Preliminary Results of Automatic P-Wave Regional Earthquake Arrival Time Picking using Machine Learning with Kurtosis and Skewness as The Input Parameters

Y H Lumban Gaol¹, R K Lobo¹, S S Angkasa¹, A Abdullah¹, I Madrinovella¹, S Widyanti¹, A Priyono¹,², S K Suhardja¹, A D Nugraha², Z Zulfakriza², S Widiyantoro²,³, M Luqman Hakim⁴, K H Palgunadi⁵, B Mujihardi⁶

¹Faculty of Exploration and Production Technology, University of Pertamina, Jl. Teuku Nyak Arief, Simprug, Kebayoran Lama, Jakarta 12220, Indonesia
²Global Geophysics Research Group, Faculty of Mining and Petroleum Engineering Institute of Technology Bandung, Jalan Ganesa No. 10, Bandung 40132, Indonesia
³Faculty of Engineering, Maranatha Christian University Bandung, Jalan Surya Sumantri No 65, Bandung 40164, Indonesia
⁴Agency for Meteorology, Climatology and Geophysics, Jalan Angkasa I, No. 2, Kemayoran, Jakarta, Indonesia
⁵Physical Science and Engineering, King Abdullah University of Science and Technology, Saudi Arabia.
⁶Geothermal Research – Innovation & New Ventures, PT. Pertamina, Jl. Mega Kuningan Barat III, Jakarta Selatan, Indonesia

*Corresponding author: s4ndy104@gmail.com

Abstract. The traditional method in determining first arrival time of earthquake dataset is time consuming process due to waveform manual inspection. Additional waveform attributes can help determine events detection. One of the widely used attribute is The Short Term Averaging/Long Term Averaging (STA/LTA) which is simply division moving average of waveform amplitude between short time and longer time. Alternatively, waveform attribute can also be built using kurtosis and skewness. The kurtosis attribute is defined as sharpness of data distribution. By definition, the maximum signal should be at or close to the P wave arrival. The skewness is typically used to show normal distribution of the data. The uniqueness of this method is that it has an ability to determine whether the first P wave arrival has positive of negative number. The skewness calculation is similar to kurtosis but it uses the power of 3 instead of 4. With the objective of generating efficient scheme to pick first time arrival, we explore use artificial neural network and a combination of kurtosis and skewness attributes. We use a collection of magnitude events with moment magnitude larger than 3 located close to Moluccas island, Indonesia. We collected all events information from the Indonesian Agency of Meteorology, Climatology and Geophysics. The process is started with manually pick all P wave arrivals using manual inspection. Next, we trained the artificial neural network with attributes numbers as inputs and arrival time we manually picked as the output. In total we used 100 regional events that has clear P wave phases. Then, we validated the results until reaching 0.99 accuracy. In the last step, we tested the once trained procedures on new waveforms contained events. Current result shows an average of 0.4s different between manual
pick and trained picked from machine learning. The accuracy can be improved by applying a band pass 0.1-2 Hz filtering with an average of 0.2s.

1. Introduction.
Machine learning is a topic that rapidly developed in the last few years and has already been applied in many different industry branches, also in seismic processing. One machine learning algorithm that can be applied for first-break picking purposes is the Artificial Neural Network (ANN). [1] used this method to predict the earthquake first-break using 11 trained input parameters so the computer can learn its pattern to estimate the first break of a new earthquake data. Another example for first break picking purposes, e.g. support vector machines [2] reinforcement learning [3] or convolutional neural networks [4],[5]. These studies motivated us to try a similar method with a more straightforward yet practical approach.

In this paper, we propose a new innovative and efficient P-wave first break picking method using Artificial Neural Network (ANN) by combining kurtosis and skewness as the input parameters that are effective enough to distinguish a first-break and not first-break. These parameters will be trained using Keras, a Python’s newest machine learning library. The result expected from this research is a Python-based program that can do automatic first break picking.

2. Data and Method
2.1. Artificial Neural Network (ANN)
ANN is a machine learning algorithm that works similarly to the human brain system [6]. ANN is a collection of connected units or nodes called neurons that can be separated into three main parts (architecture): input layer, hidden layer, and output layer. The general equation of ANN is [7]:

\[ y = f \left( \sum_{i=0}^{n} (x_i W_i) + B \right) \]  

where \( y \) is output, \( x \) is input, \( W \) is weight and \( B \) is bias

![Figure 1. ANN general architecture](image)

2.2. Kurtosis and skewness
Kurtosis
Kurtosis is a degree of peakedness or sharpness from a data distribution. Normal Gaussian distribution will have kurtosis = 0, but earthquake data is a non-Gaussian distribution and has a higher kurtosis value. With this characteristic, kurtosis can be used to identify a non-Gaussian distribution (first-break) that will be characterized by a maximum kurtosis value [8].
Skewness is a degree of distortion from normal distribution data. In a simple term, skewness measures the lack of symmetry in the data distribution. An earthquake arrival time data is characterized by an amplitude increase, which is higher than its mean amplitude value. When a significant difference in amplitude, the skewness attribute is sufficient to detect a first-break and its type (peak or trough) [8].

\[
Kurtosis = \frac{\sum_{i=1}^{M}(x_i - \bar{x})^4}{(M - 1)s_x^4} - 3
\]

\[
Skewness = \frac{\sum_{i=1}^{M}(x_i - \bar{x})^3}{(M - 1)s_x^3}
\]

where \(M\) is total samples of amplitude, \(x = \{x_1, x_2, ..., x_M\}\) is amplitude value, \(\bar{x}\) is mean of amplitude, and \(s_x\) standard deviation [9].

2.3. Mean Absolute Percentage Error (MAPE)

We calculated the error of machine learning auto pick by comparing it to manual pick (considered as true first-break) using MAPE equation [10]:

\[
MAPE = \frac{|P_m - P_o|}{P_m} \times 100\%
\]

where \(P_m\) is manual-pick sample’s position, \(P_o\) is the machine learning auto-pick sample’s position.

2.4. Data Processing and General Workflow

In this study, we used 106 local-earthquake waveforms with a sampling rate of 20 samples/second and magnitude > 3 that recorded by BMKG station from November 06, 2019 to November 15, 2019. We apply a bandpass filter with an average frequency range is 0.5 - 2.5 Hz to increase the signal-to-noise ratio of the waveforms. Next, we manually picked the first break and calculated kurtosis and skewness, used as input parameters. Then, we give a label to the data with 1 and 0; 1 for the first-break and 0 for not first-break. We used 11 samples of each waveform with detail as one sample with label 1 is the first-break, and ten samples with label 0 are non-first-break (five samples before and after the first break). The dataset will be normalized to simplify the calculation process. After that, we train the dataset and do testing on synthetic data that we downloaded from the IRIS dataset to check if our model is already good enough or not.

3. Results and Discussions

We trained 106 pairs datasets that consist of kurtosis and skewness of each event. This machine learning model’s accuracy has reached 92% after 200 iterations with model loss below 0.22. A good model loss is getting closer to 0, which means that machine learning’s predictive results have no difference to the actual value.

A. Synthetic Data

We make first break predictions on three types of synthetic data with sampling rate of 4 samples/second as a benchmark to whether the machine learning model that has been trained is good enough in predicting the first break. The accuracy of predicted results on all three types of synthetic data (Figure 2) will be analyzed from MAPE value (Table 1).
Figure 2. Auto first-break picking on Synthetic Data

Table 1. Analysis of first-break picking prediction on synthetic data

|                 | Sample position on manual first-break picking | Sample position on auto first-break picking | Differences | MAP Error |
|-----------------|-----------------------------------------------|--------------------------------------------|-------------|-----------|
| Clean synthetic data | 588th (147 second)                             | 588 (147 second)                           | 0           | 0 %       |
| Medium noise synthetic data | 588th (147 second)                             | 597 (149.25 second)                        | 9 samples (2.25 second) | 1.5 %     |
| High noise synthetic data | 588th (147 second)                             | 593 (148.25 second)                        | 5 samples (1.25 second) | 0.8 %     |

The machine learning model that has been trained before is good enough in predicting the first break on clean or noisy synthetic data. The error on clean synthetic data is 0% that means the prediction using machine learning is identical to the actual first break. Meanwhile, auto picking results error on medium and high noise data are 0.015% and 0.008%, respectively (2.25s and 1.25s differences from the actual first break).

B. Real Data
Once we are satisfied with the predicted results on the synthetic data, we continued the auto picking prediction on the unfiltered and filtered real data (Figure 3 – Figure 5).
Figure 3. Auto first-break picking on real data 1 (raw and filtered)

Figure 4. Auto first-break picking on real data 2 (raw and filtered)

Figure 5. Auto first-break picking on real data 3 (raw and filtered)
Table 2. Analysis of first-break picking prediction on real data

| Event   | MAPE of Auto-Pick on Raw Data (Differences (s)) | MAPE of Auto-Pick on Filtered Data (Differences (s)) |
|---------|-----------------------------------------------|---------------------------------------------------|
| Event 1 | 1.43% (0.45s)                                 | 1.14% (0.35s)                                     |
| Event 2 | 0.16% (0.05s)                                 | 2.52% (0.75s)                                     |
| Event 3 | 1.18% (0.15s)                                 | 0.79% (0.05s)                                     |

Auto-picking was applied to three real earthquake events and show reasonably satisfactory results. On raw (unfiltered) data, the most significant error is demonstrated by Event 1 and Event 3, with the highest deviation of 0.44s and 0.15s, respectively. Moreover, the filtered data results are better, with the highest deviation is only 0.75s on Event 2. Overall, the prediction of filtered data is better than raw data because most of the data we trained is the filtered data. We think that increasing the amount of training dataset can improve accuracy because we have more variations that can be learned by the computer.

4. Conclusions
We have conducted the first break picking of the earthquake using an artificial neural network, a machine learning algorithm. The results showed that our machine learning model is already good enough to predict the first break; however, it still has a few deviation or error in predicting raw data. We can improve our machine learning model by adding more data to be trained so our computer can learn more variation patterns of earthquake’s first break.

References
[1] Mężyk M, & Malinowski M 2019 Multi-pattern algorithm for first-break picking employing open-source machine learning libraries Journal of Applied Geophysics Vol. 170 103848
[2] Yalcinoglu L, Millahn K, Fruhwirth R 2011 Automatic first break detection using support vector machine classification 73rd Conference and Exhibition incorporating SPE EUROPEC EAGE
[3] Ma Y, Luo Y, 2018 Automatic first-arrival picking with Reinforcement Learning, CPS/SEG International Geophysical Conference
[4] Yuan S, Liu J, Wang S, Wang T, Shi P 2018 Seismic waveform classification and first-break picking using convolution neural network IEEE Geoscience and Remote Sensing Letters 15 p 272–276
[5] Tsai K C, Hu W, Wu X, Chen J, and Han Z 2018 First-break automatic picking with deep semisupervised learning neural network SEG Technical Program Expanded Abstracts : 2181-2185
[6] Brunton S L, Kutz J N 2019 Neural networks and deep learning In Data-Driven Science and Engineering pp. 195–226
[7] Aggarwal C C 2018 *Neural networks and deep learning* In Neural Networks and Deep Learning Springer International Publishing AG

[8] Gunawan H et al 2012 *Penentuan waktu tiba gelombang-p secara otomatis dengan metoda skewness dan kurtosis terintegrasi* JURNAL METEOROLOGI DAN GEOFISIKA, VOL. 13 (1) 63–69

[9] Putra J W G 2019 *Pengenalan konsep pembelajaran mesin dan deep learning*. Retrieved from: https://www.researchgate.net/publication/323700644_Pengenalan_Pembelajaran_Mesin_dan_Deep_Learning

[10] Choi Y, Ryu S, Park K, Kim H 2019 *Machine learning-based lithium-ion battery capacity estimation exploiting multi-channel charging profiles* IEEE Access 7 75143–75152