**NAANMUDHALVAN-IBM SKILL**

**ARTIFICIAL INTELLIGENCE**

**GROUP PROJECT**

**Project Title: Earthquake Prediction model using python.**

**Phase II . Submission**

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Group Members  Name | **Naan Mudhalvan** ID | Email ID |
| 1 | S Abirami | au820321106001 | veeravelabirami@gmail.com |
| 2 | M.Kalaimathi | au820321106019 | muhilvanangu@gmail.com |
| 3 | S.Amirtha | au820321106006 | amirthasaravanakumar04@gmail.com |
| 4 | S.Aishwarya | au820321106003 | aishwaryasivakumar12112003@gmail.com |
| 5 | P.Bavadharani | au820321106010 | dharanibava89@gmail.com |
| 6. | B.Gnanasri | au820321106014 | gnanasribaskar23@gmail.com |

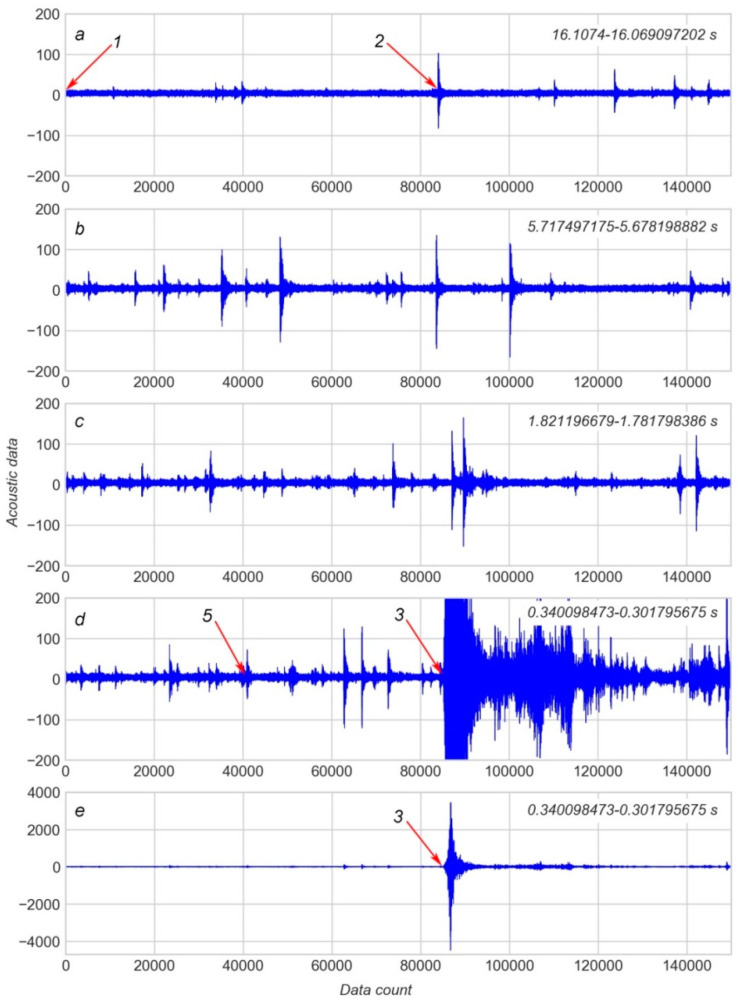
**Data and Methods**

### **Data**

A laboratory experiment that closely mimics real earthquakes is described in [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B12-sensors-20-04228)]. The main idea of the modelling technique is the slow relative motion of rigid, usually steel, plates pressed against each other and separated by a thin granular layer. This granular layer mimics the contact surface of the layer between tectonic plates in which rocks are located. A laboratory quake machine reproduces the stick-slip motion of conjunct plates; acoustic emission from granular gauge interlayer and contact stress values are constantly recorded against the time that remains to failure of the granular layer. These periodic failures are accompanied by a drastic increase in acoustic emission and a drop in contact stress, and are considered to be analogous to real earthquakes [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B26-sensors-20-04228)]. It is recognized that the greater the drop in stress, the more intense the ground motions during real earthquakes [[27](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B27-sensors-20-04228)]. Because a material emits acoustic signals in the course of work and especially before failure, a similar approach may be used for predicting not only real earthquakes but also other types of failures in nature and industry, such as landslides, avalanches, and failure of machine parts [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B12-sensors-20-04228)].

### **Methods**

The training dataset was split into 17 files, each containing an AD and TTF for one separate cycle. The first part of the work was carried out using a file for the longest cycle, i.e., the 8th piece in the training dataset. The first TTF in this cycle was 16.1074 s. [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/) shows several examples of 150 K pieces of ad for this cycle.



The beginning of the cycle is covered by the first 150 K piece ([Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)a). Data provided in [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)b,c illustrates the gradual increase of spikes of AD during the seismic cycle. The piece which contains the EQ event is shown at different levels of magnification in [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)d,e. The role of these spikes is discussed further below. In the final stage of work, modelling was performed using all datasets provided by [[25](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B25-sensors-20-04228)].

The XGBoost library providing the gradient boosted trees approach [[28](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B28-sensors-20-04228)] was used for modelling. It is currently agreed that this technique leads to the best performance compared to other modelling algorithms [[29](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B29-sensors-20-04228)]. For example, XGBoost was used to determine the dominant frequency of an eruptive tremor of the volcano Piton de la Fournaise [[30](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B30-sensors-20-04228)]. XGBoost stands for extreme Gradient Boosting. This algorithm implements an ensemble of decision trees and uses gradient boosting to build models more accurate than the single decision tree or random forest approaches.

Model quality was assessed by mean absolute error (MAE) in 6-fold cross-validation (CV). Cross-Validation is used to estimate model quality by splitting the training dataset into n-folds. One of the folds is used for model validation, while the rest n−1 folds are used for modelling. Modeling is repeated n times, so every fold is used once as a validation dataset. MAE is one of the metrics used to estimate model accuracy. It counts the mean absolute difference between the value of the target parameter (TTF in our case), which is predicted by the model, and the actual parameter value from the validation dataset. Python 3.7 and the necessary libraries, such as pandas, sklearn, and xgboost, were employed to carry out the study.

The main aim of this study was to find an appropriate set of features derived from AD that gives the least MAE in CV. A detailed approach to feature engineering is discussed further below. Since the speed of data processing is critical for the detection and early warning of earthquakes [[15](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B15-sensors-20-04228)] the goal of this work was to determine the feature(s) that are not only useful for building ML models with acceptable accuracy but also enable relatively rapid processing of real-time data.

## **Feature Engineering**

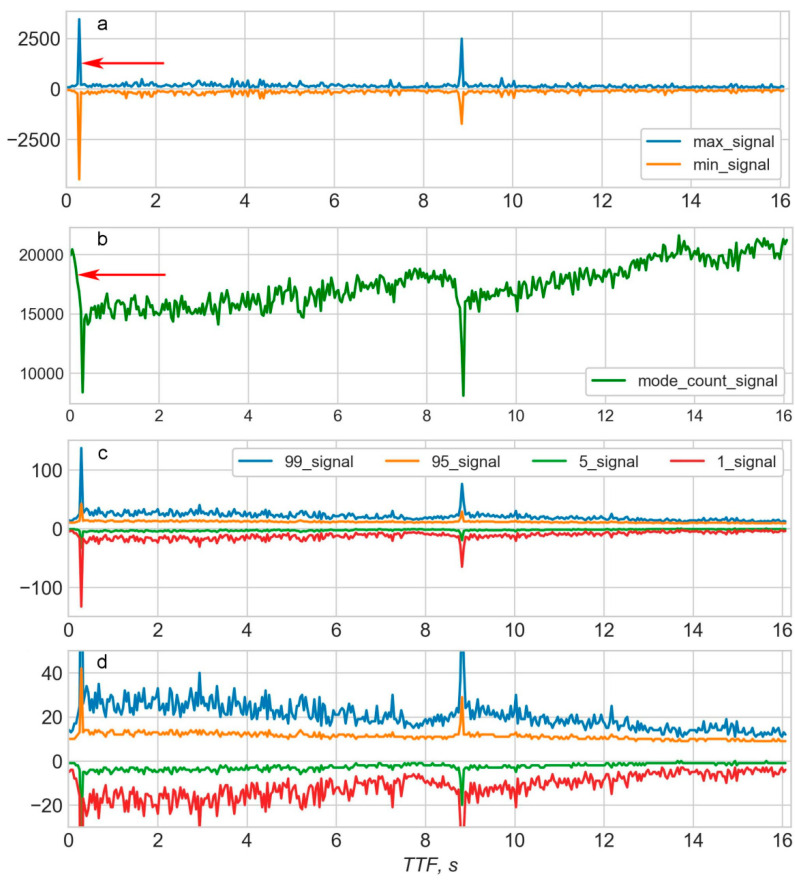
The following key approach was used for feature engineering. In the first step, it is assumed that the distribution of AD is the source of useful features. This assumption is based on “common sense” suggestions, observation of changes in AD distribution over time (see [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)) and also on results published in related works [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B12-sensors-20-04228),26].

It is evident that stick-slip failure (see arrow 3 on Figure 1) is preceded by a number of spikes of AD (see arrow 2 and similar symbols on Figure 1). These spikes appear as a result of micro failure events and may predict TTF [[12](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B12-sensors-20-04228),[13](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B13-sensors-20-04228)]. Generally, the shorter the TTF the more frequent the AD spikes. Hence, the statistical characteristics of AD may serve as features for modelling.

As shown in [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B26-sensors-20-04228)] instantaneous statistical characteristics of AD appear as a “fingerprint” of the fault zone stress state. The variance of the seismic signal is the most important feature, although other statistical characteristics are also important [[26](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B26-sensors-20-04228),[29](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B29-sensors-20-04228),[31](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B31-sensors-20-04228)]. The authors of [[13](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B13-sensors-20-04228)] stressed that the kurtosis of the acoustic signal is an additional powerful feature for the prediction of TTF.

A total of 18 statistical features were derived from each of the 150 K pieces of AD in this work. Nine of these statistical features were maximum, minimum, mean, standard deviation, (standard deviation)/(mean), skewness, kurtosis, mode, number of mode appearance. The remaining nine features were percentiles at the following percent levels: 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th. The “maximum” and “minimum” features were calculated but not used for modelling because these features were only used to indicate the main EQ event due to their outstandingly large values.

Using the initial data sequence from the 8th cycle in the LANL dataset [[25](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B25-sensors-20-04228)], a database of statistical features was created as a result of feature calculation for every subsequent 150 K portion of AD. The database contains 413 rows which cover the TTF range of 16.0691–0.0156 s. [Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f002/) shows several features plotted against TTF. It is evident that at least some of the features correlate with TTF, for example, the number of mode appearances (see [Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f002/)c, “mode\_count\_signal”). Another point to consider is that a certain portion of data is recorded after the EQ at approximately 0.3 s (arrow on [Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f002/)a). Values of any given feature 0.2–0.3 s after an EQ are similar to those long before the EQ in the early stage of a seismic cycle. If values recorded after the EQ were incorporated, then an additional error would appear in the model. This is due to the similarity in feature values at the beginning and the end of a cycle (arrow on [Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f002/)b).



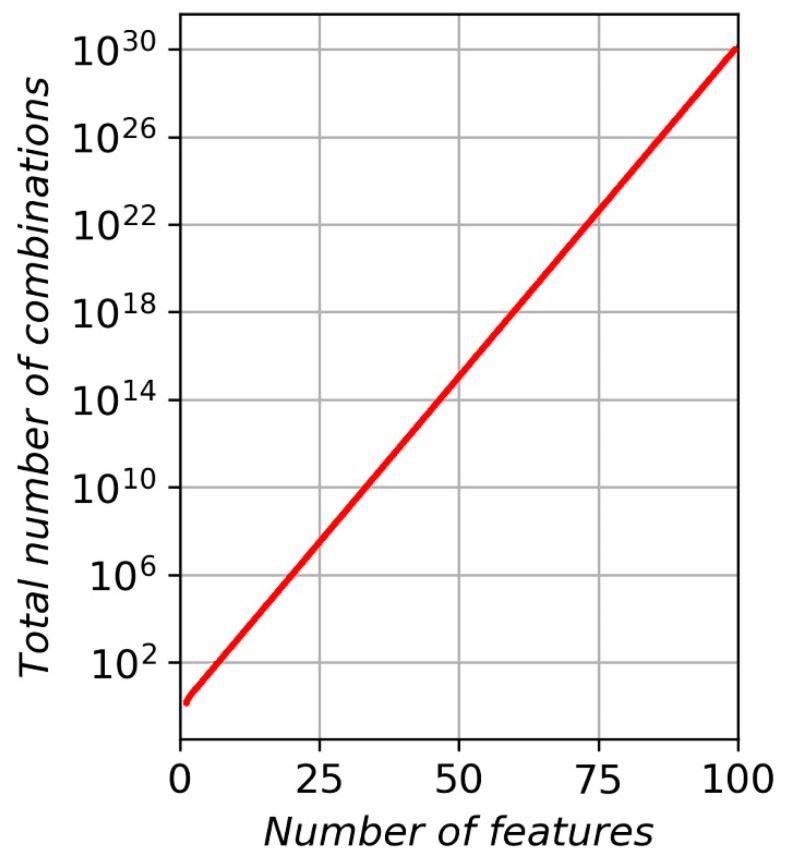
[Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f002/)

Some of the features plotted against time to failure (TTF) for the 8th “earthquake” cycle: (**a**)—maximums and minimums; (**b**)—number of mode appearance; (**c**,**d**)—99th, 95th, 5th, and 1st percentiles.

In order to increase the model accuracy, all tail rows which correspond to the period after the EQ should be deleted from the database of statistical features. Another rationale for deleting data after an EQ is that the main goal of modelling using data from laboratory EQs is to identify features that would be useful to predict real EQs. It is obvious that in reality only data before an EQ would be used for prediction. Any data after an EQ has neither a logical nor practical sense for prediction of that particular EQ.

Due to the development of modelling tools such as Python and appropriate libraries, training a model can be performed rapidly using only several lines of code. The major challenge in training any model is determining which features should be used.

In the current work, the final selection of features was based on the building of different models to compare MAEs and picking the best combination of features that gives the lowest MAE. However, according to the well-known curse of dimensionality, the total number of possible combinations of features increases far faster than the number of features in the set ([Figure 3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f003/)).

[](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=7435601_sensors-20-04228-g003.jpg" \t "tileshopwindow)

[Figure 3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f003/)

The total number of possible combinations versus the number of features.

For example, four features in the set give fifteen possible combinations, 7–127, 10–1023, 15–8191, 16–64,995, 18–262,143, and so on. The “brute force” (BF) method of feature engineering involves sequential modelling and CV score calculation for each combination of features and picking the combination with the least MAE. This method guarantees that the best combination of features is determined. However, BF is time consuming if the number of features exceeds some threshold. In general, the higher the number of features analyzed, the greater the time required to solve the model. For example, a set of 43 features [13] gives a total of 8.796 × 1012 combinations, which would require a significant amount of time to find the best combination.

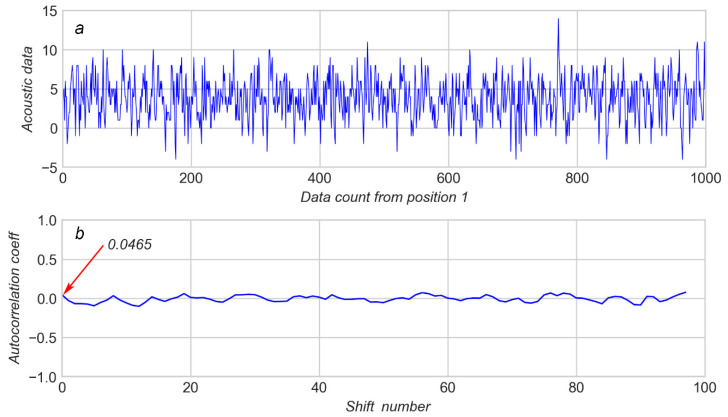
In the paper [13], only two best features were chosen from 43 for prediction of TTF. This means that most of the features are either excessive or not suitable for modelling. Therefore, it can be concluded that the first step in feature engineering is to exclude all non-significant features from the set. Every feature excluded can significantly decrease the total amount of combinations to examine during the BF approach. In our, case excluding only two features decrease the number of combinations from 262,143 to 64,995 (16 features instead of 18).

The “maximum” and “minimum” features can be excluded based on the following reasoning: The maximum values of AD in 150 K pieces is equivalent to the 100th percentile value. This study uses the “99th percentile” feature which is close to the 100th percentile; therefore the 100th percentile (i.e., “maximum”) is superfluous and can be excluded. Similarly, the “minimum” feature is equivalent to the “0 percentile” and can be excluded as the “1st percentile” feature has already been considered. The only reason for calculating “maximum” and “minimum” features is because they are needed for correct identification of the 150 K piece which contains the EQ. It also helps to correctly delete tail rows containing AD after the EQ. After excluding “maximum” and “minimum” features, 16 features remain in the model, giving a total of 64,995 possible combinations.

The backward feature elimination technique (BFE) was employed for reducing the number of features. The rationale behind using this method in the current study is that, if there is a total of n features in a set, then there are n possible combinations of (n−1) features in the subset. Assuming that the vast majority of features are either bad or neutral for model quality, it is highly probable that the MAE for the model—which uses all n features—would be bigger than the least MAE for n models using (n−1) features. If so, then only one model is needed such that it uses all n features, and MAEn is then calculated; thereafter, n models are required, each of which uses one of the possible subsets of (n−1) features. It is also important to choose the subset which results in the least MAEn−1. If a full set of n features contains at least one feature that is bad or excessive for the model, then MAEn would be greater than or equal to the least MAEn−1. This bad or excessive feature should be absent in the subset that generates the model with the least MAEn−1. Thus, this feature can be excluded from the set of features. BFE takes about a minute in semiautomatic mode to exclude one feature and can be fully automated if necessary. BFE is consistently used to reduce the number of features from n to about 10. Thereafter, the straight BF method is used to find the combination of features that gives a model with the least MAE.

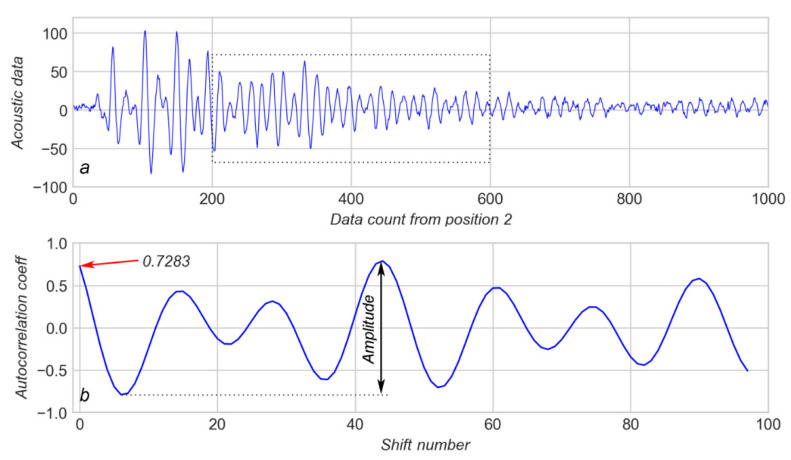
The next important point to consider is how many CV cycles are necessary for every step of the work. Each single CV cycle returns the mean MAE for only six calculations in total. Therefore, the resulting MAE varies at the second decimal point from one run to the other. In order to decrease the variance of MAE the number of repetitions (cycles) of CV should be increased. Two cycles (CV-2) were used for BFE and 500 cycles (CV-500) were used in the final calculation of MAE for the best combination of features. Using CV-500 enables MAEs that are stable to the third decimal point to be obtained.

Figure 4a shows a short portion of 1000 values for the AD–TTF diagram corresponding to arrow 1 on Figure 1a. This portion of data contains no spikes in the AD and the AD distribution seems to be random. Figure 5a shows a short portion of 1000 values for the AD–TTF diagram corresponding to arrow 2 (see Figure 1a); that is, the beginning of the first significant spike observed in the AD. The AD distribution, in this case, seems to be more or less periodic with a gradual increase of random constituents. Spikes in AD are characterized not only by an increase in AD amplitude but also by the grade of AD periodicity.

[](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=7435601_sensors-20-04228-g004.jpg" \t "tileshopwindow)

[Figure 4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f004/)

(**a**): The 1000 K window of acoustic data from arrow 1 (see [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)); (**b**): corresponding autocorrelation coefficients.

[](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=7435601_sensors-20-04228-g005.jpg" \t "tileshopwindow)

[Figure 5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f005/)

(**a**)—1000 K window of acoustic data from arrow 2 (see [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/)); (**b**)—corresponding autocorrelation coefficients.

[Figure 6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f006/)

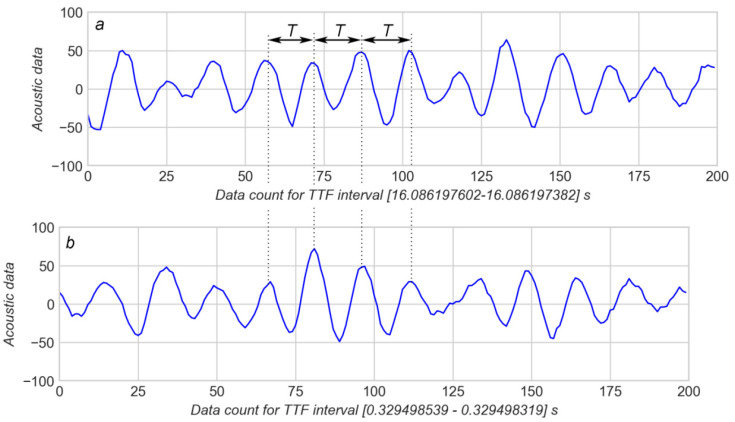
Several first calculations of autocorrelation coefficient (AC) for 13 consequent acoustic data (AD) values corresponding to arrow 2 in [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/).

A total of 98 shifts were used to calculate the ACs for every position of the “sliding window”, containing 1000 values of AD. The results from the calculation of the AC for AD in Figure 4a and Figure 5a are shown on Figure 4b and Figure 5b respectively.

It can be observed that a highly aperiodic AD (see Figure 4a) produces an AC which varies in a very narrow range of about ±0.1 (Figure 4b).

[Figure 7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f007/)

(**a**): “earthquake” event, see arrow 3 on [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f001/); (**b**): 1000 K window of acoustic data from arrow 4; (**c**): corresponding autocorrelation coefficients.

[](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=7435601_sensors-20-04228-g008.jpg" \t "tileshopwindow)

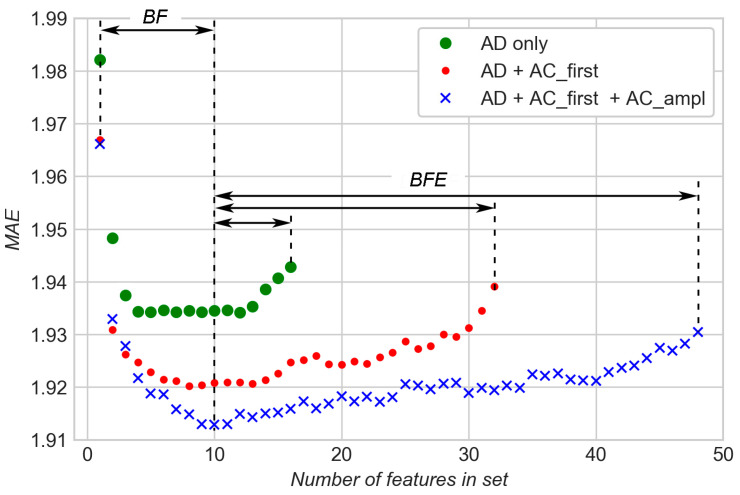
This way, three major parameters were used for modelling which are: acoustic data (AD); the first value of AC on every “sliding window” (AC\_first); the amplitude of AC on every “sliding window” (AC\_ampl). Each 150 K piece of AD contains 150 sliding windows and, therefore, 150 values of AC\_first and 150 values of AC\_ampl. Because 16 statistics were calculated for each of three parameters, the overall number of features considered was 48.

These features were calculated for every separate seismic cycle in the database provided by [[25](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/#B25-sensors-20-04228)]. For every portion containing 150 K of AD, the features were calculated and recorded with the corresponding TTF in the separate file of features. Since the TTF change during 150 K of AD was just 0.04 s, the TTF value is considered a constant during any given 150 K piece. The last value of TTF in the 150 K piece was used as this constant time. Maximum and minimum values of AD were also recorded; they allowed the row that contained the EQ event to be located. All rows after the row with EQ event were deleted according to the reasons explained above. It should be noted that in all files TTF for the EQ event was approximately the same (near 0.3 s).

Finally, all separate files with features were merged into one database.

**Results and Discussion**

[Figure 9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f009/) shows the best MAE obtained for any given number of features in the subset of features. Three sequences of dots represent three sets of features: 16 features for AD only; 32 features for AD+AC\_first; and 48 features for AD+AC\_first+AC\_ampl. Each sequence of dots is a function “MAE vs. Number of features” for the corresponding set of features.

[](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=7435601_sensors-20-04228-g009.jpg" \t "tileshopwindow)

[Figure 9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7435601/figure/sensors-20-04228-f009/)

Comparison of modelling results for a different combination of features.

It is evident that in each of the three sets of features there are good, excessive (neutral), and bad features. During BFE, bad features were gradually removed from subsets of features and “MAE vs. Number of features” functions slowly decreased for each of three sets with the decreasing number of features. As the number of features reached 10, the BF method was used to find the best combination of features that gave the least MAE. It can be seen that MAE decreases significantly for each of the three functions in the early stages of the increasing number of features during the BF stage. This means that useful features are gradually added to the subsets. After a certain number of features, MAE stabilizes and minimums are reached. Addition of more features does not lead to a decrease of MAE, so these additional features are excessive (neutral) for modelling.