Auto-REP: an automated regression pipeline approach for high-efficiency earthquake prediction using LANL data

Yang, F.; Kefalas, M.; Koch, M.; Kononova, A.V.; Qiao, Y.; Bäck, T.H.W.

2022 14th International Conference on Computer and Automation Engineering (ICCAE) | 978-1-6654-8380-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCAE55086.2022.9762437

Auto-REP: An Automated Regression Pipeline Approach for High-efficiency Earthquake Prediction Using LANL Data

Fan Yang
Xi'an Jiaotong University
Leiden University
Xi'an, China; Leiden, The Netherlands
f.yang@liacs.leidenuniv.nl

Marios Kefalas
Leiden University
Leiden, The Netherlands
m.kefalas@liacs.leidenuniv.nl

Milan Koch
Leiden University
Leiden, The Netherlands
m.koch@liacs.leidenuniv.nl

Anna V. Kononova
Leiden University
Leiden, The Netherlands
a.kononova@liacs.leidenuniv.nl

Yanan Qiao
Xi'an Jiaotong University
Xi'an, China
qiaoyanan@mail.xjtu.edu.cn

Thomas Bäck
Leiden University
Leiden, The Netherlands
t.h.w.baeck@liacs.leidenuniv.nl

***Abstract*—Earthquake prediction, which is a key issue that has long existed among seismologists, is of high scientific importance. An earthquake prediction model can output the time of earthquake occurrence in advance using machine learning methods, which is receiving increasing attention. Earthquake prediction involves a large variety of data mining steps, which requires a large amount of time for processing data and model development. Thus, an efficient and accurate prediction method is needed. Aiming to solve this problem, we propose Auto-REP, an automated machine learning-based regression model. Our contribution of Auto-REP is using laboratory seismic data to develop a regression pipeline in an automated manner, and eventually obtain the prediction results of laboratory earthquake occurrence. The automated pipeline consists of feature extraction, feature selection, modelling algorithm and optimization. With this approach we extract features from each of the earthquake channels which results in a massive feature space. The hyperparameters of the model are optimized by a Bayesian technique as part of the automated approach. The experimental results shows that the MAE and MSE of our model in the training and testing datasets are 1.48, 1.51 and 1.52, 1.59. The results demonstrate that our Auto-REP method can predict laboratory earthquakes efficiently and accurately.**

***Keywords*—Automated machine learning, Earthquake prediction, Time series regression, Hyperparameter optimisation.**

I. Introduction

The increasing frequency of major natural disasters in recent years is having a serious impact on human society, causing irreversible damage as well as sounding a clear warning for the achievement of the UN Sustainable Development Goals by 2030 [1]. The outbreak of the COVID-19 epidemic in 2020 has created a tremendous global impact. At the history of world development, natural disasters, disease epidemics and other catastrophes have always accompanied human society. In addition to epidemics, earthquakes, one of the most powerful natural disasters, have brought huge economic and property damages to human society, causing also a substantial amount of loss of life. For example, in 2018, there were 119 earthquakes of magnitude 6 or higher worldwide, including 18 earthquakes of magnitude 7 or higher and 2 earthquakes of magnitude 8 or higher (in the Gulf of Alaska and in the Fiji Islands), resulting in enormous loss of life and property damage [2]. The sudden occurrence of earthquakes often results in a failure to respond in time, bringing the loss of large quantities of materials and equipment and even a large number of casualties. Therefore, earthquake prediction has a significant and practical application value.

The era of Big data has arrived with the rapid expansion of data accumulated in all areas of society, including scientific research, production and consumption, at an unprecedented rate. The integration of Big data and machine learning techniques for solving practical problems in computer and automation engineering has gained increasing significance [3]. As stated in a review article published by Science [4], machine learning has been one of the most rapidly developing areas of information science and technology. In recent years, outstanding international research results in the field of earthquakes have generally been produced in this context, i.e., in interdisciplinary research combining earthquake precursors with artificial intelligence techniques. Therefore, with the advancement of deep learning and machine learning research, seismologists have recently been motivated to use computer and automation engineering methods to analyze seismic data and build models for earthquake prediction [5].

Current earthquake prediction research presents several challenges. First, the urgent tasks that models need to address is to accurately predict the time to event of an earthquake, thus avoiding the disaster generated by the earthquake in time. Second, earthquake prediction models need to minimize human intervention. Finding the optimal set of hyperparameter by hand is nearly impossible, therefore using automated

pipelines can result in optimal hyperparameters and as a result even better performances. Traditional methods require a large number of human settings of finding hyperparameters, the model efficiency is low. Therefore, each aspect of machine learning needs to be integrated so as to improve the automation of the model. Finally, earthquake prediction is a research task that requires high rapidity, and the model needs to be able to minimize unnecessary computation times. Although LANL seismic predictions use laboratory seismic data, LANL earthquake predictions have the same seismic physical properties as the prediction studies of real earthquakes and can provide an essential methodological reference in actual earthquake prediction. Therefore, in this work, the LANL (Los Alamos National Laboratory) dataset [6] was used to validate Auto-REP that combines feature extraction, feature selection, hyperparameter optimisation (HO), and regression modelling. Auto-REP is considered as an automated machine learning pipeline with the input data being time series and the output being the predicted time to event.

The overall structure of the paper is divided into five sections. Section II includes a literature review of current research on earthquake prediction. Section III deals with the methods used for this analysis. Section IV analyzes the experimental results of the regression pipeline and the comparison with other commonly used approaches. Finally, Section V summarizes the contribution of this paper and discusses possible future studies.

II. Related Work

Currently, many researchers have started to explore data mining and machine learning methods for earthquake prediction. Rouet et al. [7] came to the conclusion that random forests are effective for earthquake predictions, and offered preliminary results on successfully applying methods developed on lab data to field data on a specific type of earthquake known as slow earthquakes. Lubber et al. [8] investigated how machine learning regression from a laboratory earthquake event database performs as a function of completeness for earthquake catalogs , and then used lab data to predict earthquake magnitudes using random forests. Huang et al. [9] proposed a feature engineering framework for short-term earthquake prediction based on the Multi-Component Seismic Monitoring System precursor data and historic seismic events. In addition, the daily mean value of geo-sound time-series data is calculated as a feature. With this framework the researchers discovered a novel feature, the daily mean value of the geo-sound. Bray et al. [10] employed residual methods to assess the goodness of fit of earthquake forecasting models. Ogata et al. [11] utilized residual analysis to identify features in earthquake datasets. The pre-processing of the waveform is an essential aspect of seismic signal analysis. Traditional methods usually require manual frequency domain analysis of the waveform. Denoising through Robust Principal Component Analysis (R-PCA), as described in [12], [13], and dictionary learning [14] are two additional waveform approaches. Ross et al. [15] presented PhaseLink, a system for grid-free earthquake phase association based on recent developments in deep learning. The proposed method was trained entirely on millions of synthetic sequences to connect phases that share a common origin. For representation learning and earthquake simulation, Holtz et al. [16] used convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and their joint design. They also tested the Wavenet technique for audio processing in earthquake datasets. To learn the characteristics of first-arrival earthquake waves, Li et al. [17] trained a generative adversarial network (GAN). The results indicate that GANs can explore a compact and effective representation of seismic waves, which has the great potential for future seismological applications. Liu et al. [18] proposed a spatial-temporal data analysis-based earthquake prediction study. The results suggest that the strategy proposed in this research can take use of earthquake spatial-temporal correlations to improve predictions than before. In order to provide a clear presentation of the existing research using machine learning methods, Table I presents an overview of earthquake prediction research methods.

TABLE I
AN OVERVIEW OF EARTHQUAKE PREDICTION RESEARCH METHODS. (WHERE HO REFERS TO HYPERPARAMETER OPTIMIZATION)

Year	Study	Method	Whether HO is used?	Is the method automated?	Whether dealing with laboratory data?
2013	Bray et al [10]	Residual methods	✗	✗	✗
2018	Li et al. [17]	GAN	✗	✗	✗
2018	Lubber et al [8]	RF	✓	✗	✓
2019	Rouet et al. [7]	RF	✓	✗	✗
2019	Ross et al. [15]	Deep learning	✓	✗	✓
2019	Holtz et al. [16]	LSTM	✓	✗	✓

Research in automated machine learning pipelines is currently also receiving increasing attention from academics. Bruckner et al. [19] presented ML-o-scope, an interactive visualization system for exploratory analysis of convolutional neural networks, a popular pipelined model type. The ML-o-scopes time-lapse engine, which provides views into model dynamics during training, is presented, and the system is evaluated as a tool for tuning large-scale object-classification pipelines. Estevez-Velarde et al. [20] offered a new theoretical and practical perspective on AutoML. The proposal is tested in a variety of machine learning problems and compared to alternative approaches, demonstrating that it is competitive in standard benchmarks with other AutoML alternatives. Furthermore, it can be used in novel scenarios where existing alternatives are ineffective, such as several NLP tasks. Kathiravelu et al. [21] proposed Niffler, an integrated framework that allows researchers to run machine learning pipelines at research clusters by efficiently querying and retrieving radiology images from hospitals' Picture Archiving and Communication Systems (PACS). Niffler retrieves and stores imaging data using the Digital Imaging and Communications in Medicine

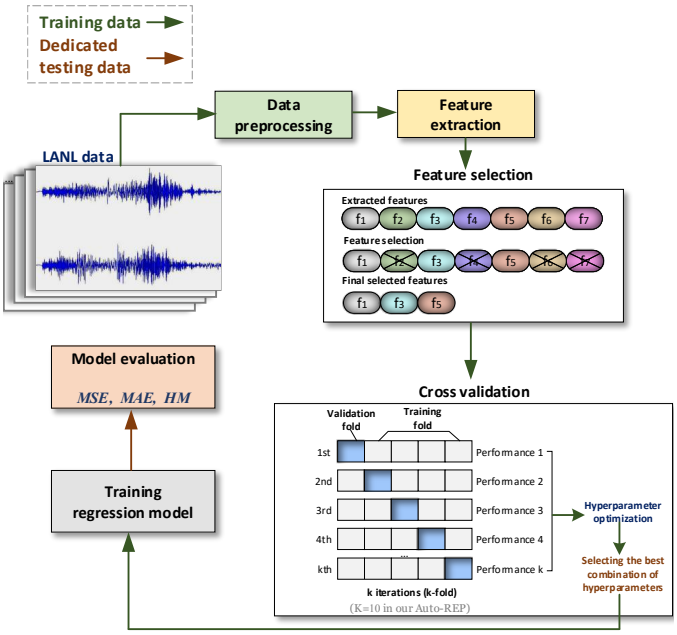


Fig. 1. Workflow of Auto-REP. The whole pipeline consists of six main parts: data pre-processing, feature extraction, feature selection, cross-validation, model training and model evaluation. The incoming seismic time series is used by Auto-RE to obtain the time-to-event of the seismic event and to evaluate the model performance.

(DICOM) protocol, as well as providing metadata extraction and application programming interfaces (APIs) for applying filters to the images. The above research revealed that automated machine learning can deliver an effective and reliable solution to classification and prediction problems.

It can be seen from the existing research that machine learning algorithms have been gradually applied to the study of earthquake prediction, and have achieved promising prediction results. However, the existing research is still focused on the seismic prediction model itself, and each part of machine learning does not integrate effectively. In addition, for a large number of non-specialists, using the methods currently studied, requires a lot of knowledge in data pre-processing, feature selection, and hyperparameter settings. This prohibits non-specialists in using the models for earthquake prediction quickly and accurately. To address these issues, our proposed Auto-REP method allows data to flow through the pipeline through an automated machine learning approach. The data are sequentially processed through feature extraction, feature selection, modelling, hyperparameter selection, and model evaluation, which significantly minimises human intervention in seismic prediction methods. Our Auto-REP consequently improves the efficiency of the model and increases the automation of the earthquake prediction method, providing an enhanced, integrated solution for earthquake prediction.

III. Methodology of Auto-REP

Due to the devastating consequences of earthquakes, forecasting them is a critical priority in Earth science. Current scientific research on earthquake forecasting focuses on three key points: estimating the location, the time of occurrence and the magnitude. Our main goal in this research is to use an automated machine learning method to predict the time-to-event of an earthquake and thus provide a comprehensive solution for ascertaining when the earthquake will occur.

The workflow of the proposed automated machine learning regression model-based approach to earthquake prediction is shown in Fig.1. The steps of this workflow are discussed in the sections below.

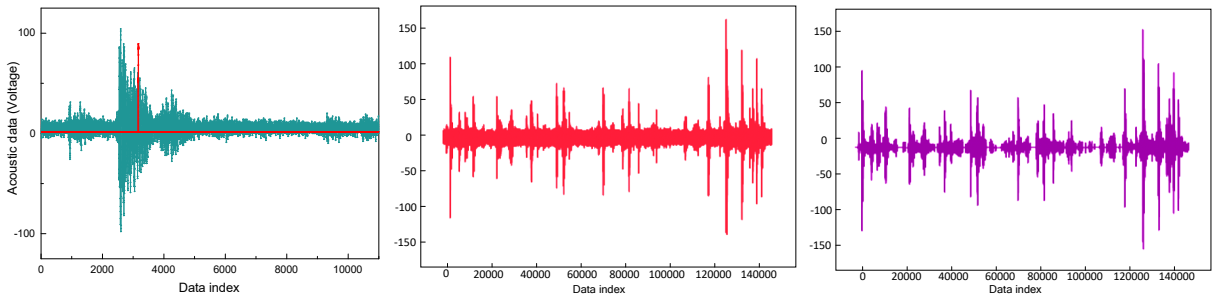
A. Data Preprocessing

The main purpose of our analysis of seismic signals is to predict the time left for the next earthquake to occur. However, seismic data are often accompanied by a large amount of noise [22], which may also have an impact on the performance of the prediction model. Therefore, we need to pre-process the seismic dataset first in order to make our pipeline have a stable and efficient prediction performance. In the pre-processing stage we mainly need to perform noise reduction on the data. The seismic data often contain a large amount of random noise, which greatly affects the quality of the data and leads to poor model recognition, so noise reduction is required to improve the signal-to-noise ratio and remove the obvious noise fluctuations in the LANL dataset. In this paper, we primarily use the wavelet noise reduction method [23], which is also provided by Kaggle, to pre-process the data set. Fig.2.(a) presents segment of LANL training data. In Fig.2.(a), the x-axis represents the data index of the intercepted training data set, and the y-axis represents the acoustic signal value corresponding to the sample points. The red line where the sudden fluctuation occurs indicates the acoustic data value at the time of the earthquake. The signal plots before and after pre-processing are shown in Fig.2.(b) and Fig.2.(c). It can be seen that after pre-processing, many noisy signals have been removed from the dataset.

B. Feature Extraction

We intend to create understandable and computationally efficient regression models in this paper by automatically generating and selecting time series features. Rather different features are important depending on the domain and the seismic data analysis. Traditionally, such features are designed by hand from time series, which requires technical expertise and expert knowledge. In our work, feature extraction is designed to obtain a large number of time series features. and then select the most appropriate features for the overall earthquake prediction task in order to establish a systematic and automated approach to generating features for seismic series. During the feature extraction process, we use the predefined parametric features functions, e.g., autocorrelation, kurtosis, and skewness from the `tsfresh` package[1]. After we get the extracted features for earthquake prediction task, and the features will go through the feature selection procedure.

[1]tsfresh version 0.18.0



(a) Segment of LANL training data (b) Seismic signal without pre-processing (c) Seismic signal after pre-processing

Fig. 2. A demonstration of LANL training data segment and pre-processing approach. where the unit of y-axis is the voltage value of "Acoustic_data" in LANL datasets and its SI unit is (mV) . The red line where the sudden fluctuation occurs indicates the acoustic data value at the time of the earthquake.

C. Feature Selection using Boruta

Feature selection is the method of selecting the most efficient features from the original features to reduce the dimensionality of the dataset and the computational overhead, thus improving the regression performance. Since we have to select features extracted from a large quantity of time series signals in the seismic dataset, it is challenging to select the most suitable combination of features for the regression model without a comprehensive evaluation of the statistical features of each seismic signal. We therefore adopt the Boruta method to evaluate all extracted features and select the most suitable features for the regression model [24].

The core idea of the Boruta algorithm is to evaluate the importance of each feature variable through a loop method [24]. By copying the original feature set, randomly mix each feature value to construct a random shadow feature. The final sample data set of the model is original feature and shadow feature. The primary reason we chose Boruta as the feature selection method is that the Boruta algorithm is based on the idea of a random forest classifier and incorporates randomness into the system by collecting the results of different feature combinations from a random set of samples. The randomness of Boruta will provide us with a clearer picture of the results of feature selection, providing us with a solid understanding of which features are truly important and thus keeping those significant features.

D. Modelling

During the modelling stage, the features that were chosen are used to train a random forest (RF) regressor model. The main reasons for our decision to choose random forest regressor as the regression model are as follows:

- Due to the relatively large training samples of seismic data, random forest regressor is more suitable than shallow neural networks for regression modeling of large-scale datasets [25].
- Random forest regressor introduces randomness, which is less likely to cause overfitting problems, and has better noise immunity, which is suitable for handling acoustic data containing partial noise signals [26].

Therefore, due to the advantages of random forests, which are less prone to overfitting, faster training and simpler to implement, we adopted the random forest regressor to introduce into our Auto-REP.

E. Hyperparameter Optimization

The proper configuration of hyperparameters is critical to the performance of a machine learning model [27]. As a result, throughout this research, hyperparameter optimization is required to improve the performance of the prediction model on seismic dataset. Grid Search [28], Evolutionary Algorithms [29], and Bayesian Optimization [30] are some of the well-established methods for prediction task [31] The Bayesian Optimization algorithm is chosen for this paper due to its efficiency when optimizing expensive problems, such as training a machine learning algorithm, which can be time consuming. Grid search is also a simple, efficient, and widely used method in hyperparametric optimization problems [32]. Another reason we adopt Bayesian optimization is that Bayesian optimization uses a Gaussian process that takes into account previous parameter information, while grid search fails to consider previous parameter information. In addition, Bayesian conditioning is fast with fewer iterations, but grid search is slow and prone to dimensionality explosion when there are many parameters.

The hyperparameter optimization procedure is used to enhance model performance in a 10-fold cross validation configuration, where the dataset is divided into ten parts, with nine of them rotating as training data and one as test data. Table II presents the hyperparameter optimization results under ten times of experiments in our pipeline. The primary reason for choosing 10-fold is that a large number of experiments using different learning techniques with a large dataset have demonstrated that 10-fold is an appropriate choice for obtaining the best error estimates while ensuring efficient training as well as not causing redundant time and space consumption [33], [34].

IV. Experiments and Analysis

Our experiments have been designed to answer the following two research questions regarding our proposed method:

TABLE II
HYPERPARAMETER OPTIMIZATION RESULT

Hyperparameter	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Is bootstrap used for training samples	True	True	True	True	True	True	True	True	True	True
Max number of features when splitting a node	auto	auto	auto	auto	auto	auto	auto	auto	auto	auto
Max depth of each tree	6	8	7	9	8	11	9	10	9	8
The number of n_estimators	900	800	700	600	900	1000	800	900	900	1000
The number of min_samples_split	16	18	15	18	14	17	16	16	18	19
The number of min_samples_leaf	7	10	8	9	7	10	7	9	7	9

- **RQ1:** How is the performance of our proposed Auto-REP method compared to other commonly used approaches?
- **RQ2:** What advantage can be gained by solving the earthquake prediction problem using an automated machine learning method compared to any existing competitive regression algorithm for the harmonic mean of efficiency.

A. LANL Dataset

LANL dataset is signal data of earthquake occurrence, and these signal data are used to predict the time-to-event of an earthquake. The acoustic signal is emitted by the movement of a fault block in the earth's crust along a fault and is considered to be the main cause of earthquakes. As such acoustic signals can usually be captured long before an earthquake occurs, machine learning techniques are able to discern specific patterns in these acoustic signals. Based on these characteristics, machine learning techniques can assess how much pressure the fault is under and how long before it ruptures, which ultimately provides a more accurate prediction of whether an earthquake will occur. LANL datasets for training and testing are publicly available on the Kaggle website[2]. Kaggle provided the testing data in the form of 2624 sequential segments. Each segment contains $150,000$ acoustic signal data points. The data all comes from a well-known earthquake physics experimental setup.

B. Performance Evaluation

For model evaluation, we used MSE (Mean Squared Error and MSE (mean absolute error) [35] as regression metrics.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{1}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m} \mid (y_i - \hat{y}_i) \mid \tag{2}$$

In Equations (1) and (2), m denotes the number of predicted TTE(time-to-event) value , $\hat{y}_i$ indicates the i-th predicted TTE value after regression procedure, and y_i represents the i-th actual TTE value.

First, Fig.3 and Fig.4 demonstrate the results of feature selection using Boruta. Among them, Fig.3 exhibits the distribution of different selected feature frequencies and Fig.4 presents the heat map results for the ranking and number of top-10 selected features under 10 times experiments. The models were used over the dataset using a 10-fold technique to evaluate performance for hyperparameter optimisation. The original data are divided into 10 groups, and each subset of data is made into a validation set separately, and the remaining subsets of data are used as training sets, and the final errors are summed and averaged to get the errors. The MSE and MAE of different methods are calculated separately to compare the performance of regression models.

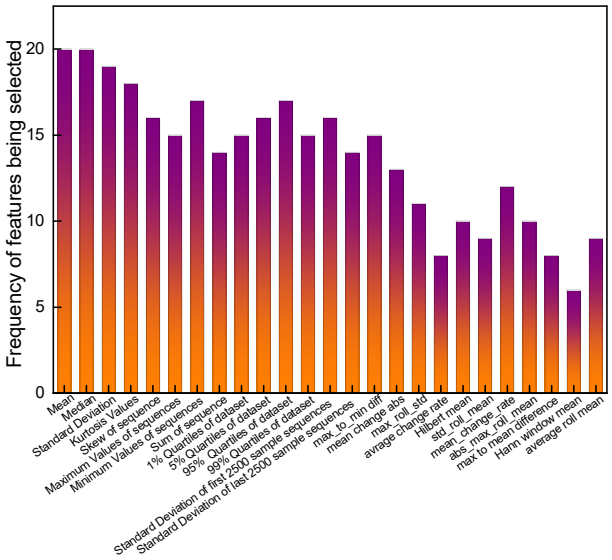


Fig. 3. Distribution of different feature frequencies using Boruta (20 feature selection experiments)

C. Comparison of Model Results

In order to compare our automated machine learning pipeline approach with the currently available regression methods [36], we list the models to be compared for testing and comparing the performance of our proposed pipeline as follows.

- **PCA+RF**: Principal Component Analysis (PCA) is one of the most widely used algorithms for data dimensionality reduction. Random Forest (RF) uses bootstrap sampling to collect several different sub-training datasets from the input training dataset to train several different decision trees in turn; in the prediction phase, the random forest

[2]https://www. kaggle.com/c/LANL-Earthquake-Prediction/data

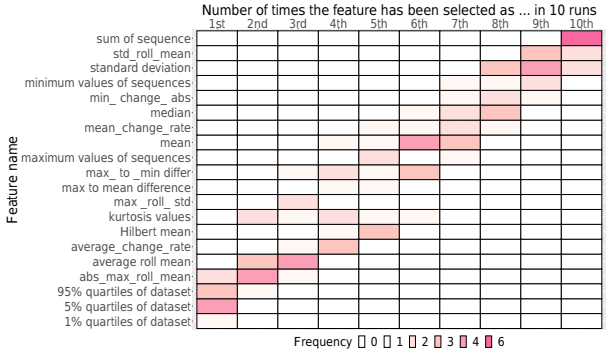


Fig. 4. Heat map results for the ranking and number of different features selected within 10 independent runs of Auto-REP

averages the prediction results of multiple internal decision trees to obtain the final result.

- **SVR**: The primary method of SVR regression is to find a regression plane such that all the data of a set are closest to that plane, and the model is regressed by the regression plane found. The SVR method is primarily a linear regression by constructing a linear decision function in a high dimensional space after dimensional enhancement.
- **CNN**: Convolutional neural networks consist of one or more convolutional layers and a top fully connected layer (corresponding to a classical neural network).
- **LSTM**: The LSTM architecture is a type of artificial recurrent neural network (RNN) that is utilized in deep learning. [37], long-term dependencies can be learned via LSTM.
- **FECNN+LSTM (Feature Engineering convolutional neural networks CNN + LSTM)**: combines feature engineering and four-layer CNN plus two-layer LSTM.

To evaluate our proposed Auto-REP based on automated machine learning pipeline, we use the LANL dataset and observe the time consumption of the whole process of feature extraction, feature selection, model construction, and hyperparameter extraction under the different regression methods.

1) Comparison results regarding RQ1: Table III gives the comparison results for different methods. The LANL dataset was used in this paper to develop an automated regression pipeline model that combines feature extraction, feature selection, and regression modelling. Our approach outperforms the competition in terms of experimental results, but most significantly, the automated machine learning approach does not need much human interaction and has considerable advantages in terms of model processing and parameter setting time.

As shown in Table III, the model of Feature Engneering CNN + LSTM achieved a MAE of 1.49 for training data and achieved a MAE of 1.51 for testing data. The MAE and MSE of our model in the training and testing datasets are 1.48, 1.51 and 1.52, 1.59, respectively, which are slightly inferior compared to the combination of CNN+LSTM, but the high automation process of the model enables the feature selection and model construction of the data in a shorter time, which

TABLE III
PERFORMANCE COMPARISON IN TERMS OF MSE, MAE

	Training data		Testing data		
Method	MSE	MAE	MSE	MAE	HM
Auto-REP	1.52	**1.48**	1.59	**1.51**	**0.79**
PCA+RF	1.74	1.71	1.73	1.80	0.62
SVR	1.85	1.87	1.88	1.94	0.63
CNN	2.22	2.41	2.29	2.51	0.72
FE CNN + LSTM	**1.51**	1.49	**1.58**	**1.51**	0.73
LSTM	1.79	1.82	1.86	1.85	0.75

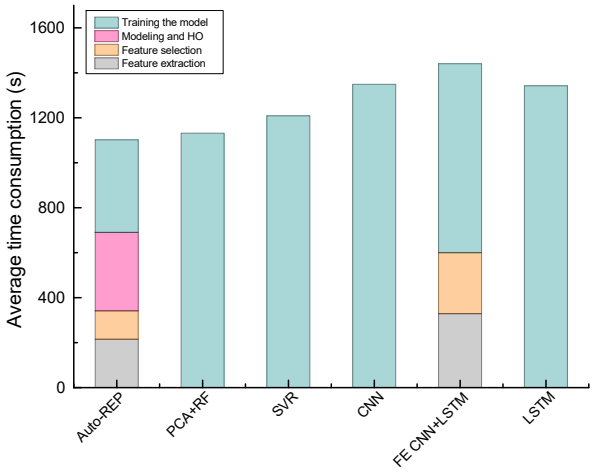


Fig. 5. Average time consumption using different methods for earthquake prediction on LANL testing dataset.

improves the training efficiency of the regression model (see also Fig.5). The superior result of our approach is largely due to the fact that our model is less affected by noise and performance of the RF model in the pipeline is most possibly due to the large number of trees in the forest, which is 800-1000 Decision Trees in our model. Therefore, the experimental results indicate that the proper number of decision trees is useful to improve the efficiency of the model.

As can be seen in Fig.5, our proposed pipeline regression method of automated machine learning has the lowest result in time consumption with 1,102s. This is also consistent with our assumptions, as the pipeline integrates the machine learning process and greatly reduces the time consumed to process data between different modules. We can also note that the feature engineering CNN + LSTM requires the most time consumption is the most time consuming with 1,440s. The major reason is that the feature extraction in the CNN is not as efficient as the feature extraction in our pipeline using tsfresh. In addition, the model combination of CNN + LSTM involves the training of convolutional neural network and RNN modules, which also increases the training time of the model. Moreover, the absence of an automated machine learning approach leads to the fact that each part of the pipeline from feature extraction to model building, is separated from each other and cannot be efficiently articulated, which additionally

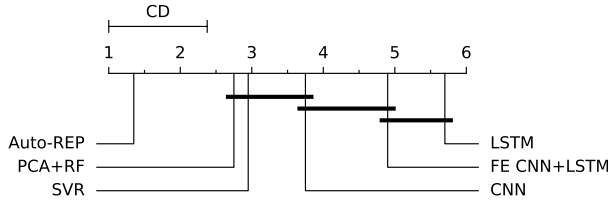


Fig. 6. CD diagram of the Nemenyi test on the HM metric. Numbers indicate mean ranks in 15 repetitive runs (lower means better). Rank means with non-significant difference are connected with a horizontal line.

causes extra time consumption.

2) Comparison results regarding RQ2: To address the second research question, we compared our automated machine learning approach with the existing methods, taking into account not only the accuracy but also the rapidness of the model for earthquake prediction. For this objective, we chose the harmonic mean (HM) [38] as suggested by Schäfer and Leser [39].

$$HM = 1 - \frac{2 \cdot (1 - C_r) \cdot (1 - C_a)}{(1 - C_r) + (1 - C_a)} \tag{3}$$

$C_r = 1 - \frac{\hat{T}_{all}}{T_{max\_fe}+T_{max\_fs}+T_{max\_ho}+T_{max\_m}}$, where T_{max_fe}, T_{max_fs}, T_{max_ho}, and T_{max_m} represent maximum time consumption of feature extraction, feature selection, hyperparameters optimisation and modeling, respectively. $\hat{T}_{all}$ indicates the the average total time consumption for each method. C_r and C_a are indicators of rapidity and accuracy, respectively. We will refer to this as the HM metric, whose values are in the range [0,1], to answer the second research question. Since the method proposed by Schäfer and Leser [39] is for the accuracy of the classification results, here we replace the accuracy C_a with $1 - RMSE/100$ value of the regression model, which more precisely reflects the predictive effect of the regression model. The value of $\hat{T}_{all}$ was obtained by conducting 15 regression experiments and obtaining the average time consumption for each method. We can also get T_{max_fe}, T_{max_fs}, T_{max_ho}, and T_{max_m} for the maximum time of each part. Then C_r for each method can be calculated through the Equation.(3).

For a comparative analysis of the results using HM metric among the different methods, we used the Friedman test to see if there were any significant differences between rank means, and then we used the Nemenyi post-hoc test [40] to see if there were any significant pairwise differences in average ranks. We used critical difference (CD) diagrams [41] to visualize the results of this test, which are commonly used for comparing different methods to show the result of a statistical comparison of ranking results. In order to ensure that the final CD diagrams obtained have more significant statistical difference, we selected 15 repetitions for the experiments for $RQ.2$. According to the comparison of methods based on the HM metric in Fig.6, Auto-REP outperforms all of the methods we evaluated. It can be shown that our proposed automated machine learning-based earthquake prediction method can balance accuracy and rapidity. Meanwhile, it can also be seen from Fig.6 that the rank means of PCA-RF, SVR, and CNN have non-significant difference. The results in Fig.6 demonstrate further that our Auto-REP method can balance advantageous prediction accuracy with less time consumption, providing stable and efficient regression performance over other comparative methods.

V. Conclusion and Future Work

Earthquake prediction research has received increasing attention and has become a hot issue that needs to be addressed. In this paper, we present a generic automated machine learning regression pipeline model and implement a time to event prediction using the LANL seismic dataset. Our regression pipeline model is able to achieve significant seismic prediction results, and is both efficient and automated.

In addition to the field of earthquake prediction, in the future, Auto-REP can also be applied in numerous computer and automation engineering fields, such as meteorological and disaster forecasting engineering. However, there is still potential to further improve the performance of our method. Therefore, in future research, we will consider data pre-training methods such as [42]–[44] in our proposed automated machine learning technique Auto-REP, which can potentially further improve the predictive performance of our method.

Acknowledgment

The authors would like to thank Furong Ye of the Leiden Institute of Advanced Computer Science (LIACS) from Leiden University for his helpful discussions on this research topic.

References

[1] J. C. Gill and F. Bullough, "Geoscience engagement in global development frameworks," *Annals of geophysics*, vol. 60, 2017.

[2] M. Kasai and T. Yamada, "Topographic effects on frequency-size distribution of landslides triggered by the hokkaido eastern iburi earthquake in 2018," *Earth, Planets and Space*, vol. 71, no. 1, pp. 1–12, 2019.

[3] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine learning for industrial applications: a comprehensive literature review," *Expert Systems with Applications*, vol. 175, p. 114820, 2021.

[4] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[5] Q. Wang, Y. Guo, L. Yu, and P. Li, "Earthquake prediction based on spatio-temporal data mining: an lstm network approach," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 1, pp. 148–158, 2017.

[6] P. A. Johnson, B. Rouet-Leduc, L. J. Pyrak-Nolte, G. C. Beroza, C. J. Marone, C. Hulbert, A. Howard, P. Singer, D. Gordeev, D. Karaflos *et al.*, "Laboratory earthquake forecasting: A machine learning competition," *Proceedings of the National Academy of Sciences*, vol. 118, no. 5, 2021.

[7] B. Rouet-Leduc, C. Hulbert, and P. A. Johnson, "Continuous chatter of the cascadia subduction zone revealed by machine learning," *Nature Geoscience*, vol. 12, no. 1, pp. 75–79, 2019.

[8] N. Lubbers, D. C. Bolton, J. Mohd-Yusof, C. Marone, K. Barros, and P. A. Johnson, "Earthquake catalog-based machine learning identification of laboratory fault states and the effects of magnitude of completeness," *Geophysical Research Letters*, vol. 45, no. 24, pp. 13–269, 2018.

[9] J. Huang, X. Wang, S. Yong, and Y. Feng, "A feature enginering framework for short-term earthquake prediction based on aeta data," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 2019, pp. 563–566.

[10] A. Bray and F. P. Schoenberg, "Assessment of point process models for earthquake forecasting," *Statistical science*, pp. 510–520, 2013.

[11] Y. Ogata, "Statistical models for earthquake occurrences and residual analysis for point processes," *Journal of the American Statistical association*, vol. 83, no. 401, pp. 9–27, 1988.

[12] H. Akhondi-Asl and J. D. Nelson, "M-estimate robust pca for seismic noise attenuation," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1853–1857.

[13] J. Cheng*, K. Chen, and M. D. Sacchi, "Application of robust principal component analysis (rpca) to suppress erratic noise in seismic records," in *SEG Technical Program Expanded Abstracts 2015*. Society of Exploration Geophysicists, 2015, pp. 4646–4651.

[14] Y. Chen, "Fast dictionary learning for noise attenuation of multidimensional seismic data," *Geophysical Journal International*, vol. 209, no. 1, pp. 21–31, 2017.

[15] Z. E. Ross, Y. Yue, M.-A. Meier, E. Hauksson, and T. H. Heaton, "Phaselink: A deep learning approach to seismic phase association," *Journal of Geophysical Research: Solid Earth*, vol. 124, no. 1, pp. 856–869, 2019.

[16] C. Holtz and V. Gokul, "Early forecasting of quakes via machine learning," 2019.

[17] Z. Li, M.-A. Meier, E. Hauksson, Z. Zhan, and J. Andrews, "Machine learning seismic wave discrimination: Application to earthquake early warning," *Geophysical Research Letters*, vol. 45, no. 10, pp. 4773–4779, 2018.

[18] J. Liu, Y. Huang, Y. Lu, and G. Zhang, "Earthquake prediction based on spatial-temporal data mining," in *International Conference on Intelligent Automation and Soft Computing*. Springer, 2021, pp. 1201–1212.

[19] D. Bruckner, "Ml-o-scope: a diagnostic visualization system for deep machine learning pipelines," CALIFORNIA UNIV BERKELEY DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCES, Tech. Rep., 2014.

[20] S. Estevez-Velarde, Y. Gutiérrez, A. Montoyo, and Y. A. Cruz, "Automatic discovery of heterogeneous machine learning pipelines: An application to natural language processing," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3558–3568.

[21] P. Kathiravelu, P. Sharma, A. Sharma, I. Banerjee, H. Trivedi, S. Purkayastha, P. Sinha, A. Cadrin-Chenevert, N. Safdar, and J. W. Gichoya, "A dicom framework for machine learning pipelines against real-time radiology images," *arXiv preprint arXiv:2004.07965*, 2020.

[22] S. Chanda and S. N. Somala, "Single-component/single-station–based machine learning for estimating magnitude and location of an earthquake: A support vector machine approach," *Pure and Applied Geophysics*, pp. 1–18, 2021.

[23] M. Han, Y. Liu, J. Xi, and W. Guo, "Noise smoothing for nonlinear time series using wavelet soft threshold," *IEEE signal processing letters*, vol. 14, no. 1, pp. 62–65, 2006.

[24] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta–a system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[25] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[26] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

[27] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated machine learning*. Springer, Cham, 2019, pp. 3–33.

[28] H. A. Fayed and A. F. Atiya, "Speed up grid-search for parameter selection of support vector machines," *Applied Soft Computing*, vol. 80, pp. 202–210, 2019.

[29] S. Mirjalili, "Evolutionary algorithms and neural networks," in *Studies in Computational Intelligence*. Springer, 2019, vol. 780.

[30] P. I. Frazier, "Bayesian optimization," in *Recent Advances in Optimization and Modeling of Contemporary Problems*. INFORMS, 2018, pp. 255–278.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[32] S. Kaur, H. Aggarwal, and R. Rani, "Hyper-parameter optimization of deep learning model for prediction of parkinson's disease," *Machine Vision and Applications*, vol. 31, no. 5, pp. 1–15, 2020.

[33] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, 2011.

[34] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1586–1594, 2019.

[35] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[36] T. Zaidi, A. Samy, M. Kocatürk, and H. F. Ateş, "Learned vs. hand-crafted features for deep learning based aperiodic laboratory earthquake time-prediction," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2020, pp. 1–4.

[37] T. M. Breuel, "Benchmarking of lstm networks," *arXiv preprint arXiv:1508.02774*, 2015.

[38] R. Beatson, M. S. Floater, and C. E. Kåshagen, "Hermite mean value interpolation on polygons," *Computer Aided Geometric Design*, vol. 60, pp. 18–27, 2018.

[39] P. Schäfer and U. Leser, "Teaser: early and accurate time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1336–1362, 2020.

[40] P. Nemenyi, "Distribution-free multiple comparisons. princeton university," *New Jersey*, 1963.

[41] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[42] Q. Xu, Q. Wen, and L. Sun, "Two-stage framework for seasonal time series forecasting," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3530–3534.

[43] H. Hu, M. Tang, and C. Bai, "Datsing: Data augmented time series forecasting with adversarial domain adaptation," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2061–2064.

[44] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2712–2721.