MIDDLE EAST TECHNICAL UNIVERSITY

DEPARTMENT OF COMPUTER ENGINEERING

CENG 562: MACHINE LEARNING

TERM PAPER OF THE PROJECT

Predicting Ground Motion Parameters Using Machine Learning Algorithms for Türkiye

Çağlar TEMİZ

MSc. Student in Civil Engineering

2227940

Due Date: 01/06/2025

Table of Contents

Abstract: 3
INTRODUCTION 4
LITERATURE REVIEW 4
Seismic Event Discrimination 5
Earthquake Signal Detection and Earthquake Early Warning 5
Seismogram Simulation and Ground Motion Characterization 6
METHODOLOGY AND IMPLEMENTATION 9
Dataset 9
Applied Machine Learning Algorithms 10
Implementation 12
RESULTS AND DISCUSSION 14
Results 14
Discussion 17
CONCLUSION 17
REFERENCES 18

Predicting Ground Motion Parameters Using Machine Learning Algorithms for Türkiye

Abstract:

Earthquakes are one of the most destructive natural disasters in the world. Due to the lack of information about the earth crust, our understanding of earthquakes remains incomplete. However, earthquake engineers and engineering seismologists are trying to estimate the major impact parameters of an earthquake, in other words ground motion parameters. There are various methods that have been developed to estimate these ground motion parameters, and they are generally categorized into stochastic methods, deterministic methods, and artificial intelligence-based approaches which include machine learning or deep learning algorithms. In this study, selected machine learning algorithms are applied to the dataset that includes the key parameters obtained from station records of past earthquakes in the Türkiye region, specifically for events with magnitudes between $6 \leq M_w \leq 8$. For this study, the dataset is introduced and the machine learning algorithms and their usage is summarized. Despite data limitations and incomplete results, the current implementation is summarized and planned, future directions are discussed.

INTRODUCTION

Earthquakes are among the most significant and destructive natural disasters in the world, cause severe material and moral damage. Türkiye is one of the country's most seriously affected by earthquakes due to its location on active fault zones such as the North Anatolian Fault Zone (NAFZ) and the East Anatolian Fault Zone (EAFZ). Due to unknown nature of the earthquake, there are numerous studies to estimate the impact of the earthquakes, specifically the ground motion parameters.

The main ground motion parameters are PGA (Peak Ground Acceleration), PGV (Peak Ground Velocity), $PSA_{T=n}$ (Pseudo Spectral Acceleration at a period of *n* seconds), as well as Arias and Housner Intensities. These parameters can be changed or increased depending on regional seismic characteristics and the structural conditions. To calculate or estimate these parameters, station recordings or some methods are used. These methods can be grouped as stochastic methods, deterministic methods, and artificial intelligence-based approaches.

In this study, artificial intelligence-based approaches are investigated as a literature survey and through a literature survey, and PGA estimation is performed as a toy problem using bagging (Random Forest) and boosting (Gradient Boosting, XGBoost, AdaBoost, CatBoost, and LightGBM) machine learning algorithms, and additional multilayer perceptron algorithm (ANN – Artificial Neural Network). The dataset used in the analysis is obtained from the Disaster and Emergency Management Authority (AFAD) database. The results are presented in the corresponding section; however, the outcomes did not meet the expected level of success, primarily due to the limited size of the dataset. Despite, this study provides a valuable starting point for integrating artificial intelligence methods into earthquake-related research.

LITERATURE REVIEW

In recent years, the number of studies investigating earthquake studies using machine learning or generally artificial intelligence techniques has been increasing. Machine learning applications for seismic event monitoring can be divided into several areas such as, seismic event discrimination, earthquake signal detection, seismogram simulation, ground motion characterization, earthquake forecasting and more (Mousavi & Beroza, 2022). The studies conducted on some of these topics, and their aim, methodology and implementation are shown below in general.

Seismic Event Discrimination

Identifying whether a signal originates from an earthquake or a different source is often crucial, particularly in mining areas or quarry blasts. In the study done by Linville et al. (2019), they propose a model of two neural network (NN) architectures (CNN and RNN) on the task of binary event classification for tectonic earthquakes and quarry blasts at local scales. The dataset is waveforms collected from the database of University of Utah Seismograph Stations (UUSS), and the algorithms are applied on them shown in Figure 1. The base parameter of the algorithms is the nature of the seismic wave (P-waves are faster and always arrive before S with a lower amplitude).

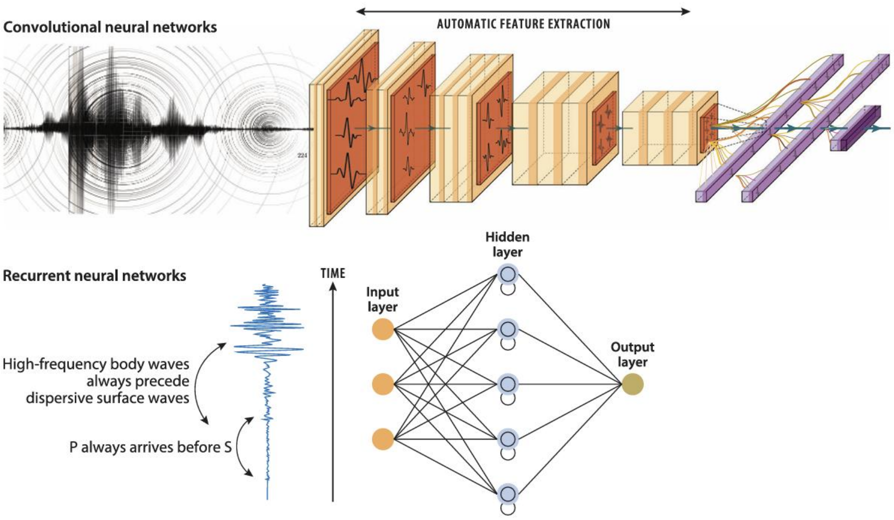


Figure 1. A scheme of the study done by Linville et al. (Mousavi & Beroza, 2022)

The study presents their findings have more than 99% accuracies while allowing them to access both source and path-controlled information, in addition to ambient background noise and null data (zero fill for vertical-only stations), requiring each model to learn, through training, which aspects of the input domain are most important for prediction.

Earthquake Signal Detection and Earthquake Early Warning

Understanding an earthquake signal’s property has an importance especially in earthquake early warning studies. A study done by Wang & Teng (1995) shows that using the dataset of the ratio in between STA (short time average) and LTA (long time average) of the signals, earthquakes can be detected before the full signal arrives with ANN network. The schematic representation of ANN and the algorithm of the recommended study shown in Figure 2 and 3.

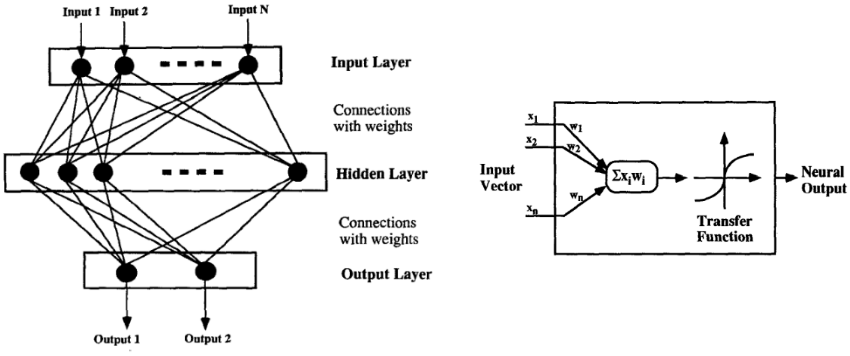


Figure 2. ANN network representation.

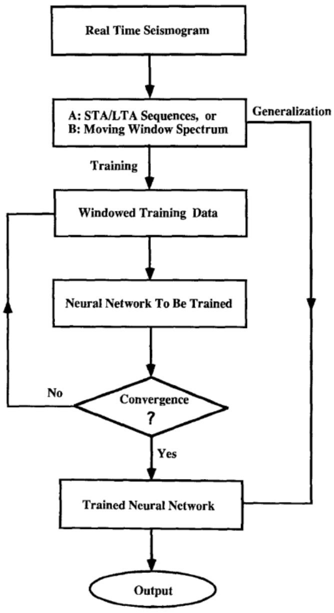


Figure 3. Earthquake signal detection algorithm (Wang & Teng, 1995).

Seismogram Simulation and Ground Motion Characterization

Seismogram simulation has a key role used to estimate potential earthquake impacts and support regional preparedness, even in areas lacking historical seismic records. Ground Motion Models (GMMs) serve as one of the key validation tools for these simulated seismograms. Therefore, this topic encompasses both seismogram simulation and ground motion characterization. Additionally, fault and site properties play a significant role in informing and training artificial intelligence algorithms.

A study conducted by Karimzadeh et al. (2024), shows a GMM (Ground Motion Model, a model that shows ground motion parameters in a range of given standard deviations) with a two layered neural network model with the inputs of stochastic simulation shown in Figure 4 and this model is compared with the existing GMMs for different Joyner-Boore distances in Figure 5.

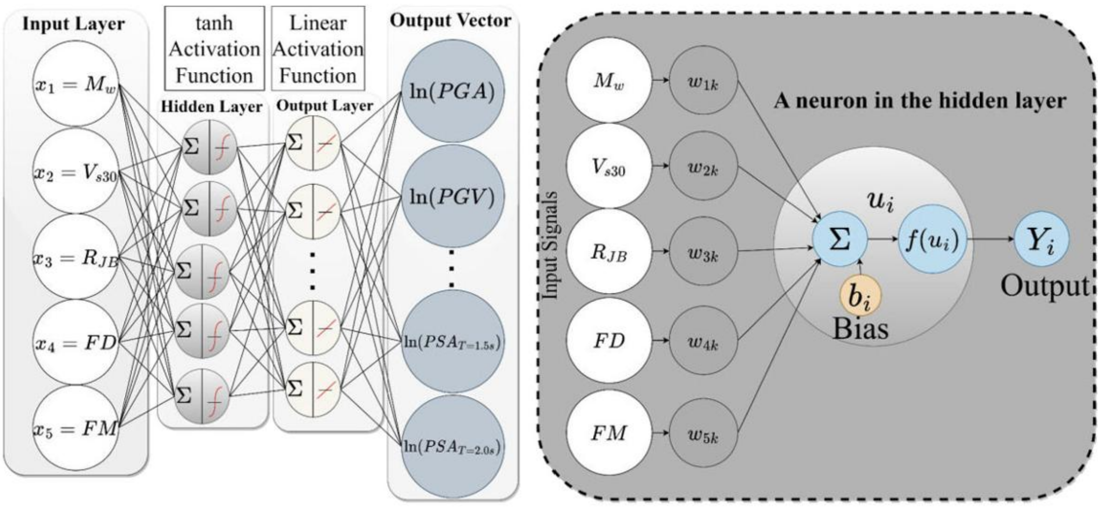


Figure 4. Structure of the ANN model and illustration of artificial neurons of the hidden layer (Karimzadeh et al., 2024).

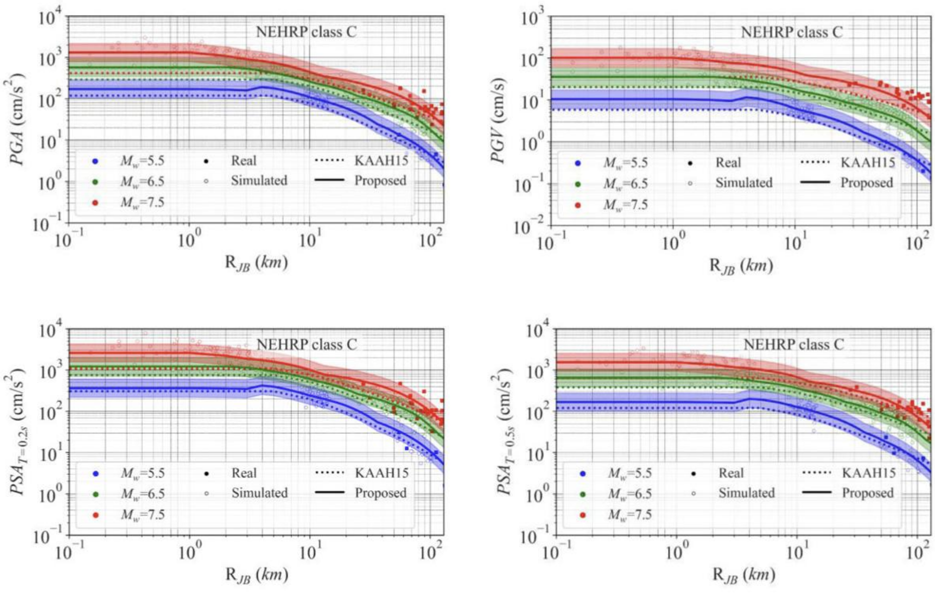


Figure 5. Variation of PGA, PGV and PSA at periods of 0.2, 0.5 s (Karimzadeh et al., 2024).

Kuran et al., (2024) presented another study in which machine learning algorithms were employed to estimate PGV (Peak Ground Velocity) to assess potential damage to mid-rise and high-rise buildings. the dataset used in the study includes over 950 earthquakes that occurred between 1983 and 2023, sourced from the Turkish Strong Motion Database. The applied algorithms are Random Forest, Support Vector Machine, Linear Regression, Artificial Neural Network, Gradient Boosting, and Bayesian Ridge Regression. Figure 6 and 7 shows the flow chart of the study and comparison of the results with actual values.

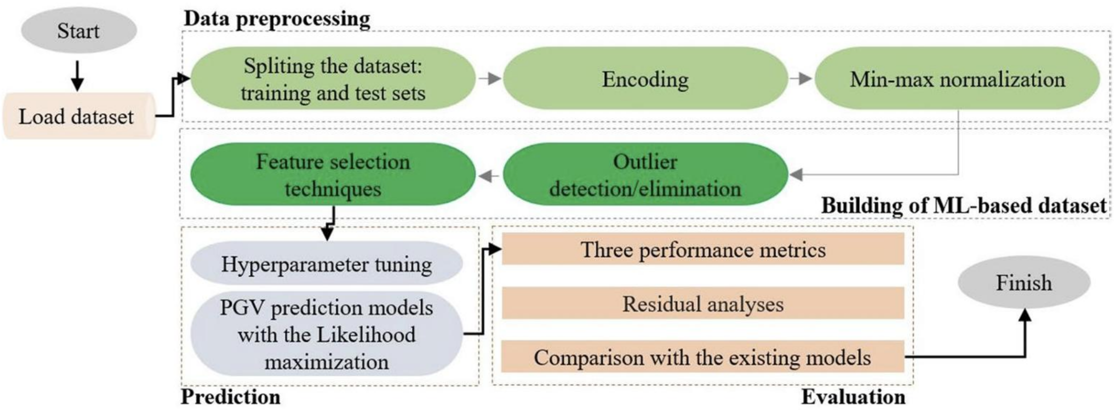


Figure 6. Flowchart followed in Kuran et al. (2024).

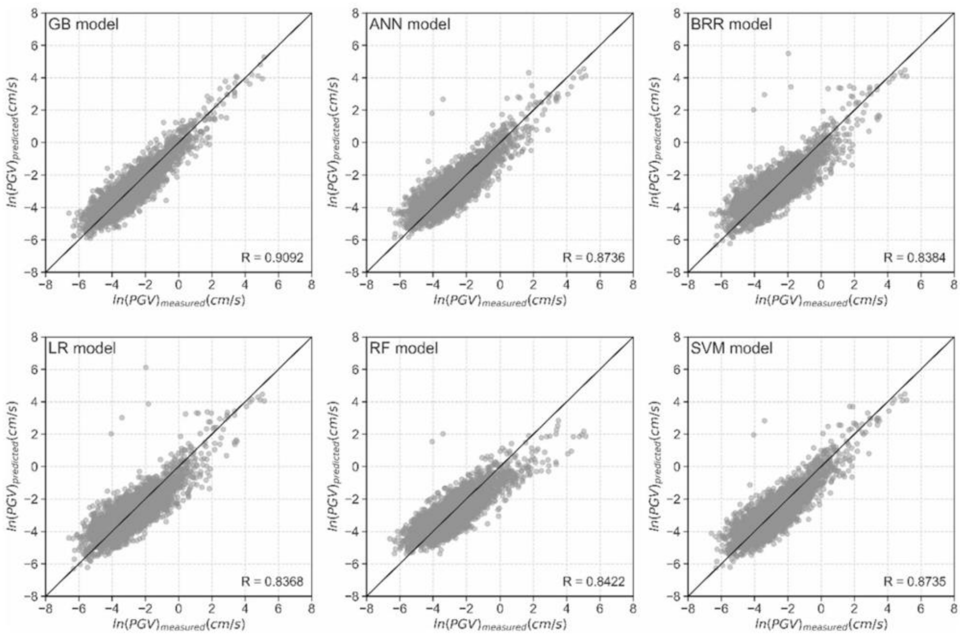


Figure 7. Comparison of measured and predicted ln(PGV)s in Kuran et al. (2024).

METHODOLOGY AND IMPLEMENTATION

The methodology and implementations of the top problem that is predicting PGA values for horizontal (N-S or E-W) and vertical directions (U-D) from the selected earthquakes with $6 \leq M_w \leq 8$ are explained in this part. The reason for this magnitude value selection is the estimation of PGA values from strong ground motion. However, this selection causes some problems shown in the results and discussed in the discussion part. The dataset and applied algorithms are explained below. Used dataset and codes are published in GitHub repo retrieved from github.com/CaglarTemiz/CENG562-Project.

Dataset

The dataset contains 2177 raw station data from the earthquakes that have magnitude values from M_w = 6.0 to M_w =7.7 and their PGA values. Figure 8 and Table 1 shows the selected earthquakes and their parameters whereas Figure 9 shows the station network of AFAD TADAS that the data is taken from.

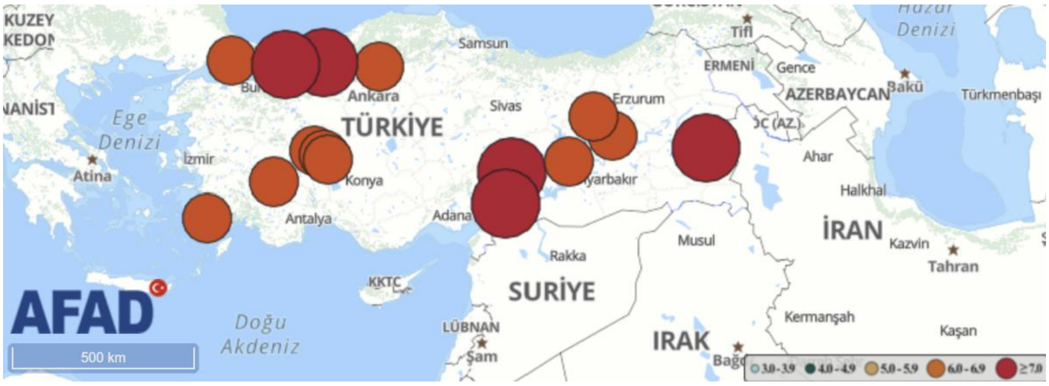


Figure 8. Selected earthquakes from AFAD TADAS website.

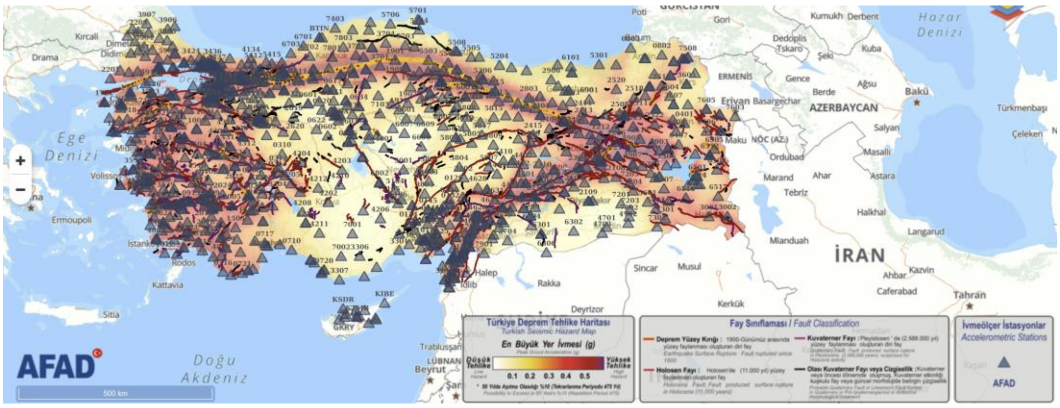


Figure 9. Accelerometer stations from AFAD TADAS website.

Table 1. Selected earthquakes and their main parameters.

Event ID	Event Date		Epicenter Agency	Epicenter Lon	Epicenter Lat	Magnitude	Depth	Location
543428	06-02-2023	01:17	AFAD	37.043	37.288	7.7	8.6	Pazarcık (Kahramanmaraş)
543593	06-02-2023	10:24	AFAD	37.239	38.089	7.6	7	Elbistan (Kahramanmaraş)
247730	17-08-1999	00:01	ISC	29.955	40.756	7.6	17	GÖLCÜK (IZMIT)
246572	12-11-1999	16:57	ISC	31.187	40.806	7.1	10.4	DÜZCE (BOLU)
457758	24-01-2020	17:55	DDA	39.063	38.3593	6.8	8.06	Sivrice (Elazig)
543431	06-02-2023	01:28	AFAD	36.92	37.304	6.6	6.2	Nurdağı (Gaziantep)
483762	30-10-2020	11:51	AFAD	26.703	37.879	6.6	14.9	Ege Denizi, 17.26 km Seferihisar (Izmir)
381491	20-07-2017	22:31	AFAD	27.4435	36.9198	6.5	19.44	Ege Denizi, 5.09 km Bodrum (Mugla)
658148	23-04-2025	09:49	AFAD	28.22639	40.85361	6.2	4.91	Marmara Denizi - [23.16 km] Silivri (İstanbul)
375576	12-06-2017	12:28	AFAD	26.3126	38.8488	6.2	15.86	Ege Denizi, 20.73 km Karaburun (Izmir)
444581	08-08-2019	11:25	DDA	29.584	37.851	6	10.92	Bozkurt (Denizli)
141933	23-10-2011	10:41	AFAD	43.4657	38.689	7	19.02	MERKEZ (VAN)
236848	01-05-2003	00:27	ISC	40.4637	38.9987	6.3	10	MERKEZ (BINGÖL)
246572	12-11-1999	16:57	ISC	31.187	40.806	7.1	10.4	DÜZCE (BOLU)
247730	17-08-1999	00:01	ISC	29.955	40.756	7.6	17	GÖLCÜK (IZMIT)

On Table 2, there is an example representation of station data belonging to the specific earthquake. Full Dataset is obtained from these station data including earthquake parameters.

Table 2. Example of station data.

Location	Code	Longitude	Latitude	Province	District	PGA_NS	PGA_EW	PGA_UD	Repi
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3429	28.25964	41.08492	İstanbul	Silivri	72.66854	106.3631	34.49335	25.0425
Marmara Denizi - [23.16 km] Silivri (İstanbul)	5906	27.93164	40.97338	Tekirdağ	Marmaraereğlisi	107.3819	68.33629	29.0576	29.15053
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3428	28.7296	40.98455	İstanbul	Avcılar	99.26741	81.12464	64.24406	43.05064
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3415	28.75848	41.02729	İstanbul	Küçükçekmece	210.1979	138.988	71.59369	47.01171
Marmara Denizi - [23.16 km] Silivri (İstanbul)	5917	28.00535	41.2706	Tekirdağ	Çerkezköy	24.96344	25.73455	12.33026	49.86276
Marmara Denizi - [23.16 km] Silivri (İstanbul)	5907	27.77633	41.1418	Tekirdağ	Çorlu	18.40127	21.99868	12.65696	50.24959
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3416	28.83635	40.97466	İstanbul	Bakırköy	37.94964	27.06303	33.33609	51.34333
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3432	28.79038	41.10586	İstanbul	Başakşehir	50.91311	64.34055	45.17266	53.36192
Marmara Denizi - [23.16 km] Silivri (İstanbul)	3431	28.71567	41.18623	İstanbul	Arnavutköy	100.6746	160.5748	56.03554	53.65241
Marmara Denizi - [23.16 km] Silivri (İstanbul)	1659	28.39153	40.37506	Bursa	Karacabey	8.021781	9.236179	4.6668	55.33324

Applied Machine Learning Algorithms

In the toy problem, six machine learning algorithms are used. These algorithms are used just to try and obtain some results for this preliminary study. Algorithm parameters are selected randomly and shown in Table 4. The best-case sscenarios are saved for future implementations and shown in the results part. Python programming language is used with numpy, pandas and scikit-learn libraries for the implementation.

Table 4. Applied algorithms and their parameters

Gradient Boosting	**n_estimators:** [50, 100, 200] **learning_rate:** [0.01, 0.1, 0.2] **max_depth:** [3, 4, 5] **min_samples_split:** [2, 5, 10] **min_samples_leaf:** [1, 2, 4]
XGBoost	**n_estimators:** [50, 100, 200] **learning_rate:** [0.01, 0.1, 0.2] **max_depth:** [3, 6, 9] **min_child_weight:** [1, 5, 10] **subsample:** [0.8, 0.9, 1.0] **colsample_bytree:** [0.8, 0.9, 1.0]
AdaBoost	**n_estimators:** [50, 100, 200] **learning_rate:** [0.01, 0.1, 0.2] **loss:** ['linear', 'square']
CatBoost	**iterations:** [100, 200] **depth:** [4, 6, 8] **learning_rate:** [0.01, 0.05] **l2_leaf_reg:** [3, 5] **border_count:** [32, 50] **bagging_temperature:** [0.0, 0.5]
LightGBM	**num_leaves:** [15, 31] **learning_rate:** [0.05] **n_estimators:** [50] **max_depth:** [3, 5] **min_child_samples:** [50] **subsample:** [0.8] **colsample_bytree:** [0.8] **max_bin:** [63] **lambda_l1:** [0] **lambda_l2:** [0]
Random Forest	**n_estimators:** [50, 100, 200] **max_depth:** [None, 10, 20, 30] **min_samples_split:** [2, 5, 10] **min_samples_leaf:** [1, 2, 4] **bootstrap:** [True, False]
ANN	**hidden_layer_sizes** = (100, 50) **activation** = 'tanh' **learning_rate** = 'adaptive' **max_iter** = 5000 **random_state** = 42

Implementation

From the 2177 raw data with 18 columns, unnecessary columns are eliminated. Geometric means are calculated for N-S and E-W components of PGA values. And for both vertical and horizontal components of PGA, z-score is calculated and if the absolute value of the z-score is above 3, this data is accepted as an outlier. For vertical PGA values, 2143 data and for horizontal ones 2134 data is used. Table 3 and Figures 10 and 11 show the last version of the dataset after elimination.

Table 4. Dataset after elimination process

Magnitude	Depth	PGA_NS	PGA_EW	PGA_UD	Repi	PGA_H	z_score
6.2	4.91	72.67	106.36	34.49	25.04	87.92	0.65
6.2	4.91	107.38	68.34	29.06	29.15	85.66	0.62
6.2	4.91	99.27	81.12	64.24	43.05	89.74	0.67
6.2	4.91	210.20	138.99	71.59	47.01	170.92	1.51
6.2	4.91	24.96	25.73	12.33	49.86	25.35	0.00
...	...	...	...	...	...	...	...
7.6	17	5.92	11.69	3.69	334.73	8.32	-0.18
7.6	17	11.69	8.91	4.43	345.24	10.21	-0.16
7.6	17	9.89	10.80	3.33	346.53	10.33	-0.16
7.6	17	5.98	5.25	3.30	371.91	5.60	-0.21
7.6	17	0.85	1.16	0.37	561.02	0.99	-0.25

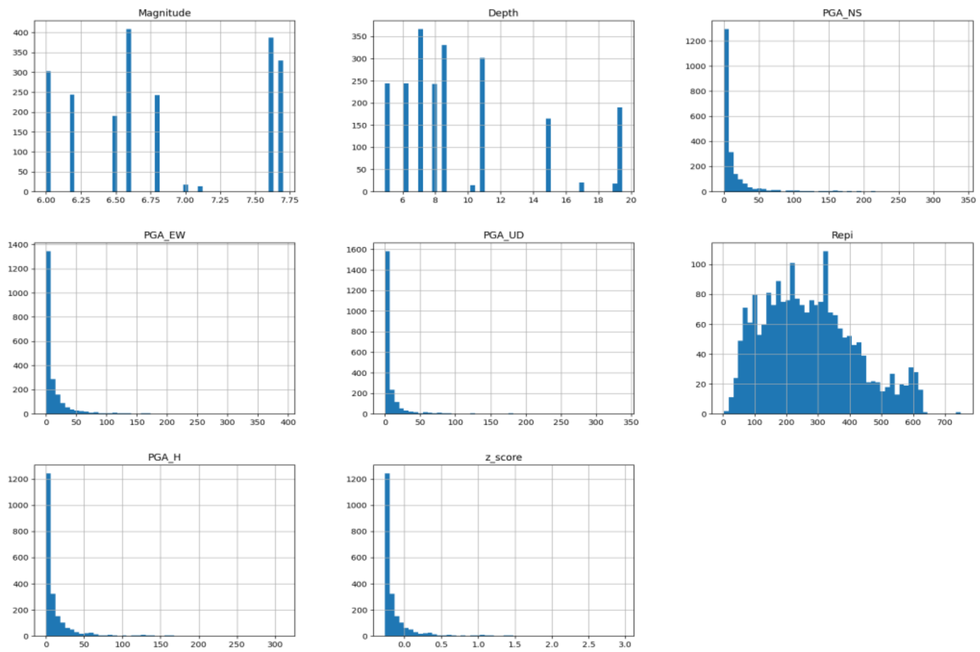
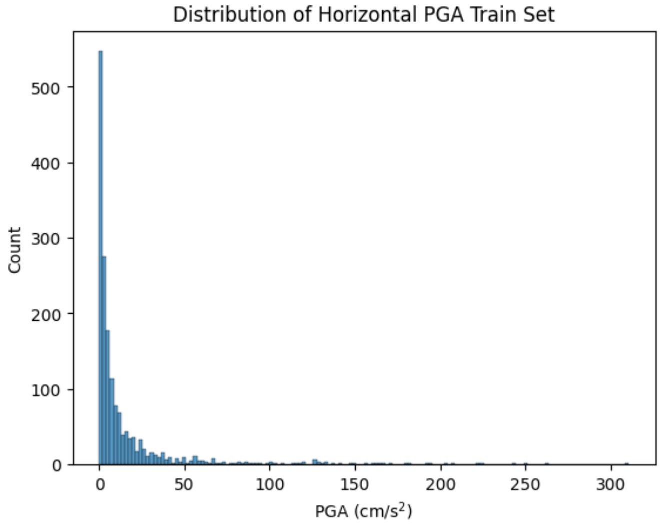


Figure 10. Histogram plots of the variables in eliminated dataset.



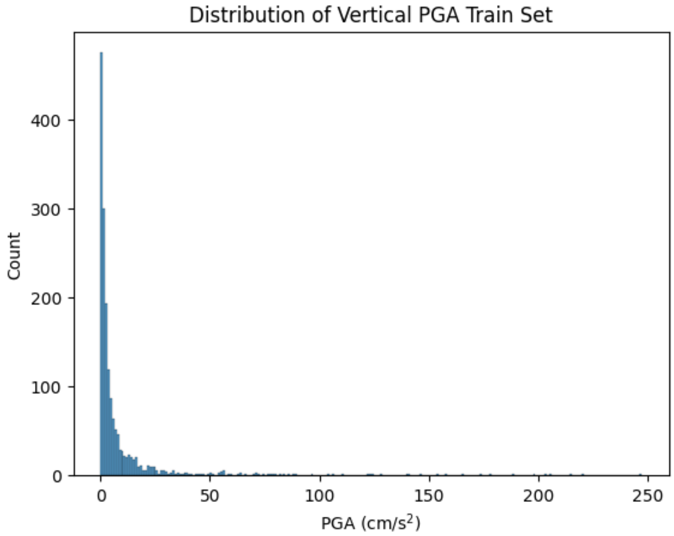


Figure 11. Vertical and horizontal PGA distributions.

Afterwards, the data is separated as 80% for train and 20% for test data before application of the machine learning algorithms. Magnitude value, Depth of the event and Epicentral distance (R_{epi}) is selected as an input whereas PGA values are selected outputs.

RESULTS AND DISCUSSION

Results

In Figure 12 and 13, comparison of the algorithm outputs and test data results can be clearly seen. For learning effectiveness parameters, Figure 13 and 14 demonstrates algorithms R square (R2) and Root Mean Square Error (RMSE) result for both vertical and horizontal PGA.

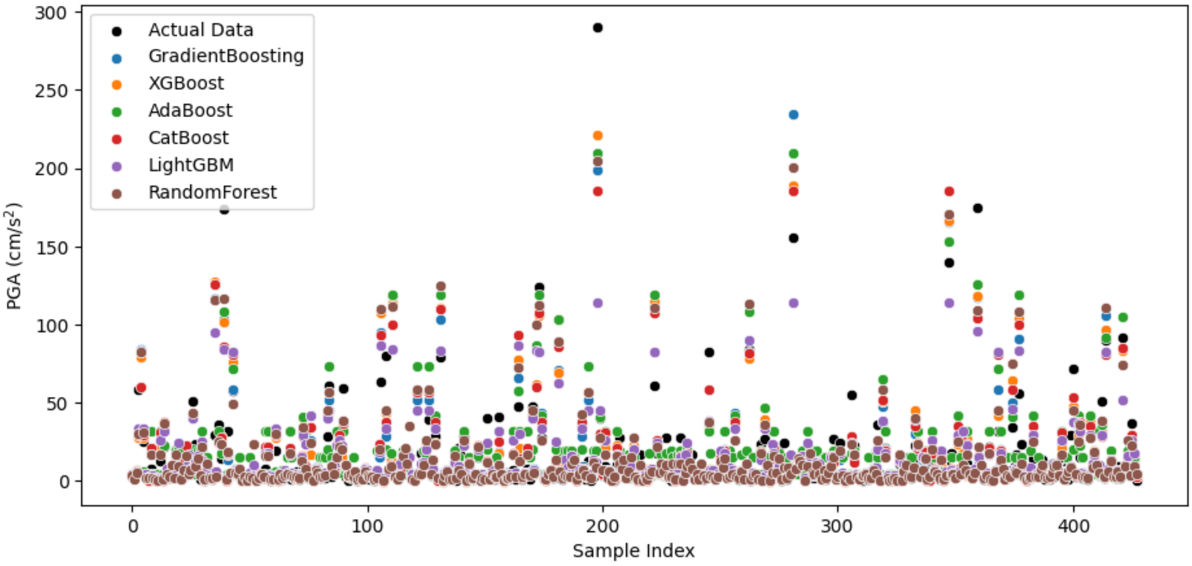


Figure 12. Comparison of algorithms' horizontal PGA test results with actual data

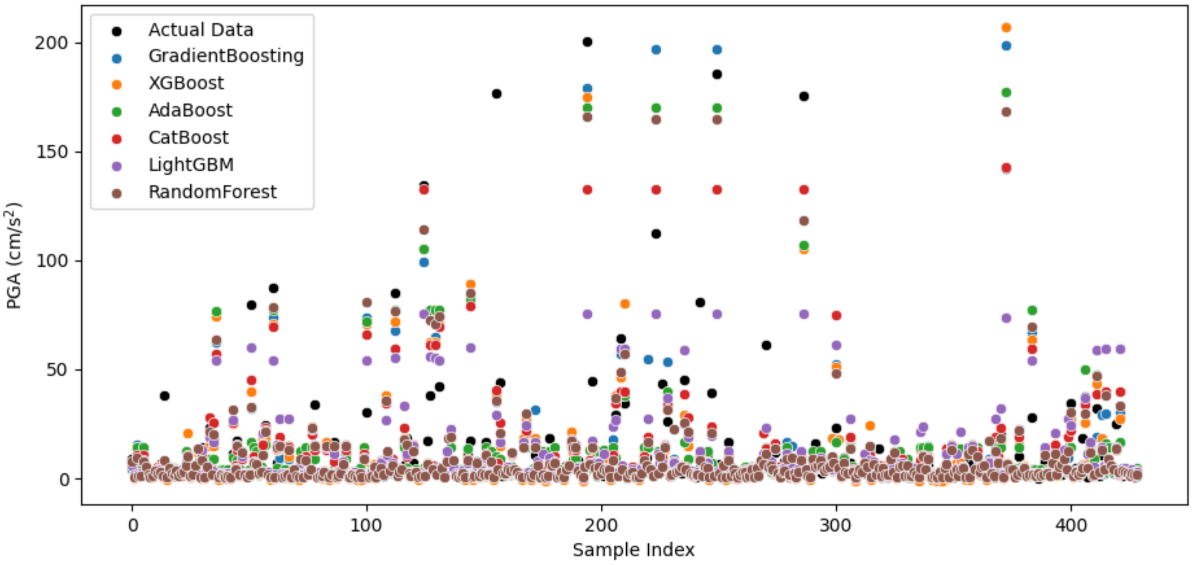


Figure 13. Comparison of algorithms' vertical PGA test results with actual data

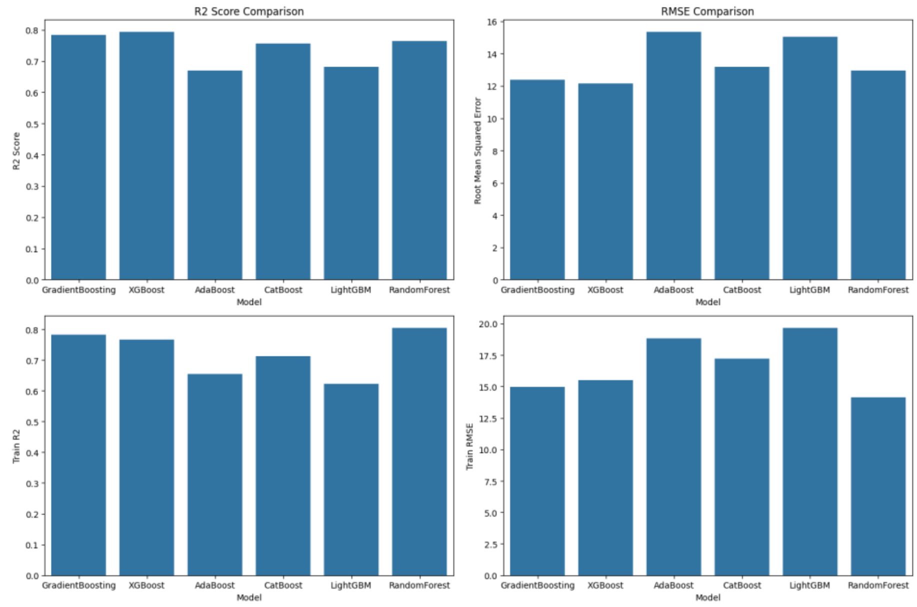


Figure 14. Learning effectiveness parameters for horizontal PGA.

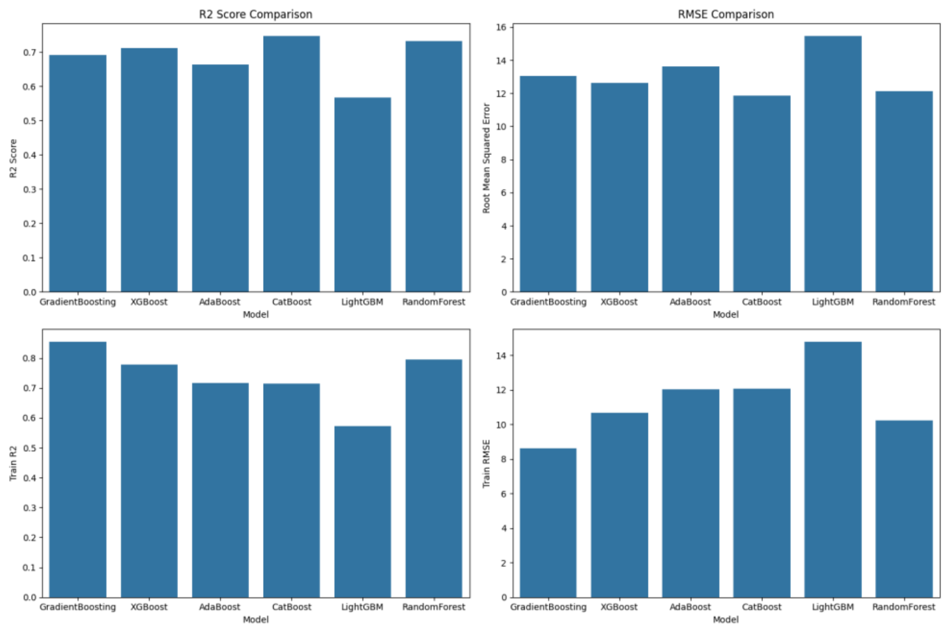


Figure 15. Learning effectiveness parameters for vertical PGA.

The results of the horizontal PGA values are generated from the preliminary study. PGA distributions are divided into two groups and ANN implementation is done for this final paper. The results of ANN implementation of the train and test dataset in Figure 15 and 16.

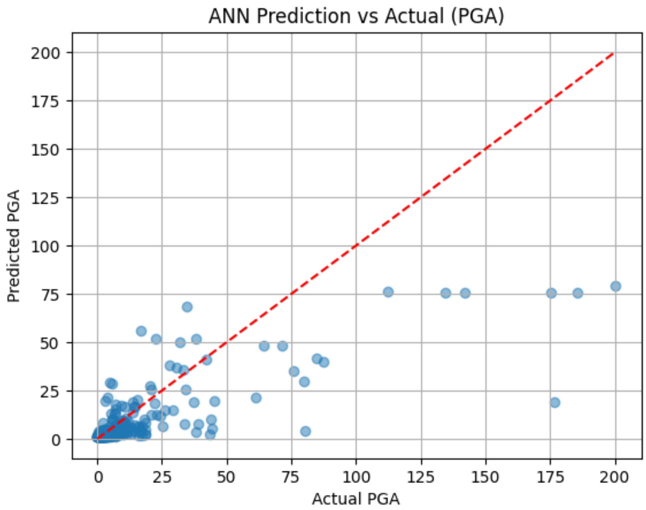


Figure 15. ANN implementation and comparison with actual data for vertical PGA.

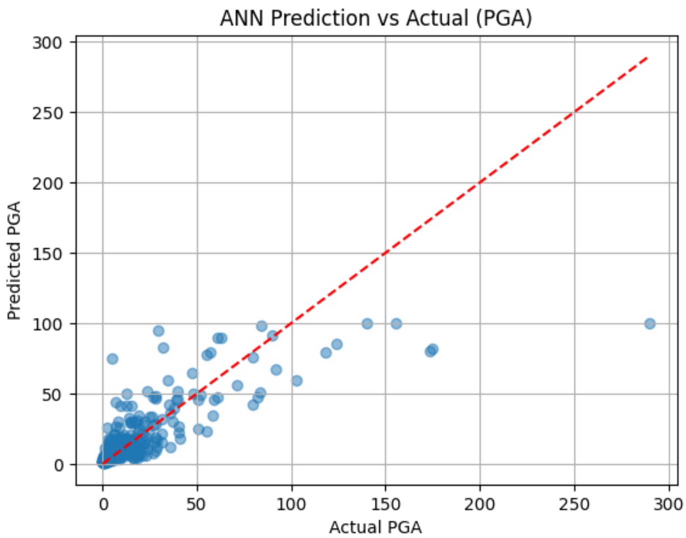


Figure 16. ANN implementation and comparison with actual data for horizontal PGA.

Table 5. Learning effectiveness parameters for ANN implementation.

	PGA (Horizontal)	PGA (Vertical)
R2	0.649	0.548
RMSE	15.803	15.778
MAPE	6.989	5.458

Discussion

For both horizontal and vertical PGA values, the XGBoost and Random Forest algorithms yielded the best results based on R^2 and RMSE metrics, while AdaBoost performed the worst. Additionally, for vertical PGA values, CatBoost produced better results compared to LightGBM.

Artificial Neural Network (ANN) implementations performed better for horizontal PGA values; however, the R^2 values for both ANN models remained low.

Overall, the results are not competitive with those reported in the literature. Nevertheless, this implementation is acceptable as a preliminary study or 'toy problem,' and, as noted, it serves as an initial step into the application of artificial intelligence in earthquake research.

CONCLUSION

This study demonstrates the application of machine learning algorithms to estimate Peak Ground Acceleration (PGA) values for earthquakes in Türkiye, focusing on events with magnitudes ranging from Mw 6.0 to 7.7. The algorithms including Random Forest (RF), XGBoost, CatBoost, LightGBM, AdaBoost, and Artificial Neural Networks (ANNs) were implemented on a curated dataset derived from AFAD strong motion records.

While the model performances do not yet match in the existing machine learning literature, the study shows a path for the use of machine learning in earthquake engineering. It serves as a preliminary investigation and the foundation for my future research career.

REFERENCES

Karimzadeh, S., Mohammadi, A., Sajad Hussaini, S. M., Caicedo, D., Askan, A., & Lourenço, P. B. (2024). ANN-based ground motion model for Turkey using stochastic simulation of earthquakes. *Geophysical Journal International*, *236*(1), 413–429. https://doi.org/10.1093/gji/ggad432

Kuran, F., Tanırcan, G., & Pashaei, E. (2024). Developing machine learning-based ground motion models to predict peak ground velocity in Turkiye. *Journal of Seismology*. https://doi.org/10.1007/s10950-024-10239-y

Linville, L., Pankow, K., & Draelos, T. (2019). Deep Learning Models Augment Analyst Decisions for Event Discrimination. *Geophysical Research Letters*, *46*(7), 3643–3651. https://doi.org/10.1029/2018GL081119

Mousavi, S. M., & Beroza, G. C. (2022). Machine Learning in Earthquake Seismology. *Annu. Rev. Earth Planet Sci.*, *37*, 23. https://doi.org/10.1146/annurev-earth-071822

Wang, J., & Teng, T.-L. (1995). Artificial Neural Network-Based Seismic Detector. In *Bulletin of the Seismological Society of America* (Vol. 85, Issue 1). http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/85/1/308/5342203/bssa0850010308.pdf

The dataset and codes are retrieved from github.com/CaglarTemiz/CENG562-Project.