results require effective waveform quality control to reduce artifacts from signal complexity and noise, a task typically completed by human analysts. We explore automated approaches to improve the efficiency of waveform quality control processing by investigating logistic regression, support vector machines, k-nearest neighbors, random forests (RF), and artificial neural networks (ANN) algorithms. Trained using nearly 400,000 waveforms with human-assigned quality labels, the ANN and RF models outperformed other algorithms with a test accuracy of 92%. We evaluated the trained models using seismic events from geographic regions not used for training. The results show the trained models agree with labels from human analysts, but required only 0.5% time. Although the quality assignments assessed general waveform signal-to-noise, the ANN or RF labels can help facilitate detailed waveform analysis, reducing surface-wave measurement outliers without human intervention.

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], and 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.

Corresponding author: Chengping Chai (chaic@ornl.gov)

Key Points:

- We applied five machine learning algorithms to a waveform quality control problem using a labeled dataset of 400,000 surface-wave samples
- Neural networks and random forests outperformed other algorithms with a higher accuracy, a faster execution speed, and a smaller storage
- The trained neural network and random forest performed equally to human analysts but used only 0.5% of time of human analysts

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Abstract

Surface-wave seismograms are widely used by researchers to study Earth's interior and earthquakes. Reliable results require effective waveform quality control to reduce artifacts from signal complexity and noise, a task typically completed by human analysts. We explore automated approaches to improve the efficiency of waveform quality control processing by investigating logistic regression, support vector machines, k-nearest neighbors, random forests (RF), and artificial neural networks (ANN) algorithms. Trained using nearly 400,000 waveforms with human-assigned quality labels, the ANN and RF models outperformed other algorithms with a test accuracy of 92%. We evaluated the trained models using seismic events from geographic regions not used for training. The results show the trained models agree with labels from human analysts, but required only 0.5% time. Although the quality assignments assessed general waveform signal-to-noise, the ANN or RF labels can help facilitate detailed waveform analysis, reducing surface-wave measurement outliers without human intervention.

Plain Language Summary

Surface waves generated by earthquakes carry valuable information about Earth's subsurface and the sources that generate them. To reliably and robustly extract information from a suite of surface waveforms, the signals require quality control screening. This process has typically been done by experts labeling each data sample visually, which is time-consuming and tedious for large datasets. To speed up signal quality assessment, we trained machine learning methods using a large set of human-labeled waveforms. We compared five techniques: logistic regression, support vector machines, k-nearest neighbors, random forests, and artificial neural networks. The artificial neural networks performed the best and achieved an accuracy of 92%. Once trained, the neural network model matched human performance but reduced the time cost by 99.5% when applied to data it had never seen. Our analyses demonstrate the capability of automated processing to improve quality in surface-wave-related measurements without human quality control screening.

1 Introduction

Surface waves have long been used for subsurface imaging (e.g., Ekström, 2011) and earthquake source studies (e.g., Ammon, 2005). Recently, double-difference seismic source location derived using surface wave cross-correlations at globally-distributed stations has proven successful in various geological settings (Chai et al., 2019; Cleveland et al., 2015, 2018; Cleveland & Ammon, 2013; Howe et al., 2019; Kintner et al., 2018, 2019, 2020, 2021). These techniques require reliable surface-wave measurements, which is usually assured through the careful visual inspection of seismograms. With seismic network deployments increasing in frequency and size, the amount of available surface-waveforms is also increasing. More data is unequivocally a good thing, but quality control of the ever-growing data volumes requires substantial time and effort. The complexity of surface-wave signals and the spatially and temporally varying character of seismic background noise makes reliable automation of the quality control process a challenge. In some cases, data quality control becomes the most time-consuming part of a seismological analysis.

Machine learning (ML) has shown promise when applied to a variety of seismological research problems. This includes body-wave detection and arrival-time picking (e.g., Chai et al., 2020; Mousavi et al., 2020; Perol et al., 2018; Ross et al., 2018; Yoon et al., 2015; L. Zhu et al.,

2019; W. Zhu & Beroza, 2018) and signal association (e.g., McBrearty et al., 2019; Ross et al., 2019). ML has also been used for seismic source studies that include earthquake location (e.g., X. Zhang et al., 2020), earthquake magnitude estimation (e.g., Mousavi & Beroza, 2020), earthquake focal mechanism determination (e.g., Kuang et al., 2021), and seismic signal discrimination (e.g., Li et al., 2018; Meier et al., 2019; Seydoux et al., 2020). ML algorithms have also been developed for seismic tomography (e.g., Bianco & Gerstoft, 2018; Z. Zhang & Lin, 2020), and laboratory earthquake prediction (e.g., Rouet-Leduc et al., 2017). Most existing work has focused on body-wave analysis, few studies have focused on applying ML to the quality control of regional and teleseismic intermediate-period surface-waveforms.

An important application of ML in geophysics is to reduce the burden of seismic processing to a level that allows more observations (more earthquakes, more seismograms, etc.) to be included in seismic analyses. We develop automated quality control processes that decrease the data quality assessment burden and increase overall data quality applicable to research efforts into earth structure (Herrmann et al., 2021) and seismic source analysis (e.g., Lay et al., 2018), while also being a source of data for long standing projects that quantify earthquake sources from regional to global scales (e.g., Ekström et al., 2012). No automated process is perfect, but application of ML approaches can effectively and efficiently identify the best and worst data and allow human attention to focus on marginal-quality and unexpected observations that require more understanding and experience to assess.

In this work, we explore the opportunities of ML to aid in the analysis of intermediate-period regional and teleseismic seismic surface waves. We compiled roughly 400,000 surface-wave signals and associated quality labels from stations around the globe. The quality labels are from past studies that focused on events in various tectonic settings. We trained five ML models including logistic regression (LR, Hosmer Jr et al., 2013), support vector machine (SVM, Suykens & Vandewalle, 1999), K-nearest neighbors (KNN, Keller et al., 1985), random forests (RF, Breiman, 2001), and artificial neural networks (ANN, Jain et al., 1996) to perform automated quality control processing of intermediate-period surface-wave seismograms. We compared the performance, speed, and disk usage of these ML techniques. We also tested the general applicability of the best-performing model to events from other geographic regions.

2 Data

The data consist of seismic waveforms (along with metadata) and quality labels. The seismograms were downloaded from the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC) archive. Each waveform is associated with a particular seismic event that has known location and origin time information. The seismograms start six minutes before the origin time and end 200 minutes after the origin time. We removed the instrument response from the seismograms and rotated the horizontal components to the radial and transverse coordinate system from the original north-south east-west coordinates. To isolate intermediate-period Love and Rayleigh waves, seismograms were bandpass filtered to isolate signals with periods between 30 and 60 s.

2.1 Seismic data

During the model construction stage, we used observations from 759 seismic events and 4,502 seismic stations (Figure 1). The seismograms were analyzed for previous earthquake relocation efforts (Cleveland et al., 2018; Cleveland & Ammon, 2013, 2015; Kintner et al., 2018,

2019). The origin times of these seismic events range from May 1989 to October 2016 (Figure S1a). The magnitudes of the events range from roughly 4.5 to 7.8 (Figure S1b). The event-station distance spans a wide range from 10- to 180-degree (Figure S1c). Using a group velocity range from 5.0 to 2.5 km/s, the expected surface-wave window length ranges from 222 s to 3979 s (Figure S1d). We refer to these seismograms as dataset DA.

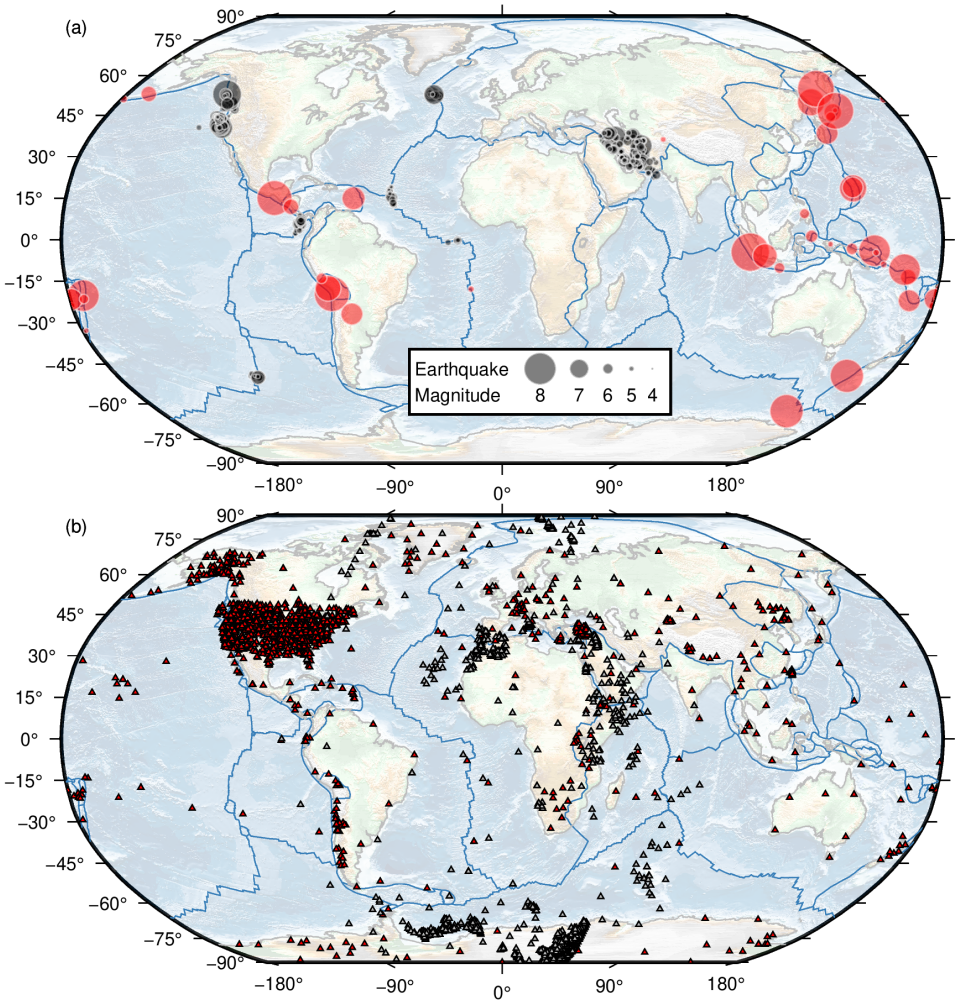

Figure 1. Maps of the (a) earthquakes and (b) seismic stations used. The size of each circle in (a) is proportional to an event's earthquake magnitude. The gray circles and triangles are used for training (dataset DA), whereas red symbols are used to evaluate the ML model after training is completed (dataset DB). Thick lines are tectonic plate boundaries (Bird, 2003). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

During the model construction stage, we selected 40 seismic events from the United States Geological Survey (USGS) ComCat catalog (between January 1990 and January 2019) with the following criteria. (a) The events were located at least one arc degree away from any seismic events of dataset DA. (b) The events were randomly selected from four magnitude bins (between magnitudes 5 and 6; 6 and 7; 7 and 8; 8 and 9) with 10 events in each bin. Seismic stations including long period high gain seismometers (LH channels, 1 Hz sampling rate) and

located between 10- and 180-degree distance were selected. Data from temporary network deployments were excluded. These seismograms will be referred to as dataset DB. After the ML models were trained, we also downloaded seismograms from 184 seismic events (Figure S2) with magnitude 6.0 and larger between January 2018 and May 2020 recorded at the station SSPA located near Standing Stone, Pennsylvania, USA. These seismograms comprise dataset DC.

2.2 Waveform quality labels

We compiled the quality labels for dataset DA from several earthquake relocation studies (Cleveland et al., 2018; Cleveland & Ammon, 2013, 2015; Kintner et al., 2018, 2019). Due to personal preferences, the original quality labels have either five or four categories. When using five categories, the three highest categories were considered acceptable (Figure S3a). For four categories, the two highest were considered acceptable (Figure S3b). To combine the datasets and maximize the number of labels, we mapped the quality labels into two categories, either accepted or rejected (see Figure S4 for waveform examples). The spatial distributions of quality labels show significant variations for different seismic events (Figure S5) due to earthquake source differences and background noise variations.

In addition, after the ML models were trained, three human analysts re-labeled 1,000 seismograms randomly selected from dataset DA. Half of them were assigned the same quality label by both a human analyst and the ANN model and referred to as Dataset 1. The other half were assigned different quality labels by a human analyst and the ANN model and referred to as Dataset 2. The human analysts also labeled 2,000 seismograms from dataset DB after we applied the ANN model to all the seismograms in dataset DB. These seismograms were randomly selected such that (1) each of the 40 distributed-magnitude earthquakes has 50 waveforms and (2) 1,000 seismograms were accepted by the ANN model whereas the other 1,000 seismograms were rejected. We consider the majority vote of the three analysts as ground truth, which is more reliable (but costly) than the labels used in the model construction stage.

3 Methods

Our analyses consisted of two stages (Figure S6), model construction and deployment. During the model construction stage, we compute statistical features from the surface-waveforms and link them with manually-assigned quality labels. These features and labels are then used to train an ML model. During the deployment stage, we obtain and compare ML-derived quality labels by applying the ML model directly to a test set of surface-waveforms not used in model construction.

3.1 Feature engineering

Surface waveforms are one of the most recognizable part of a seismic event's wavefield, but also one of the most variable. The character of the signal changes with source-to-station distance, geology along the wave's path, as well as the earthquake rupture characteristics and faulting geometry. To capture these complexities in a reasonable number of parameters, we employed a total of 301 features for each waveform (data sample). The features were computed from waveform segments that include: (1) the expected surface-wave arrival window (defined using a group velocity window from 5.0 to 2.5 km/s); (2) a time window with common duration before the surface waves; (3) ten evenly divided time windows spanning the entire data sample. For each time window, we calculated absolute energy (sum of all time samples squared), the sum

of absolute derivatives, kurtosis, skewness, maximum, minimum, mean, standard deviation, nine quantiles (10%-90%), and number of time samples. In additional to absolute values, we also included ratios of these statistical features (excluding the number of samples and absolute sum of changes) for two time-window pairs. The first pair includes the surface-wave window and the time window ahead of the surface wave. The second pair includes the two windows from the ten evenly sized windows that have the maximum and minimum absolute energy. We also included the magnitude of the earthquake, event depth, azimuth, and distance between the station and seismic event as signal-related features. As with all signal classification studies, we explored these features guided by our experience with surface wave analysis as well as numerical experiments using the training and validation sets.

3.2 Machine learning

In the model construction stage, we used data from dataset DA and randomly split it into three sets. We used 277,213 samples (waveforms) for training (70% of total), 39,601 samples for validation (10% of total), and 79,205 samples for testing (20% of total). The validation set was used to choose training parameters and features, the test set was used to evaluate the performance of the ML models. We used scikit-learn's implementation of the LR, SVM, KNN, and RF. The ANN was implemented with Keras. For the SVM algorithm, we used both a linear kernel (SVM-Linear) and a nonlinear kernel (SVM-Gaussian). The KNN model used five closest neighbors. The RF model contains 100 trees. The ANN model has six fully connected hidden layers (256 neurons), which used the rectified linear units (ReLU) activation function and followed by a dropout layer (10% dropout rate) to reduce overfitting. We set the batch size as 20 and the learning rate as 0.00001.

3.3 ML model generality

We tested the generality of the ANN model using a collection of seismic events located in different regions than the events used in the model construction stage. The qualities of a subset of seismograms were visually assessed by three analysts and compared against the ANN model results. The original research objective for assigning a signal's quality label was to decide whether it had the bandwidth and signal-to-noise ratio to perform well in a cross-correlation analysis, as well as to recognize interference with other arrivals, instrument issues, nodal signals, etc.. We tested the ANN's generality using it as a screening procedure for an automated measurement of surface-wave group velocities. The model was applied to surface-wave seismograms in Dataset DB (see next section). Surface-wave group velocities were automatically estimated from seismograms in Dataset DC. Many of the group velocities estimated from seismograms rejected by the ANN model were clear outliers.

4 Results

4.1 ML model construction and assessment

The performance of a classifier can be measured in a number of different ways, but most essential metrics are constructed using the numbers of positive and negative success and failure rates of the classifier. When trained using all the training samples, RF and ANN model out-performed LR, SVM-Linear, KNN, SVM-Gaussian when applied to the test dataset in terms of accuracy score, F1 score (see Text S1 for detailed definition), and area under the receiver operating characteristic curve (AUC, see Text S1 for detailed definition) as shown in Figure 2a.

The receiver operating characteristic curves show the same pattern as AUC (Figure S7). The accuracy score, F1 score, and AUC for the ANN model are 0.92, 0.89, and 0.97, respectively. The performance of LR and SVM-Linear was the poorest. The confusion matrices also show that the RF and ANN models performed better than others (Figure S8).

We visually checked waveforms that the ANN model assigned different labels than a human analyst using interactive visualization tools similar to Chai et al. (2018). We observed both human quality assignment errors as well as errors by the ANN model (see Figure S9 and S10 for examples). The results indicate that the ANN was working at least as accurately as human analysts. Mislabeling by human analysts is not surprising given the tediousness of the task and the natural inclination for humans to tire during the process. Mislabeling by the ANN represents the appearance of a signal with characteristics that are not in the training set, or combinations of features that contradict the general patterns in the training data.

The runtime (which includes loading the trained model and computing quality labels) of the LR, RF, and ANN model are among the fastest for 100,000 seismograms (using six 2.9 GHz CPU cores) (Figure 2b). SVM models are the slowest since the algorithm used was not parallelized. The trained KNN model uses the most disk space (1.4 GB), the LR model required the least disk space (3 KB) (Figure 3d). SVM-Linear, RF, and SVM-Gaussian require comparable storage. The ANN model requires 5 MB of storage. Considering performance, runtime, and disk space, we prefer the ANN model and the RF model for assigning a quality control value to surface-wave seismograms.

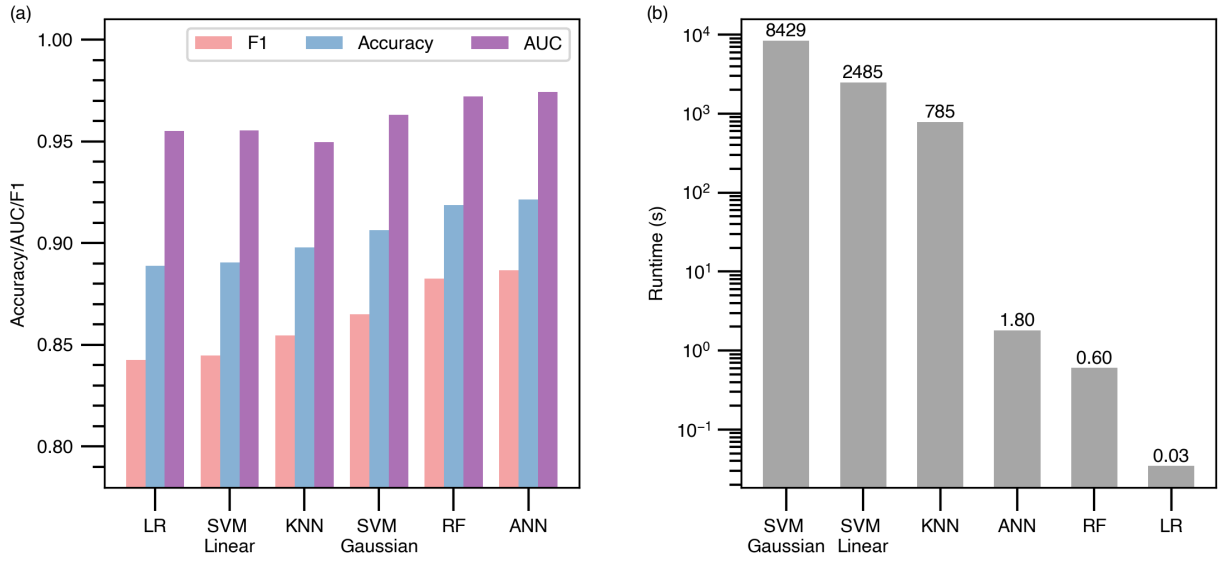

Figure 2. A comparison of (a) performance and (b) runtime for the test set from dataset DA. The performance analysis include all training samples in the dataset. The runtime is calculated by recording the time it takes for different ML algorithms to load the trained model and compute quality labels for 100,000 seismograms.

We also constructed ML models using subsets of the complete training set to investigate the model performance as a function of the number of training samples. This analysis consisted of training sets built using 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, and 100000 waveforms. As expected, the F1-score for all the algorithms improved with an increasing number

of training samples (Figure 3a and 3b). However, as model performance increases, more training samples are needed to improve the model performance by the same percentage. That is, initial improvement occurs rapidly, but as the dataset grows and accuracy increases, significantly more data are needed to make a substantial performance improvement. The RF algorithm has the best accuracy and F1 score when the number of training samples is less than or equal to 20,000. The ANN algorithm surpassed the RF method when the training samples exceed 20,000. As shown in Figure 3c, the training time (using thirty-two 2.1-GHz Intel Xeon cores) for LR, KNN, and RF algorithms is less than the other ML techniques. The training time for the SVM models increases rapidly with the number of training samples. The ANN model took longer to train, but the training time increases more slowly with the number of training samples.

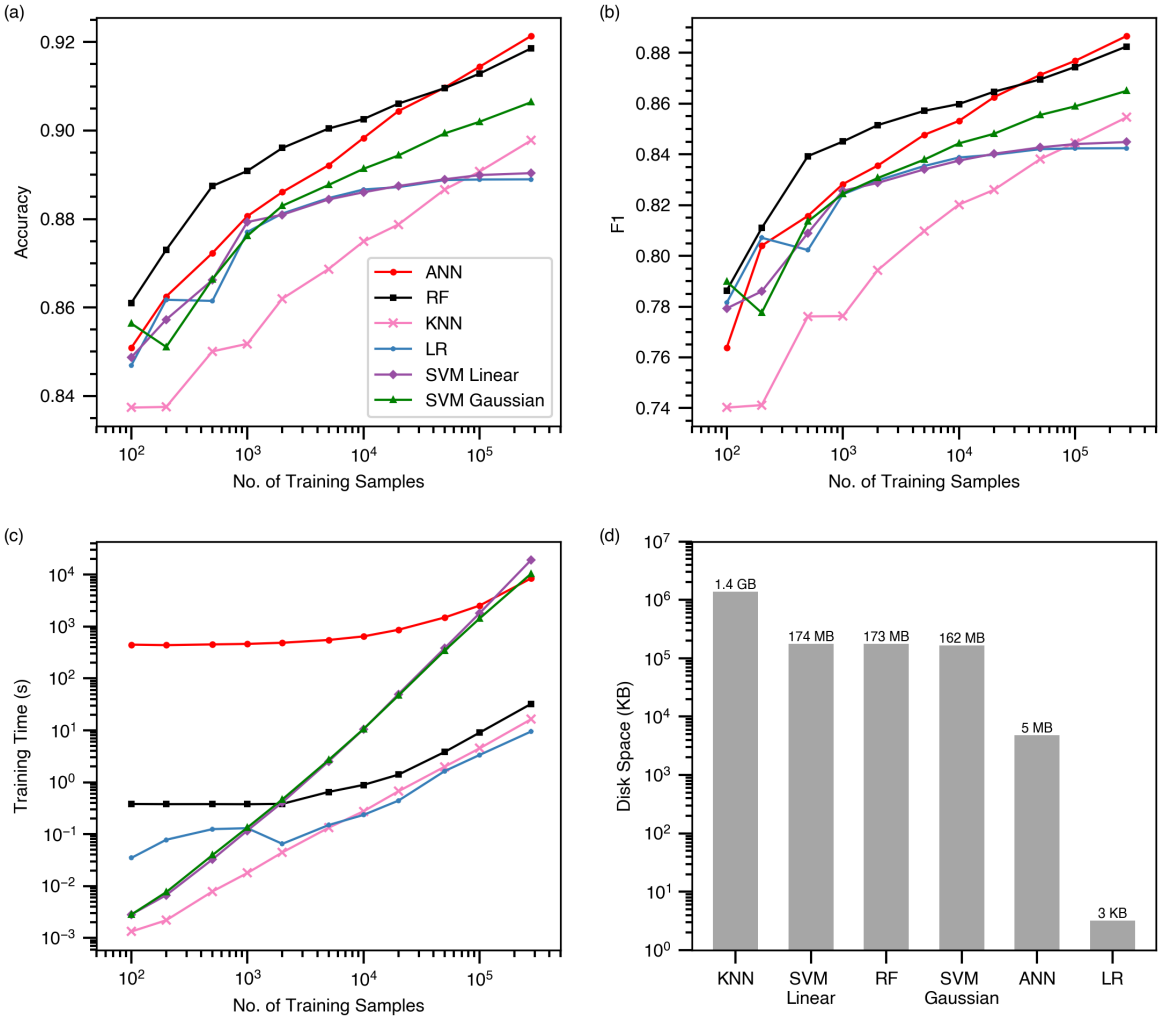

Figure 3. A comparison of performance (a) and (b), training time (c), and disk space usage (d) for different algorithms. The legends of (b) and (c) are the same as (a).

4.2 Model Applications

We compared the performance of the ANN and RF models against three human analysts using datasets 1, 2, and 3. The results shown in Figure 4 indicate that the ANN and RF models performed similarly to human analysts for all three datasets. Of course the ANN and RF models

only used 0.5% of the average human processing time (Figure 4b). In some cases, the ANN and RF models identified useable data that were rejected by one of human analysts (see Figure 4e for an example). The direct outputs of the ANN and RF models are probability scores (range from 0 to 1), which are then converted into two categories using a default threshold of 0.5, accepted (larger than or equal to 0.5) or rejected (smaller than 0.5). The probability threshold can be adjusted for a stricter screening. Increasing the threshold can improve the performance as shown in Figure 4c and 4d. When the threshold is larger than 0.5, three categories can be assigned to a seismogram instead of two. For example, a signal can be rejected if its probability score is smaller than 0.4, accepted if the probability is larger than or equal to 0.6, or considered marginal if its probability is between 0.4 and 0.6. The marginal seismograms can be further inspected by human analysts. As expected, a higher threshold leads to a smaller number of nonmarginal (accepted or rejected) labels (Figure 4c and 4d) or in other words more waveforms for human analysts to inspect. Similar to human analysts, the ANN and RF models sometimes agree and other times disagree. For dataset 3, the ANN or RF models combined mislabeled 540 seismograms out of a total of 2000. Both methods incorrectly labeled a subset of 186 seismograms (9% of the total); the ANN model mislabeled an additional 207 seismograms (393 total, overall 80% correct); the RF model mislabeled another 147 seismograms (333 total, overall 83% correct).

Though not directly trained for the quality control of group velocity estimation, we tested the ANN model to determine whether it would reduce outliers in automated group velocity measurements. The ANN model performed reasonably well for dataset DC reducing the number of unrealistic group velocity values using the ANN-based quality control (Figure S11). The result is not perfect but the operational burden of inspecting the outlier observations is substantially reduced. Transfer learning (e.g., Chai et al., 2020) may further improve the performance of the ANN model for the quality control of group velocities.

5 Conclusions and Discussion

Using nearly 400,000 waveforms and corresponding quality labels, we applied and compared five ML algorithms (LR, SVM, KNN, RF, and ANN) intended to improve the efficiency of the quality control of surface-wave seismograms. Considering performance, processing speed, and storage requirements, the ANN achieved an accuracy of 0.92, an F1 score of 0.89, and an AUC of 0.97. The RF model follows the ANN closely with slightly lower performance and higher storage requirements, but faster processing times. We prefer the ANN and RF models over the other algorithms tested. The performances of both the ANN and RF model match human analysts for data they have never seen while also reducing the time invested in surface-wave quality control by 99.5%. We also show that quality labels from the ANN model helps reduce outliers in group velocity measurements, despite the training labels originally being generated for the purposes of signal cross-correlation analysis. The improved processing speed of the ANN model compared to human analysts and a demonstration of this method to independent surface-wave measurements shows that this technique can be used to reduce the burden of quality control screening for large volumes of seismic data.

The trained ANN and RF models can be incorporated into an existing workflow that uses intermediate-period surface wave seismograms for earthquake and/or earth-structure studies. For fast-response applications, these two trained ML models can be applied automatically to identify good-quality data rapidly without human intervention. The execution speed of the two ML

models can be easily increased with more computing resources. For more comprehensive studies, the trained models can be used to pre-screen a large amount of data and allow researchers to focus on a subset of data ranked by ML labels. The numeric quality scores from the RF and ANN ML models could also be used as initial quality weights in seismological analysis.

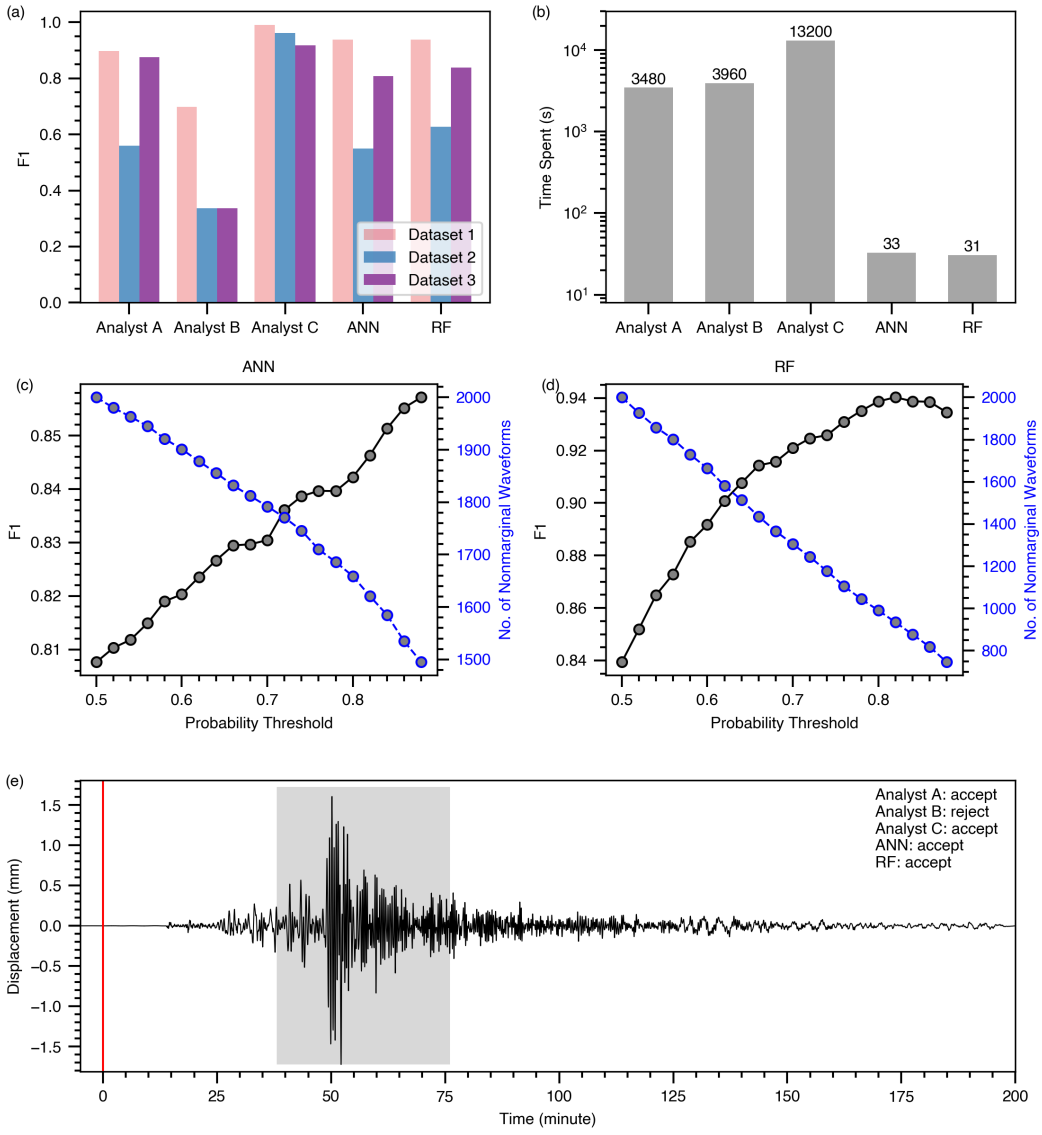

Figure 4. Additional evaluation of the ANN model after training. Panels (a) and (b) compare the ANN model against three analysts A, B, and C using a subset of 3000 seismograms from Dataset DA and DB. Note the time spent by the ANN model in (b) includes the entire processing workflow from raw seismograms to quality labels. Panels (c) and (d) show F1 and number of ML model labeled seismograms as a function of probability threshold using dataset 3. The sample seismogram in (e) was rejected by Analyst B and accepted by Analyst A, Analyst C, and the ANN model. The vertical line indicates the origin time of the 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.

Acknowledgments and Data

This work was supported by the U.S. Department of Energy (DOE), Office of Fossil Energy, Carbon Storage Program through the Science-informed Machine Learning for Accelerating Real-Time Decisions in Subsurface Applications (SMART) Initiative. This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US DOE. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan, last accessed in January 2021). We thank helpful discussions with Kipton Barros, Singanallur Venkatakrishnan, and Derek Rose. The authors declare that there is no conflict of interest regarding the publication of this article.

The authors thank the developers of GMT version 5.4.4 (Paul Wessel et al., 2013) and version 6.1.1 (P. Wessel et al., 2019), Obspy version 1.2.2 (Beyreuther et al., 2010; Krischer et al., 2015; Megies et al., 2011), Numpy (Van Der Walt et al., 2011), Matplotlib version 3.4.2 (Hunter, 2007), Scikit-learn version 0.23.2 (Pedregosa et al., 2011), Keras version 2.4.3 (https://keras.io/, last accessed in January 2021), and Google Earth (https://www.google.com/earth/, last accessed in January 2021). The facilities of the Incorporated Research Institutions for Seismology (IRIS) Data Services, and specifically the IRIS Data Management Center (https://ds.iris.edu/ds/nodes/dmc/, last accessed in January 2021), were used for access to waveforms and related metadata required for waveform data. See Table S1 for a full list of seismic networks used in this study. We thank United States Geological Survey for making the ComCat catalog (https://earthquake.usgs.gov/earthquakes/search/, last accessed in January 2021) openly available.

References

Ammon, C. J. (2005). Rupture Process of the 2004 Sumatra-Andaman Earthquake. *Science*, *308*(5725), 1133–1139. https://doi.org/10.1126/science.1112260

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010). ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, *81*(3), 530–533. https://doi.org/10.1785/gssrl.81.3.530

Bianco, M. J., & Gerstoft, P. (2018). Travel Time Tomography With Adaptive Dictionaries. *IEEE Transactions on Computational Imaging*, *4*(4), 499–511. https://doi.org/10.1109/TCI.2018.2862644

Bird, P. (2003). An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems*, *4*(3), 1027. https://doi.org/10.1029/2001GC000252

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chai, C., Ammon, C. J., Maceira, M., & Herrmann, R. B. (2018). Interactive Visualization of Complex Seismic Data and Models Using Bokeh. *Seismological Research Letters*, *89*(2A),

668–676. https://doi.org/10.1785/0220170132

Chai, C., Ammon, C. J., & Cleveland, K. M. (2019). Aftershocks of the 2012 Off-Coast of Sumatra Earthquake Sequence. *Tectonophysics*, *763*(April), 61–72. https://doi.org/10.1016/j.tecto.2019.04.028

Chai, C., Maceira, M., Santos-Villalobos, H. J., Venkatakrishnan, S. V., Schoenball, M., Zhu, W., et al. (2020). Using a Deep Neural Network and Transfer Learning to Bridge Scales for Seismic Phase Picking. *Geophysical Research Letters*, *47*(16), e2020GL088651. https://doi.org/10.1029/2020GL088651

Cleveland, K. M., & Ammon, C. J. (2013). Precise relative earthquake location using surface waves. *Journal of Geophysical Research: Solid Earth*. https://doi.org/10.1002/jgrb.50146

Cleveland, K. M., & Ammon, C. J. (2015). Precise Relative Earthquake Magnitudes from Cross Correlation. *Bulletin of the Seismological Society of America*, *105*(3), 1792–1796. https://doi.org/10.1785/0120140329

Cleveland, K. M., VanDeMark, T. F., & Ammon, C. J. (2015). Precise relative locations for earthquakes in the northeast Pacific region. *Journal of Geophysical Research: Solid Earth*, *120*(10), 6960–6976. https://doi.org/10.1002/2015JB012161

Cleveland, K. M., Ammon, C. J., & Kintner, J. (2018). Relocation of Light and Moderate-Magnitude (M 4-6) Seismicity Along the Central Mid-Atlantic. *Geochemistry, Geophysics, Geosystems*, *19*(8), 2843–2856. https://doi.org/10.1029/2018GC007573

Ekström, G., Nettles, M., & Dziewoński, A. M. (2012). The global CMT project 2004–2010: Centroid-moment tensors for 13,017 earthquakes. *Physics of the Earth and Planetary Interiors*, *200–201*, 1–9. https://doi.org/10.1016/j.pepi.2012.04.002

Ekström, Göran. (2011). A global model of Love and Rayleigh surface wave dispersion and anisotropy, 25-250 s. *Geophysical Journal International*, *187*(3), 1668–1686. https://doi.org/10.1111/j.1365-246X.2011.05225.x

Herrmann, R. B., Ammon, C. J., Benz, H. M., Aziz-Zanjani, A., & Boschelli, J. (2021). Short-Period Surface-Wave Tomography in the Continental United States-A Resource for Research. *Seismological Research Letters*. https://doi.org/10.1785/0220200462

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Howe, M., Ekström, G., & Nettles, M. (2019). Improving relative earthquake locations using surface-wave source corrections. *Geophysical Journal International*, *219*(1), 297–312. https://doi.org/10.1093/gji/ggz291

Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jain, A. K., Jianchang Mao, & Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer*, *29*(3), 31–44. https://doi.org/10.1109/2.485891

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(4), 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Kintner, J. A., Ammon, C. J., Cleveland, K. M., & Herman, M. (2018). Rupture processes of the 2013–2014 Minab earthquake sequence, Iran. *Geophysical Journal International*, *213*(3), 1898–1911. https://doi.org/10.1093/gji/ggy085

Kintner, J. A., Wauthier, C., & Ammon, C. J. (2019). InSAR and seismic analyses of the 2014–15 earthquake sequence near Bushkan, Iran: shallow faulting in the core of an anticline fold. *Geophysical Journal International*, *217*(2), 1011–1023. https://doi.org/10.1093/gji/ggz065

Kintner, J. A., Ammon, C. J., Homman, K., & Nyblade, A. (2020). Precise Relative Magnitude and Relative Location Estimates of Low-Yield Industrial Blasts in Pennsylvania. *Bulletin of the Seismological Society of America*, *110*(1), 226–240. https://doi.org/10.1785/012019163

Kintner, J. A., Cleveland, K. M., Ammon, C. J., & Nyblade, A. (2021). Local-Distance Seismic Event Relocation and Relative Magnitude Estimation, Applications to Mining Related Seismicity in the Powder River Basin, Wyoming. *Bulletin of the Seismological Society of America*, *111*(3), 1347–1364. https://doi.org/10.1785/0120200369

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

Kuang, W., Yuan, C., & Zhang, J. (2021). Real-time determination of earthquake focal mechanism via deep learning. *Nature Communications*, *12*(1), 1432. https://doi.org/10.1038/s41467-021-21670-x

Lay, T., Ye, L., Kanamori, H., & Satake, K. (2018). Constraining the Dip of Shallow, Shallowly Dipping Thrust Events Using Long-Period Love Wave Radiation Patterns: Applications to the 25 October 2010 Mentawai, Indonesia, and 4 May 2018 Hawaii Island Earthquakes. *Geophysical Research Letters*, *45*(19), 10,342-10,349. https://doi.org/10.1029/2018GL080042

Li, Z., Meier, M. A., Hauksson, E., Zhan, Z., & Andrews, J. (2018). Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning. *Geophysical Research Letters*, *45*(10), 4773–4779. https://doi.org/10.1029/2018GL077870

McBrearty, I. W., Delorey, A. A., & Johnson, P. A. (2019). Pairwise Association of Seismic Arrivals with Convolutional Neural Networks. *Seismological Research Letters*, *90*(2A), 503–509. https://doi.org/10.1785/0220180326

Megies, T., Beyreuther, M., Barsch, R., Krischer, L., & Wassermann, J. (2011). ObsPy - what can it do for data centers and observatories? *Annals of Geophysics*. https://doi.org/10.4401/ag-4838

Meier, M. A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., et al. (2019). Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning. *Journal of Geophysical Research: Solid Earth*, *124*(1), 788–800. https://doi.org/10.1029/2018JB016661

Mousavi, S. M., & Beroza, G. C. (2020). A Machine-Learning Approach for Earthquake Magnitude Estimation. *Geophysical Research Letters*, *47*(1), 1–7. https://doi.org/10.1029/2019GL085976

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1), 3952. https://doi.org/10.1038/s41467-020-17591-w

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578. https://doi.org/10.1126/sciadv.1700578

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5), 2894–2901. https://doi.org/10.1785/0120180080

Ross, Z. E., Yue, Y., Meier, M., Hauksson, E., & Heaton, T. H. (2019). PhaseLink: A Deep Learning Approach to Seismic Phase Association. *Journal of Geophysical Research: Solid Earth*, *124*(1), 856–869. https://doi.org/10.1029/2018JB016674

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine Learning Predicts Laboratory Earthquakes. *Geophysical Research Letters*, *44*(18), 9276–9282. https://doi.org/10.1002/2017GL074677

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. de, Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17841-x

Suykens, J. A. K., & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Applied and Computational Harmonic Analysis*, *9*, 293–300. https://doi.org/10.1023/A:1018628609742

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., & Tian, D. (2019). The Generic Mapping Tools Version 6. *Geochemistry, Geophysics, Geosystems*, *20*(11), 5556–5564. https://doi.org/10.1029/2019GC008515

Wessel, Paul, Smith, W. H. F., Scharroo, R., Luis, J., & Wobbe, F. (2013). Generic Mapping Tools: improved version released. *Eos, Transactions American Geophysical Union*, *94*(45), 409–410. https://doi.org/10.1002/2013EO450001

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). Earthquake detection through computationally efficient similarity search. *Science Advances*, *1*(11), e1501057. https://doi.org/10.1126/sciadv.1501057

Zhang, X., Zhang, J., Yuan, C., Liu, S., Chen, Z., & Li, W. (2020). Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-58908-5

Zhang, Z., & Lin, Y. (2020). Data-Driven Seismic Waveform Inversion: A Study on the Robustness and Generalization. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(10), 6900–6913. https://doi.org/10.1109/TGRS.2020.2977635

Zhu, L., Peng, Z., McClellan, J., Li, C., Yao, D., Li, Z., & Fang, L. (2019). Deep learning for seismic phase detection and picking in the aftershock zone of 2008 M7.9 Wenchuan Earthquake. *Physics of the Earth and Planetary Interiors*, *293*(May 2018), 106261. https://doi.org/10.1016/j.pepi.2019.05.004

Zhu, W., & Beroza, G. C. (2018). PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. *Geophysical Journal International*, *216*(1), 261–273. https://doi.org/10.1093/gji/ggy423

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.

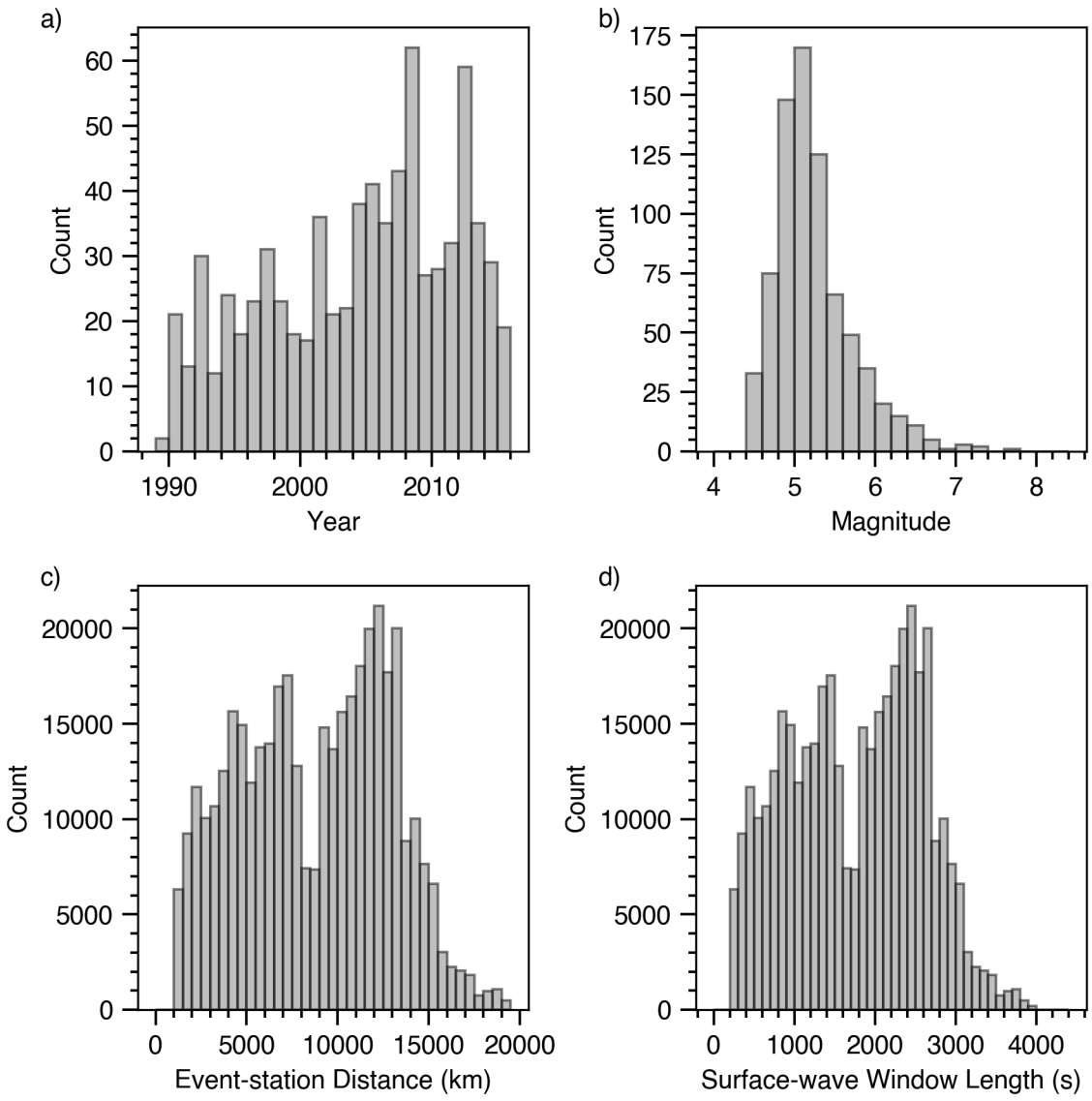

Figure S2. Histograms characterizing the properties of the training dataset DA: (a) origin year of earthquakes; (b) magnitude of earthquakes; (c) the distance between each earthquake and observing seismic station; and (d) the length of surface-wave window defined by a group velocity range from 5.0 to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

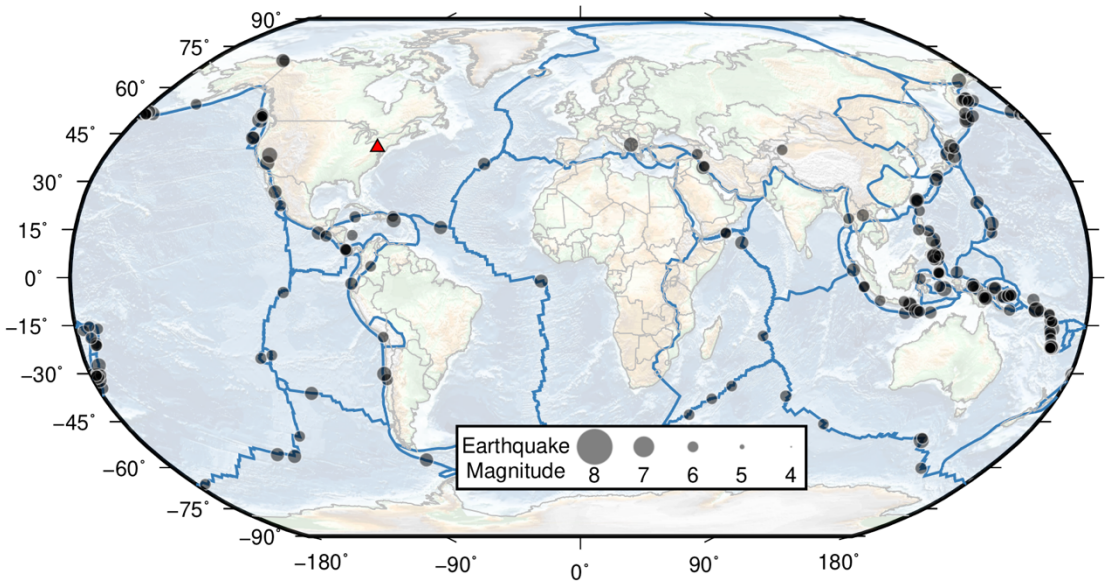

Figure S2. A map of seismic events (gray circles) and the location of seismic station SSPA (red triangle) that were used in the dataset DC.

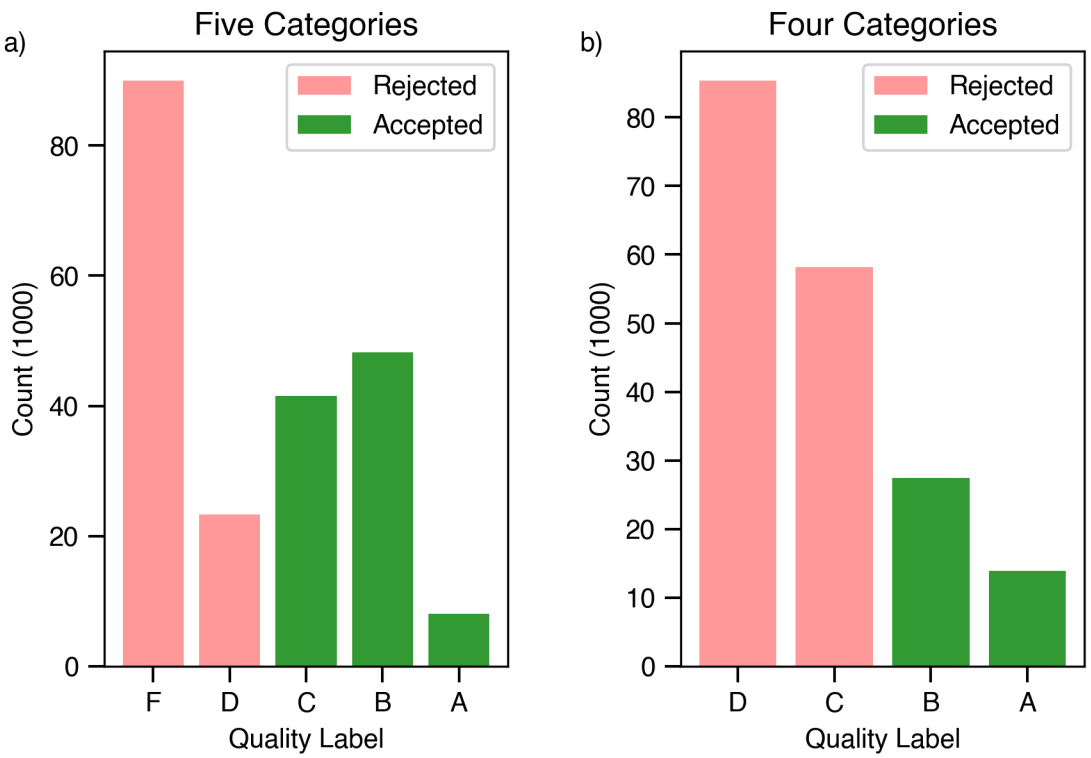

Figure S3. Distributions of original quality labels in dataset DA for (a) five categories and (b) four categories.

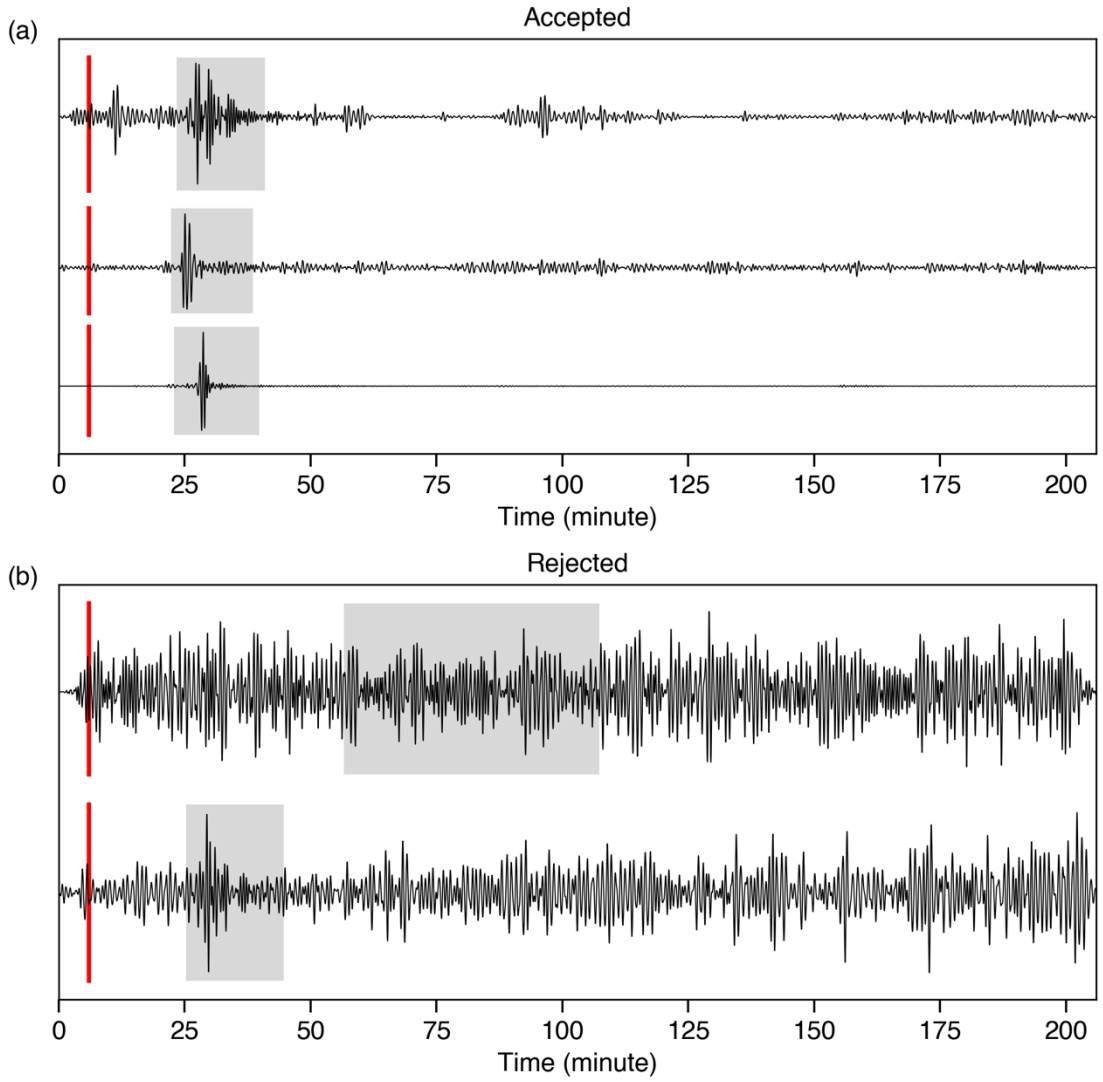

Figure S4. Example displacement waveforms in dataset DA that were (a) accepted and (b) rejected by a human analyst. The red vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity of 2.5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

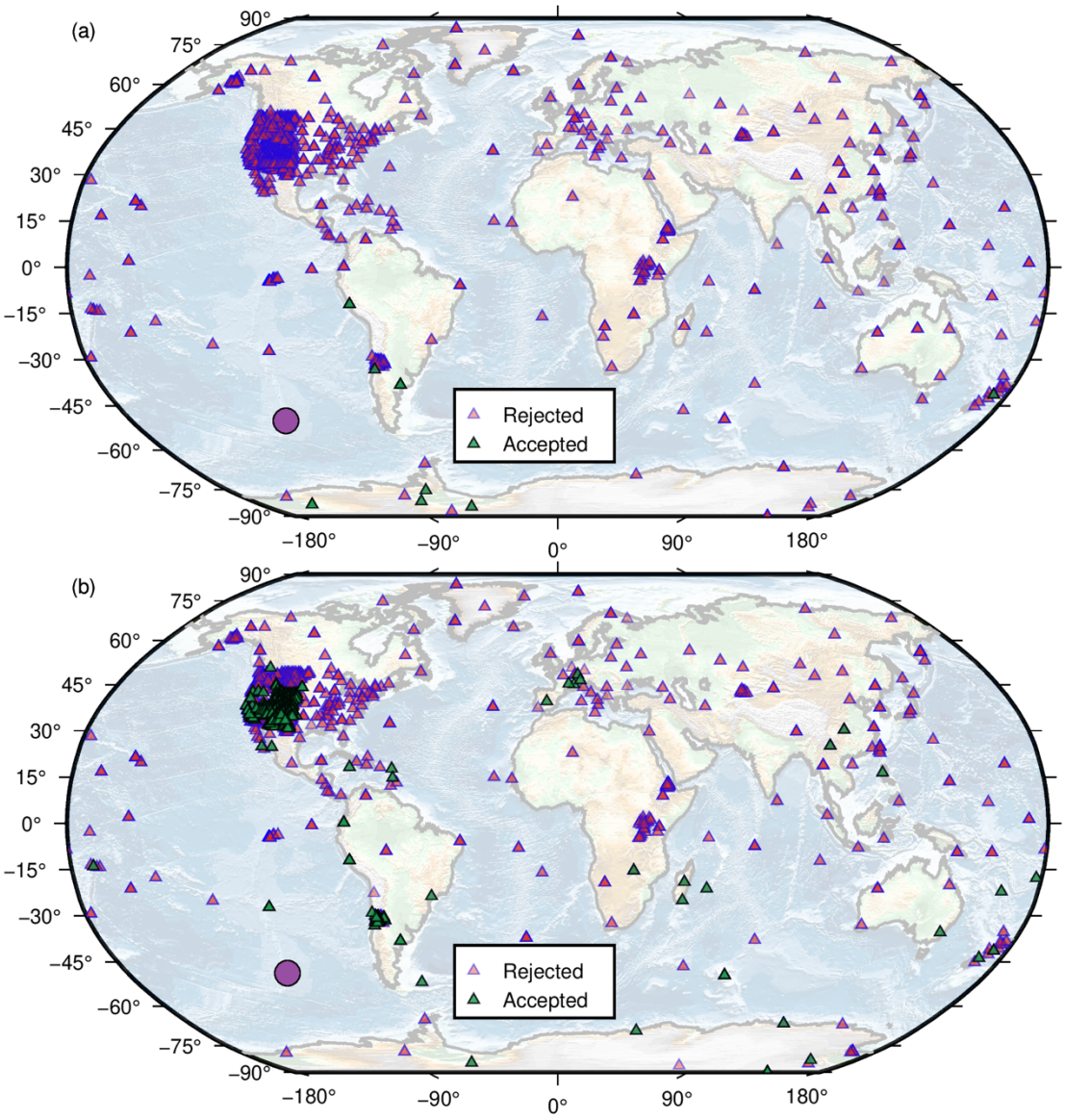

Figure S5. Spatial distributions of quality labels (triangles) for two sample earthquakes (circles) in dataset DA. The event in (a) occurred on 2018/06/12T16:53:34 UTC with a magnitude of 5. The event in (b) occurred on 2018/09/13T15:45:26 UTC with a magnitude of 5.2.

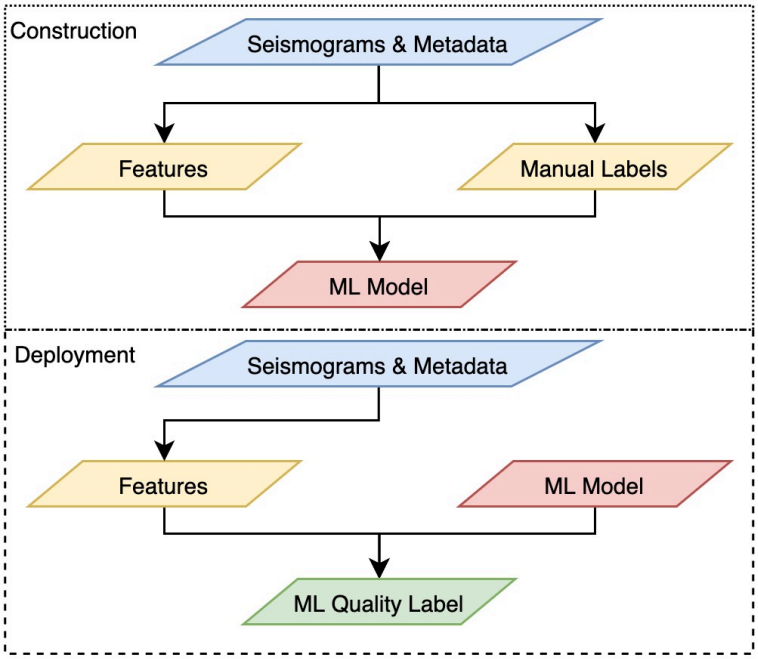

Figure S6. A flowchart illustrating the major steps of the (top) model construction and (bottom) model deployment stages. ML represents machine learning.

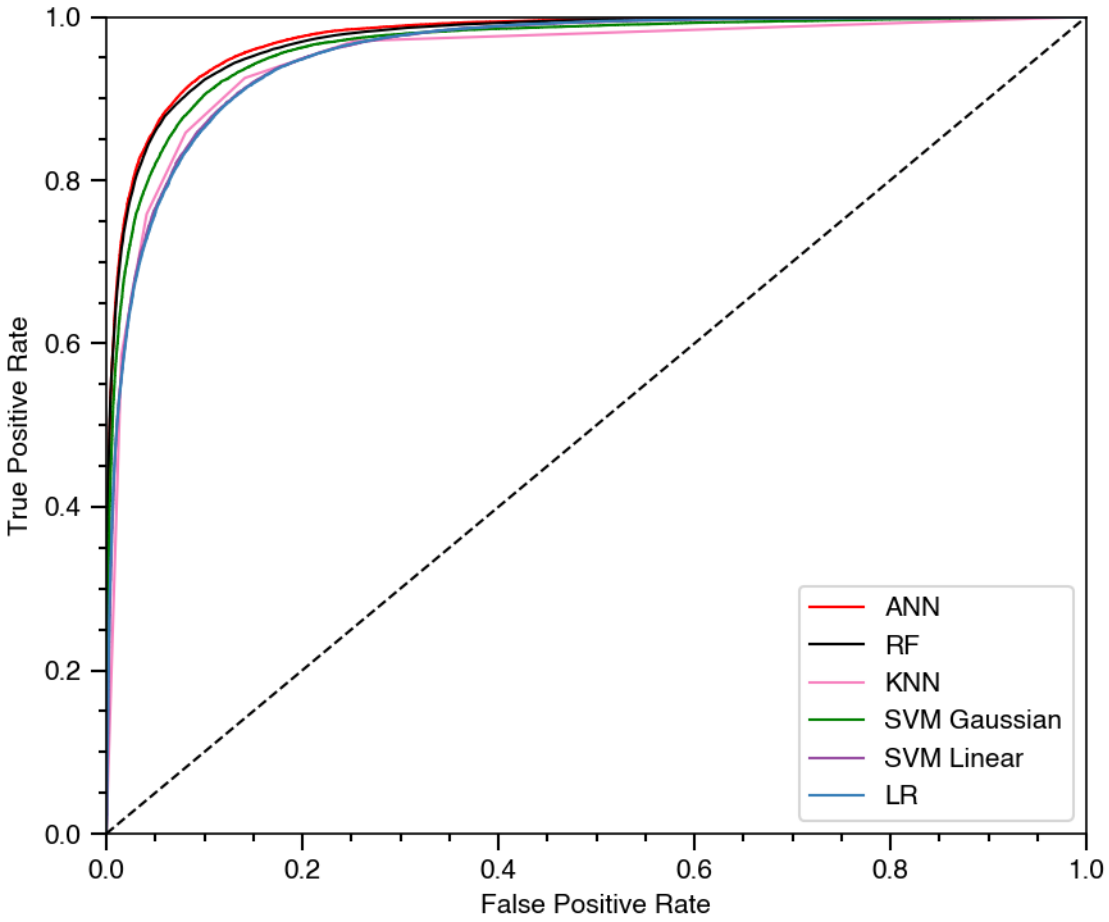

Figure S7. A comparison of the Receiver Operating Characteristic (ROC) Curves for the examined machine learning algorithms constructed using the test set of dataset DA. LR stands for logistic regression. SVM means support vector machine, KNN represents K-nearest neighbors, RF is in short for random forests,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

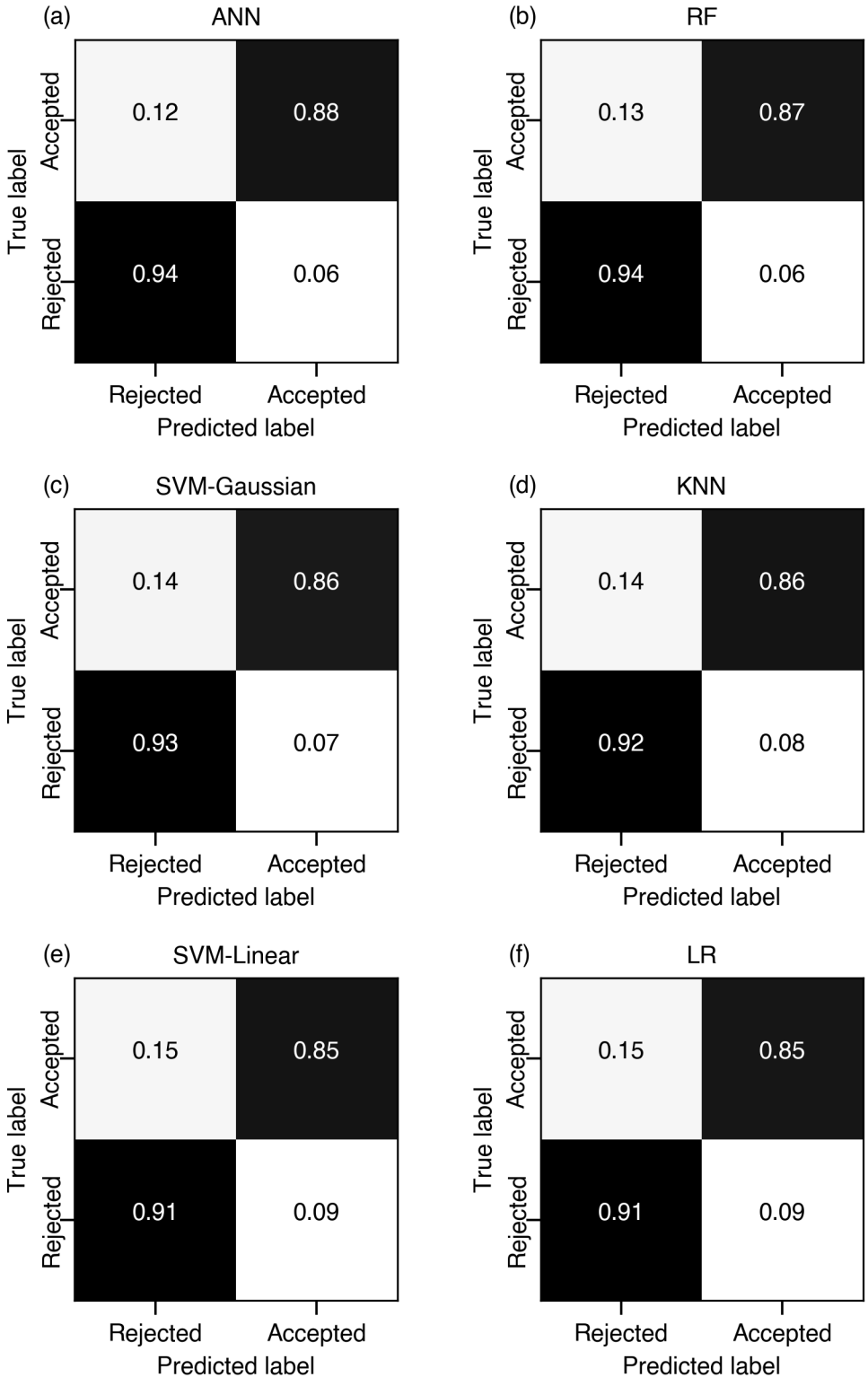

Figure S8. A comparison of confusion matrices for different machine learning algorithms using the test set of dataset DA.

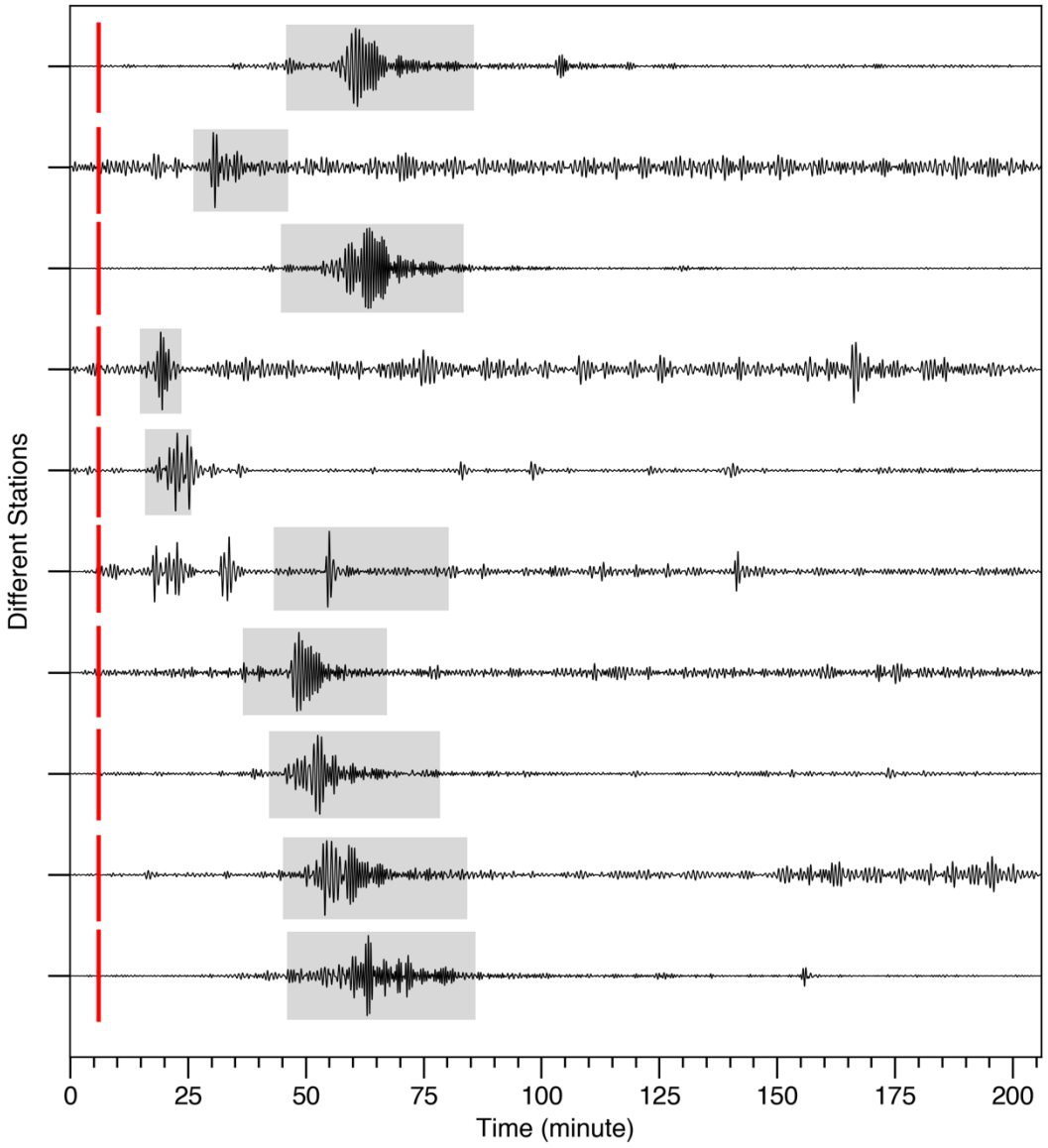

Figure S9. Waveform examples from the test set of dataset DA that were rejected by a human analyst but accepted by the ANN model. The vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity 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.

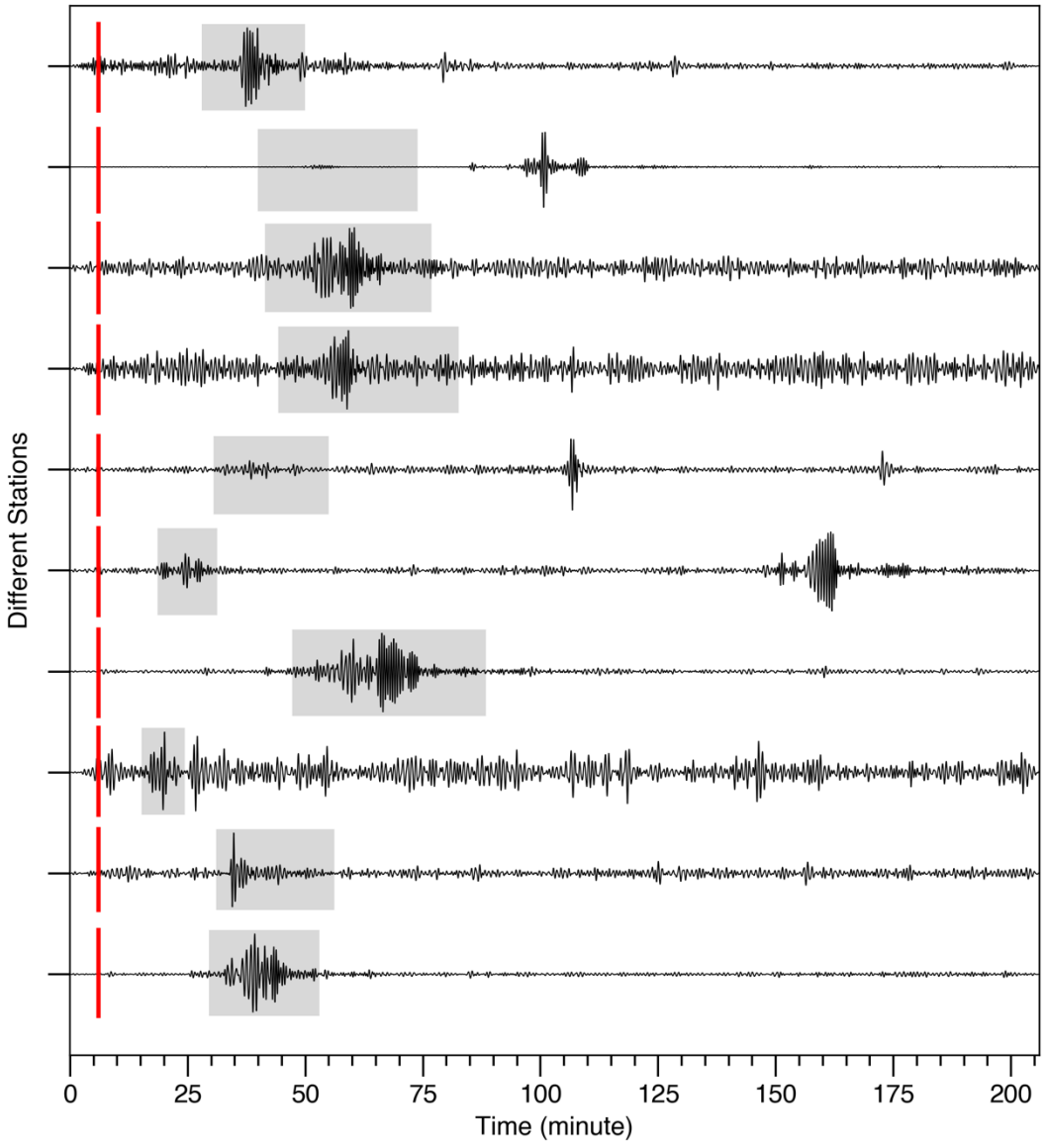

Figure S10. Waveform examples from the test set of dataset DA that were accepted by a human analyst but rejected by the ANN model. The vertical line indicates the origin time of a seismic event. The gray box represents the expected arrival time window of surface waves defined by a minimum group velocity 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.

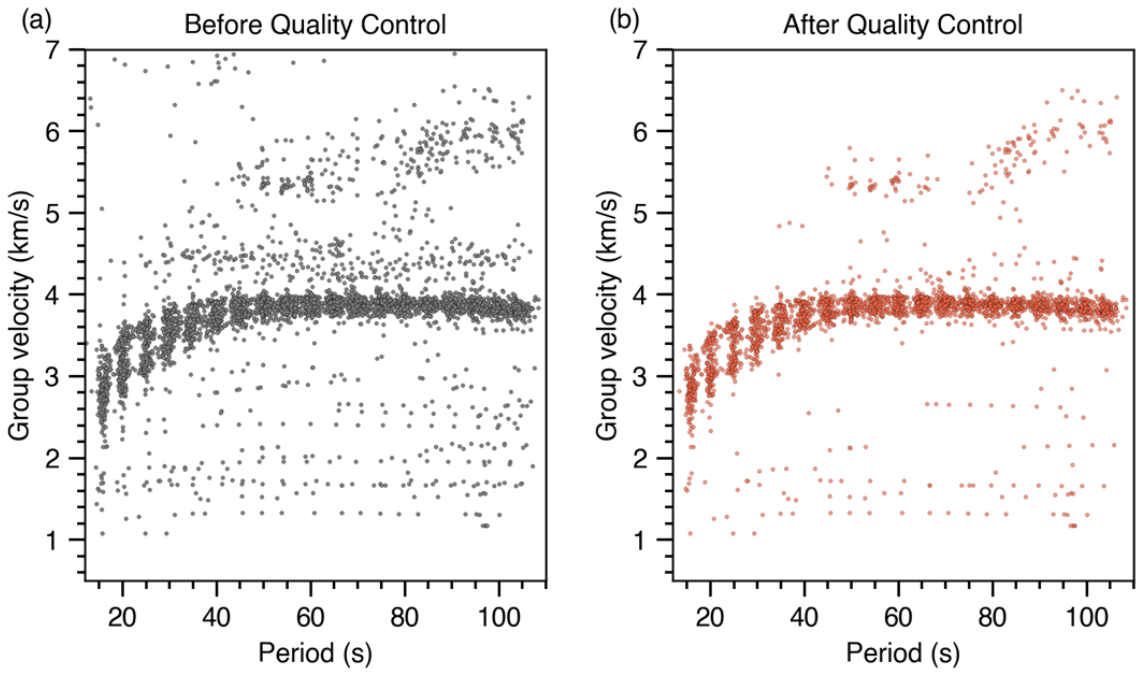

Figure S11. Automatic group velocity measurements (a) before and (b) after using the ANN model for quality control. Automated group velocities are estimated using a simple multiple filter analysis and automated identification of the time of the maximum in a Gaussian-filtered surface waveform. An unrealistic 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.