PREDICTION OF STRONG GROUND MOTION PARAMETERS USING MACHINE LEARNING TECHNIQUES

by

Fahrettin Kuran

B.S., Civil Engineering, Pamukkale University, 2020

Submitted to the Kandilli Observatory and Earthquake Research Institute

In partial fulfillment of the requirements for the degree of

Master of Science

Graduate Program in Earthquake Engineering Department

Boğaziçi University

2024

PREDICTION OF STRONG GROUND MOTION PARAMETERS USING MACHINE LEARNING TECHNIQUES

APPROVED BY:

Assoc. Prof. Gülüm Tanırcan (Thesis Supervisor)	…….
Assoc. Prof. Elham Pashaei (Thesis Co-Supervisor)	…….
Prof. Ayşe Edinçliler	…….
Assist. Prof. Gökçe Tönük	…….

DATE OF APPROVAL: 29.04.2024

ACKNOWLEDGEMENTS

I would like to thank those who significantly affected my academic life up to and during my master's education.

Firstly, I want to express my deepest gratitude to my supervisors, Assoc. Prof. Dr. Gülüm TANIRCAN and Assoc. Prof. Dr. Elham PASHAEI, for their ongoing support, professional criticisms, and sincere attitude throughout the development of this research. They always provided constructive comments and feedback on this thesis. They are valuable role models who shine a light on the academic path for me. I have always felt their support by my side and I know I will. In addition, I would like to thank Assist. Prof. Dr. Gökçe TÖNÜK, who is one of the thesis jury members, for her inspiring comments and feedback.

Moreover, I am thankful to Prof. Dr. Ali PINAR, Prof. Dr. Ayşe EDİNÇLİLER, Prof. Dr. Eser ÇAKTI, Assoc. Prof. Dr. Gülüm TANIRCAN, Assist. Prof. Dr. Karin ŞEŞETYAN, and Assoc. Prof. Dr. Ufuk HANCILAR from whom I have had the opportunity to take courses during my master's education. The knowledge and experience they conveyed have paved the way for me in the field of earthquake engineering.

Furthermore, I would like to thank Bilgen ÖZBAY, Caner GÜLENÇ, Esma ARSLAN, Dr. Hakan ALÇIK, Hasibe GÜL, and Rıza PEKTAŞ, who work at the Kandilli Observatory and Earthquake Research Institute, for accompanying me on this path of my life. I am so lucky to have them with me on this journey.

I would also like to thank Assist. Prof. Dr. Ahmad Reshad NOORI, for his valuable support and encouragement during my academic process. He always guides me with his goodwill in my academic life.

Additionally, I am grateful to Prof. Dr. Şevket Murat ŞENEL and Dr. Ali KALKAN, who mentored me with their knowledge and experience throughout my undergraduate education at Pamukkale University.

ABSTRACT

PREDICTION OF STRONG GROUND MOTION PARAMETERS USING MACHINE LEARNING TECHNIQUES

Peak ground velocity (PGV) and cumulative absolute velocity (CAV) are powerful intensity measures for quantifying potential earthquake damage. Reliable prediction of those parameters is of essential importance in the precise calculation of seismic hazard. Machine learning can provide accurate and reliable predictions of PGV and CAV due to handling nonlinear relationships, adaptability to changing conditions, automation, efficiency, and the potential for real-time predictions. This study aims to comprehensively compare the relative performance of six different machine learning algorithms for PGV and CAV prediction and develop the Turkiye-specific ground motion models with the most recent strong motion dataset. Support Vector Machine, Linear Regression, Random Forest, Artificial Neural Network, Bayesian Ridge Regression, and Gradient Boosting algorithms are evaluated and compared. The Turkish strong motion database, which consists of over 950 earthquakes occurring from 1983 to 2023, and a worldwide near-field earthquake database (moment magnitude≥5.5) are used for shaping the models' ability to learn and make accurate predictions. Various source, site, distance, and faulting parameters are considered as estimator parameters. Three feature selection methods, embedded, filter, and wrapper, are applied to determine the most suitable estimator parameters to predict PGV and CAV. Statistical evaluation metrics are employed to measure the performance of the models. Among algorithms, the Gradient Boosting algorithm shows remarkable success in the prediction of both PGV and CAV. PGV prediction performance is better when all estimator parameters are used together, while CAV prediction ability is better when user-selected parameters are utilized. Outlier elimination processes are found to be redundant in predicting the PGVs and CAVs of large earthquakes. The proposed PGV and CAV models are applicable to shallow crustal strike-slip and normal faulting earthquakes with moment magnitude ranging from 3.5 to 7.8 and Joyner and Boore distance up to 200 km.

ÖZET

MAKİNE ÖĞRENMESİ TEKNİKLERİNİ KULLANARAK KUVVETLİ YER HAREKETİ PARAMETRELERİNİN TAHMİNİ

En büyük yer hızı (PGV) ve kümülatif mutlak hız (CAV), potansiyel deprem hasarını ölçmek için kullanılan önemli kuvvetli yer hareketi parametrelerindendir. Makine öğrenmesi, doğrusal olmayan ilişkilerin ele alınması, otomasyon, verimlilik ve gerçek zamanlı tahmin potansiyeli sayesinde PGV ve CAV parametrelerinin daha doğru ve güvenilir tahminlerini sağlayabilir. Bu çalışma, altı farklı makine öğrenmesi algoritmalarının PGV ve CAV tahmini için göreceli performansını kapsamlı bir şekilde karşılaştırmayı ve en güncel kuvvetli yer hareketi veri seti ile Türkiye'ye özgü yer hareketi modellerini geliştirmeyi amaçlamaktadır. Destek Vektör Makinesi (Support Vector Machine), Doğrusal Regresyon (Linear Regression), Rastgele Orman (Random Forest), Yapay Sinir Ağı (Artificial Neural Network), Bayesian Sırt Yolu Regresyonu (Bayesian Ridge Regression) ve Gradyan Artırma (Gradient Boosting) algoritmaları değerlendirilerek karşılaştırılmıştır. Modellerin öğrenme ve doğru tahmin yapma yeteneklerinin şekillendirilmesinde, 1983'ten 2023'e kadar meydana gelmiş 950'den fazla depremden oluşan Türkiye kuvvetli yer hareketi veri tabanı ve dünya çapındaki yakın alan deprem veri tabanı (moment büyüklüğü≥5.5) kullanılmıştır. PGV ve CAV parametrelerinin tahmin edilmesi amacıyla, en uygun girdi parametreleri gömülü (embedded), filtre (filter) ve kapsayıcı (wrapper) olmak üzere üç öznitelik seçim (feature selection) yöntemi ile belirlenmiştir. Algoritmalar arasında, Gradyan Artırma algoritması hem PGV hem de CAV tahmininde dikkate değer bir başarı göstermiştir. PGV tahmin performansı, tüm girdi (all) parametreleri bir arada kullanıldığında daha iyiyken, CAV tahmin yeteneği, kullanıcı tarafından seçilen (user-selected) parametreler kullanıldığında daha iyidir. Büyük depremlerin PGV ve CAV parametrelerini tahmin etmede, aykırı değerleri eleme (outlier elimination) işlemlerinin gereksiz olduğu bulunmuştur. Önerilen PGV ve CAV modelleri, moment büyüklüğü 3.5 ile 7.8 arasında değişen ve Joyner ve Boore mesafesi 200 km'ye kadar olan sığ kabuklu doğrultu atımlı ve normal faylanma depremlerine uygulanabilir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

ABSTRACT iv

ÖZET v

LIST OF FIGURES viii

LIST OF TABLES xi

LIST OF SYMBOLS xiv

LIST OF ACRONYMS/ABBREVIATIONS xvi

1. INTRODUCTION 1

2. STRONG GROUND MOTION DATABASE 9

3. METHODOLOGY 13

3.1. Data Preprocessing 13

3.2. Feature Selection 17

3.3. A Brief Overview of ML Algorithms 18

3.3.1. Linear Regression 19

3.3.2. Random Forest 19

3.3.3. Support Vector Machine 19

3.3.4. Bayesian Ridge Regression 20

3.3.5. Artificial Neural Network 20

3.3.6. Gradient Boosting 20

3.4. Performance Evaluation Metrics 21

3.5. Hyperparameter Tuning 21

4. ANALYSIS AND RESULT 23

4.1. First Evaluation: Effect of Dataset and Outlier Elimination on Prediction . 23

4.1.1. PGV Result of the First Evaluation 24

4.1.2. CAV Result of the First Evaluation 27
4.1.3. Verification Effort with the NESS Dataset 30
4.2. Second Evaluation: Comparison of Feature Selection and ML Methods ... 36
4.2.1. PGV Result of the Second Evaluation 40
4.2.2. CAV Result of the Second Evaluation 49
4.3. Development of Turkiye-specific GMMs for PGV and CAV Using the Expanded N-TSMD 54
4.4. Comparison of the Proposed Model with the Existing GMMs 57
5. CONCLUSION 62
REFERENCES 65
APPENDIX A: THE DEVELOPMENT OF THE WEBSITE FOR THE APPLICATION OF THE PGV PROPOSED MODEL 75
APPENDIX B: THE DEVELOPMENT OF THE WEBSITE FOR THE APPLICATION OF THE CAV PROPOSED MODEL 76

LIST OF FIGURES

Figure 2.1. Epicentral distribution of earthquakes in the expanded N-TSMD. Grey lines are the fault traces [74]. 11

Figure 2.2. Style-of-faulting distribution of earthquakes (a) and site class of recordings according to NEHRP site classification (b) in the N-TSMD, SMD-TR, and NESS (A: Vs_{30}>1500 m/s, B: 760 m/s<Vs_{30}≤1500 m/s, C: 360 m/s<Vs_{30}≤760 m/s, D: 180 m/s<Vs_{30}≤360 m/s, and E: Vs_{30}<180 m/s). .. 11

Figure 2.3. Distribution of CAVs/PGVs vs. distance parameters; R_{JB}, R_{rup}, and R_{epi}. 12

Figure 3.1. Flowchart followed in this study. 13

Figure 3.2. The representation of interquartile range (IQR). 14

Figure 3.3. Outlier detection in the N-TSMD for (a) the single dataset (15,110 recordings), (b) the small-magnitude separated dataset (14,102 recordings), and (c) the moderate-to-large magnitude separated dataset (1008 recordings) (Red dots define outliers). 16

Figure 3.4. Schematic diagram of datasets used in the ML algorithms (The number of recordings indicated in parenthesis. Blue, orange, and green boxes define original, outlier-eliminated, and worldwide test datasets, respectively). 17

Figure 4.1. Performance metrics (in normalized form) of six ML algorithms trained with seven estimator parameters using the N-TSMD original and outlier-eliminated datasets for the prediction of PGV. 26

Figure 4.2. Performance metrics (in normalized form) of six ML algorithms trained with seven estimator parameters using the N-TSMD original and outlier-eliminated datasets for the prediction of CAV. 29

Figure 4.3. Distribution of performance metrics of six trained ML algorithms for PGV prediction of the NESS dataset. 33

Figure 4.4. Performance metrics of six trained ML algorithms for CAV prediction of the NESS dataset. 35

Figure 4.5. Pearson correlation coefficients of features for (a) the $5.5 \leq M_w \leq 7.6$, (b) the PGV $3.5 \leq M_w \leq 5.4$, and (c) the $3.5 \leq M_w \leq 7.6$ datasets. 37

Figure 4.6. Pearson correlation coefficients of features for (a) the CAV $5.5 \leq M_w \leq 7.6$ and (b) the CAV $3.5 \leq M_w \leq 5.4$ datasets. 37

Figure 4.7. Feature importance results obtained by LASSO for (a) the PGV $5.5 \leq M_w \leq 7.6$, (b) the PGV $3.5 \leq M_w \leq 5.4$, and (c) the PGV $3.5 \leq M_w \leq 7.6$ datasets. 38

Figure 4.8. Feature importance results obtained by LASSO for (a) the CAV $5.5 \leq M_w \leq 7.6$ and (b) the CAV $3.5 \leq M_w \leq 5.4$ datasets. 39

Figure 4.9. Results of PGV prediction analyses performed with the original $3.5 \leq M_w \leq 7.6$ dataset according to performance metrics. 43

Figure 4.10. Results of PGV prediction analyses performed with the original $5.5 \leq M_w \leq 7.6$ dataset according to performance metrics. 45

Figure 4.11. Results of PGV prediction analyses performed with $3.5 \leq M_w \leq 5.4$ dataset according to performance metrics. 47

Figure 4.12. Results of CAV prediction analyses performed with the original $5.5 \leq M_w \leq 7.6$ dataset according to performance metrics. 50

Figure 4.13. Results of CAV prediction analyses performed with $3.5 \leq M_w \leq 5.4$ dataset according to performance metrics. 52

Figure 4.14. Residuals of PGVs predicted with the proposed model (Yellow squares show the mean residuals and yellow error lines are their standard deviations. Blue and red dots represent $M_w \leq 5.4$ and $M_w \geq 5.5$ recordings, respectively). 56

Figure 4.15. Residuals of CAVs predicted with the proposed model (Yellow squares show the mean residuals and yellow error lines are their standard deviations. Blue and red dots represent $M_w \leq 5.4$ and $M_w \geq 5.5$ recordings, respectively). 56

Figure 4.16. CAVs predicted with the TR-CB19, TR-SA17, and proposed model for SoF= SS (δ=90° and λ=0°) and N (δ=60° and λ=-90°) with R_{JB}=[1:200], M_w=[5.0, 6.0, 7.0], and Vs_{30}=[550 m/s]. 59

Figure 4.17. Distribution of the near-field earthquakes (10 km<R_{JB}) in the expanded N-TSMD. 60

Figure 4.18. PGVs predicted with the BSSA14, CY14, KMHCAL23, and proposed model for SoF= SS (δ=90° and λ=0°) and N (δ=60° and λ=-90°) with R_{JB}=[1:200], M_w=[5.0, 6.0, 7.0], Vs_{30}=[550 m/s], and FD=[7 km]. 61

Figure A.1. The software interface of Turkiye-specific ground motion model for peak ground velocity. 75

Figure B.1. The software interface of Turkiye-specific ground motion model for cumulative absolute velocity. 76

LIST OF TABLES

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters. 6

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters (cont.). 7

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters (cont.). 8

Table 2.1. Minimum and maximum values of parameters in the N-TSMD and NESS (* the expanded N-TSMD with the SMD-TR). 10

Table 3.1. Hyperparameter tuning parameters of ML techniques for (a) CAV and (b) PGV output parameters. 22

Table 4.1. Performance metrics (in normalized form) of PGV prediction with six ML algorithms trained with the N-TSMD original datasets. Bold values represent the best results. 25

Table 4.2. Performance metrics (in normalized form) of PGV prediction with six ML algorithms trained with the N-TSMD outlier-eliminated datasets. Bold values represent the best results. 25

Table 4.3. Performance metrics (in normalized form) of CAV prediction with six ML algorithms trained with the N-TSMD original datasets. Bold values represent the best results. 28

Table 4.4. Performance metrics (in normalized form) of CAV prediction with six ML algorithms trained with the N-TSMD outlier-eliminated datasets. Bold values represent the best results. 28

Table 4.5. Performance metrics (in normalized form) of six trained ML algorithms for PGV prediction of the NESS dataset. 32

Table 4.6. Performance metrics (in normalized form) of six trained ML algorithms for CAV prediction of the NESS dataset. 34

Table 4.7. Summary of determining feature sets for (a) the PGV $3.5 \leq M_w \leq 5.4$, (b) the PGV $5.5 \leq M_w \leq 7.6$, and (c) the PGV $3.5 \leq M_w \leq 7.6$ datasets..... 39

Table 4.8. Summary of determining feature sets for (a) the CAV $3.5 \leq M_w \leq 5.4$ and (b) the CAV $5.5 \leq M_w \leq 7.6$ separated datasets..... 40

Table 4.9. Summary of comparison of ML techniques for PGV on $3.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models. 44

Table 4.10. Summary of comparison of ML techniques for PGV on $5.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models. 46

Table 4.11. Summary of comparison of ML techniques for PGV on $3.5 \leq M_w \leq 5.4$ datasets using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models. 48

Table 4.12. Summary of comparison of ML techniques for CAV on $5.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d)

MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models. 51

Table 4.13. Summary of comparison of ML techniques for CAV on $3.5 \leq M_w \leq 5.4$ datasets using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models. 53

Table 4.14. Comparison of the readvanced model and the model obtained with only the N-TSMD for PGV according to performance metrics (in normalized form). 54

Table 4.15. Comparison of the readvanced model and the model obtained with only the N-TSMD for CAV according to performance metrics (in normalized form). 55

Table 4.16. Summary of estimator parameters used in GMMs for CAV. 58

Table 4.17. Summary of estimator parameters used in GMMs for PGV. 58

LIST OF SYMBOLS

D_{vs}=800	The depth of the soil when the shear wave velocity reaches 800 m/s profile depth
FD	Focal depth (km)
F_{HW}	Hanging wall effect
HW	Hanging wall indicator
I_A	Arias intensity
M_L	Local magnitude
M_{ms}	Moment magnitude of M_S
M_S	Surface wave magnitude
M_w	Moment magnitude
R	Correlation coefficient
R^2	Coefficient of determination
R_{ClstD}	Closest distance between the station and the rupture surface (km)
R_{epi}	Epicentral distance (km); distance from the epicenter to a specified site
R_{hyp}	Hypocenter distance (km); closest distance between focus and site
R_{JB}	Joyner-Boore distance (km); closest distance to surface projection of rupture plane
R_{ms}	Joyner–Boore distance of M_S
R_{rup}	Rupture distance (km); closest distance to the ruptured plane
R_x	Horizontal distance from top of rupture measured perpendicular to fault strike (km)
S_a	Spectral acceleration
Sa_{as}	Spectral acceleration of A_S
Sa_{ms}	Spectral acceleration of M_S
SoF	Style-of-faulting
Vs_{20}	Shear wave velocity of the top 20 m of the soil

Vs_{30}	Shear wave velocity of the top 30 m of the soil
V_{se}	Equivalent shear wave velocity
W	Fault rupture width
$Z_{1.0}$	Basin depth (km); depth from the ground surface to the 1km/s shear-wave horizon
Z_{hyp}	Hypocenter depth (km)
Z_{TOR}	Depth to the top of ruptured plane (km)
δ	Dip angle (°)
∇_M	M_S–A_S magnitude ratio
∇_M/R_{ms}	Simulated saturation effects in the near-field region
∇_R	M_S–A_S distance ratio
λ	Rake angle (°)

LIST OF ACRONYMS/ABBREVIATIONS

AFAD	Disaster and Emergency Management Authority
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
BSSA14	Boore et al., 2014
BRR	Bayesian ridge regression
CAV	Cumulative absolute velocity
CGAN	Conditional generative adversarial network
COSMOSs	Consortium of Organizations for Strong Motion Observation Systems
CWBSN	Central Weather Bureau Seismological Center of Taiwan
CY14	Chiou and Youngs, 2014
DNN	Deep neural network
DT	Decision tree
EPRI	Electric Power Research Institute
ESM	European Strong Motion database
FP	Predominant frequency
GB	Gradient boosting
GDDA	General Directorate of Disaster Affairs
GMIMs	Ground motion intensity measures
GMMs	Ground motion models
GMPEs	Ground motion prediction equations
I. I. D	Non-independent and identically distributed
IMs	Intensity measures
IQR	Interquartile range
IRIS	Incorporated Research Institutions for Research
KMHCAL23	Karimzadeh et al., 2023
K-NET/Kik-NET	Strong-motion seismograph Networks
KOREI	Kandilli Observatory and Earthquake Research Institute
LASSO	Least absolute shrinkage and selection operator
LR	Linear regression

MAE	Mean absolute error
MAPE	Mean absolute percent error
ML	Machine learning
MLP	Multilayer perceptron
MSE	Mean squared error
N	Normal
NECIS	National Earthquake Comprehensive Information System
NEHRP	National Earthquake Hazards Reduction Program
NESS	NEar-Source Strong-motion flat-file
N-TSMD	New Turkish Strong Motion Database
OBT	Overburden thickness
PBA	Peak bedrock
PEER NGA-West2	PEER Shallow Crustal Earthquakes in Active Tectonic Regimes
PGA	Peak ground acceleration
PGD	Peak ground displacement
PGV	Peak ground velocity
PSAs	Pseudo-spectral accelerations
PSHA	Probabilistic seismic hazard assessment
R	Reverse
RANFIS	Randomized adaptive neuro-fuzzy inference system
RESORCE	Reference database for Seismic grOund-motion pRediction in Europe
RF	Random forest
RFE	Recursive feature elimination
RMSE	Root mean squared error
RSO-DNN	Refined second-order deep neural network
SD	Significant duration
SFP	Site fundamental period
SMD-TR	Strong-motion database of Turkiye
SS	Strike-slip
SVM	Support vector machine
SVR	Support vector regression
TR-CB19	Turkiye-adjusted Campbell & Bozorgnia, 2019
TR-SA17	Turkiye-adjusted Sandıkkaya & Akkar, 2017

USGS	United States Geological Survey
XGBoost	Extreme gradient boosting

1. INTRODUCTION

Peak ground velocity (PGV) and cumulative absolute velocity (CAV) are two frequently used intensity metrics in earthquake engineering studies. They are employed in a variety of different manners such as seismic hazard analysis [1], seismic risk analysis [2], soil structure interaction problems [3], soil liquefaction/buried pipelines [4-8], microzonation of earthquake hazard area [9], earthquake early warning [10-15], damage and shaking intensity [16], structural health monitoring [17], and liquefaction hazard evaluation [18].

PGV can be defined as the maximum absolute amplitude of the ground velocity during a time, which is commonly calculated by integrating the acceleration data, as seen equation 1.1. PGV is one of the most important intensity measures used as a threshold quantity for various early warning algorithms [13-15] and the most reliable ground motion parameter for predicting the dynamic performance of the soil-pile system in liquefying ground [4]. CAV was put forward as an alternative ground motion intensity measure (GMIM) for damage in 1988 by the Electric Power Research Institute (EPRI). Equation 1.2 provides the most commonly used mathematical definition of CAV [19]. Recent research in earthquake engineering [1], [20] has demonstrated that CAV is a valuable intensity metric in a range of structural and geotechnical engineering applications. Campbell & Bozorgnia [21] suggest that CAV empirical models have much smaller standard deviations compared to other Intensity Measures (IMs) such as Arias intensity (I_A) [22], which makes them appealing for usage in different engineering applications.

$$\mathrm{PGV} = \max(\left|\acute{u}_g(t)\right|) \tag{1.1}$$

$$\mathrm{CAV} = \int_0^{tmax} |a(t)|dt \tag{1.2}$$

$\acute{u}_g(t)$, t_{max}, and $|a(t)|$ are the ground velocity at time t, the total duration of the time series, and the absolute value of acceleration time series at time t, respectively.

Many Ground Motion Prediction Equations (GMPE)s are available for PGV (Zhang et al., 2022 [23]; Montalva et al., 2022 [24]; Bindi et al., 2014 [25]; Boore et al., 2014 [26] (hereafter BSSA14); Chiou and Youngs, 2014 [27] (hereafter CY14); Boore et al., 2008 [28]; Akkar et al., 2007 [29]) and CAV estimation (Bullock et al., 2021 [30]; Campbell & Bozorgnia, 2010 [31], 2014 [32], 2019 [33] (hereafter CB10, CB14, and CB19); Danciu & Tselentis, 2007 [34]; Du & Wang, 2013 [35] (hereafter DW13); Foulser-Piggott & Goda, 2014 [36], 2015 [37] (hereafter FPG14 and FPG15); Sandıkkaya & Akkar, 2017 [38] (hereafter SA17); Tetik et al., 2022 [39]), but relatively few ground motion models (GMM)s (specifically, CB19, SA17, BSSA14, and CY14) which are based on a large dataset, are available in the literature for PGV and CAV predictions.

BSSA14 [26] is one of the most comprehensive GMMs for PGV derived from a subset of the NGA-West2 database consisting of $3.0 \leq M_w \leq 7.9$ earthquakes [40]. Their model is appropriate for $3.0 \leq M_w \leq 8.5$ Strike-slip (SS) and $3.0 \leq M_w \leq 7.0$ Normal (N) earthquakes at sites with shear wave velocity of the top 30 m of the soil (Vs_{30}) in the range of 150 m/s and 1500 m/s and distances between 0 and 400 km. M_w, Joyner-Boore distance (R_{JB}), Vs_{30}, and style-of-faulting (SoF) are the estimator parameters utilized for PGV prediction. CY14 [27] PGV ground motion model is an update of their previous equation [41] with an enlarged database [40]. The essential estimator variables used in their analyses are M_w, distances (R_{rup}, R_{JB}, R_x), shear wave velocity parameters (including Vs_{30}, depth from the ground surface to the 1km/s shear-wave horizon ($Z_{1.0}$), depth to the top of the ruptured plane (Z_{TOR})), hanging wall parameters, faulting properties as well as directivity effect properties. Their model is appropriate for $3.5 \leq M_w \leq 8.5$ SS, $3.5 \leq M_w \leq 8.0$ N and reverse (R) earthquakes in active tectonic regions at sites with Vs_{30} in the range of 180 m/s and 1500 m/s and rupture distance (R_{rup}) between 0 and 300 km.

In CB19 [33] GMMs for CAV, a subset of the NGA-West2 database is used [40]. The model was developed using the functional form of CB14 for the CAV and I_A. It is applicable to M_w between 3.0 and 8.0, R_{rup} shorter than 300 km, and Vs_{30} between 100 m/s and 1200 m/s. It contains hanging wall, magnitude, geometric attenuation, shallow site response, style of faulting, anelastic attenuation basin response, fault dip, and hypocentral depth terms. CAV has the best goodness-of-fit statistics of all GMIMs they have thus far studied. SA17 [38] model was developed for CAV, I_A, and Significant Duration (SD) using a subset of the Pan-

European RESORCE database [42]. The model is appropriate for $4.0 \leq M_w \leq 8.0$, distances (R_{JB}, Epicentral distance (R_{epi}), and Hypocenter distance (R_{hyp})) shorter than 200 km, and Vs_{30} between 150 m/s and 1200 m/s. The functional form is rather straightforward. It is composed of M_w, a selected distance parameter among R_{JB}, R_{epi}, and R_{hyp}, Vs_{30}, SoF (R, SS, or N faulting) effects, and regression coefficients. The shallow crustal earthquake recordings in the RESORCE strong motion database are from worldwide but approximately one-third of the earthquakes and recordings are compiled from the Turkish strong motion networks. Recently, Tetik et al. [39] evaluated the performance of CB19 and SA17 models using the New Turkish Strong Motion Database (N-TSMD) [43] by implementing a data-driven testing approach. These models were adjusted with constant shifts and applied to the inter-event residuals (hereafter TR-CB19 and TR-SA17).

On one hand, conventional regression equations, which are easily applied with just a few parameters and have a high overall prediction success, are often used in various engineering applications. On the other hand, it is difficult to fully identify the complex and nonlinear features of the data in conventional GMPEs produced via parametric regression due to the presumption of a fixed functional form. It is particularly well known that, at near-fault regions, large PGVs formed by pulse-type low-frequency motions may require higher seismic demands from structures than those specified in design codes. Hence, the prediction of PGV is of utmost importance for the seismic design of long-period structures. However, PGVs in the near-field have yet to be predicted precisely due to unexpected properties of the source rupture process. For instance, observed geometric mean PGVs at 10-20 km distance from fault trace are four times larger than the median predictions of the most recent 2023 (M_w7.8) Kahramanmaraş doublet earthquakes of Turkiye.

Recently, the concept of machine learning (ML) has taken its place in the field of earthquake engineering, as in every field. In this regard, researchers use different ML algorithms for the prediction of GMIMs. Detailed studies about the prediction of strong ground motion parameters with ML techniques are given in Table 1.1. Among them, only two studies [44, 45] are available for CAV prediction using ML techniques. There are relatively more studies [46-57] for PGV prediction using ML techniques.

Dhanya and Raghukanth [49] studied Artificial Neural Network (ANN) to obtain a plausible GMPE for peak ground acceleration (PGA), PGV, and pseudo-spectral accelerations (PSA)s. They utilized 13,552 recordings from 288 shallow crustal earthquakes with M_w, R_{rup}, Vs_{30}, and SoF as estimator parameters. Their sensitivity study revealed that R_{rup} and M_w are the controlling estimators for the prediction of IMs. They concluded that their model's performance is seen to fall within the acceptable limits of error. Junjie et al. [46] used random forest (RF) and ANN algorithms with 6944 recordings from 32 earthquakes ($4.8 \leq M_w \leq 9.0$) for the prediction of PGV. In addition, they developed new attenuation relationships for PGV and compared their equation with RF and ANN methods. Their test results express that ML techniques predict PGV for near-fault earthquakes more accurately than the prediction equation. Nevertheless, they observed that ML models relatively overestimate PGV.

Ji et al. [45] utilized a Refined Second-Order Deep Neural Network (RSO-DNN) to obtain a GMM for the prediction of CAV by considering a variety of estimator parameters including fault geometry, fault type, earthquake source size in addition to magnitude, distance, and site measures. The utilized PEER NGA-West2 dataset comprises 17,684 recordings from 466 shallow crustal earthquakes ($3.0 \leq M_w \leq 8.0$). They also compared their ML-based predictions with regression-based GMPEs (CB19, SA17, DW13, CB10) in terms of mean absolute error (MAE), mean square error (MSE), and coefficient of determination (R^2) performance evaluation metrics. Hu et al. [44] employed the Support Vector Regression (SVR) method to produce regional GMMs for the prediction of CAV, SD, and I_A with three estimator parameters including M_w, R_{rup}, and Vs_{30}. They compared their SVR-GMMs with observed data and the existing GMMs such as FPG14 and CB10. The dataset used in their study consists of 15,960 recordings of 130 earthquakes with a magnitude of $5.0 \leq M_{JMA} \leq 8.0$ and $R_{rup} \leq 200$ km in the Kanto region, Japan. Seismic sources in their study consisted of 36 shallow-crustal, 48 upper mantle, 14 subduction interfaces, and 36 subduction slab events. The correlation coefficient (R) was used to assess the prediction performance. They found that amplitude behavior with increasing distance of SVR-GMMs is close to the FPG14.

More recently, Karimzadeh et al. [57] studied the ANN algorithm with 7359 simulated recordings for the prediction of PGA, PGV, and PSAs. The focal depth (FD), SoF, M_w (ranging from 5.0 to 7.5), R_{JB} (ranging from 0 to 272 km), and Vs_{30} are used in their

GMM for Turkiye (KMHCAL23). Karimzadeh et al. [55] also used the eXtreme Gradient Boosting (XGBoost) algorithm to obtain a local GMM for PGA, PGV, and PSAs. They utilized 247,710 simulated recordings with M_w (ranging from 5.0 to 6.8), R_{JB} (ranging from 0 to 150 km), and FD (ranging from 5 to 17 km) as estimator parameters for Portugal. Their XGBoost model indicated a persuasive performance; however, the attenuation rate of the model decreases rapidly. Mohammadi et al. [56] also studied ANN and XGBoost algorithms to develop Turkiye-based GMMs for PGA, PGV, and PSAs. They utilized 1166 real recordings of 383 earthquakes from the Disaster and Emergency Management Presidency of Turkiye (AFAD) and the RESORCE database to Turkiye-specific GMMs. Their estimator parameters are M_w, R_{JB}, Vs_{30}, and SoF. Their findings show that the XGBoost algorithm produces an unbiased model with a more significant correlation and a lower residual than ANN.

Considering the overall advantages of ML-based prediction algorithms, this study aims to provide to build ML-based Turkiye-specific GMMs for PGV and CAV using real earthquake recordings. Even though ML-based GMM studies are increasing rapidly in the literature, a detailed investigation of the processing steps to be applied to the dataset and additional techniques to increase prediction accuracy have yet to be carried out for PGV and CAV so far. In addition to creating Turkiye-specific GMM, possible applications of the dataset, the consideration of outliers in the model development, and the efficiency of feature selection methods are discussed in detail. It is aimed to advance the field and pave the way for future studies. The evaluation process has involved testing the obtained models using a global strong motion dataset consisting of moderate-to-strong earthquakes from an independent source. Six well-known ML techniques—Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GB), Bayesian Ridge Regression (BRR), Linear Regression (LR), and Artificial Neural Network (ANN)—were adapted to predict the geomean PGV and CAV of two horizontal components. The primary components of the assessment include the acceptability of the models, reliability, precision in the output, and validity for future data.

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters.

Literature	ML Technique	Data	Region	Source of Data	Estimator Parameters	Output	Evaluation metric(s)
Karimzadeh et al. (2023) [55]	XGBoost	247,710 recordings	Portugal	Simulated	M_w, R_{JB}, FD	PGA, PGV, PSAs	RMSE MAPE R^2 R
Mohammadi et al. (2023) [56]	ANN, XGBoost	1166 recordings from 383 earthquakes	Turkiye	AFAD RESORCE	M_w, R_{JB}, Vs_{30}, SoF	PGA, PGV, PSAs	RMSE R^2 R
Karimzadeh et al. (2023) [57]	ANN	7359 recordings	Turkiye	Simulated	SoF, FD, M_w, R_{JB}, Vs_{30}	PGA, PGV, PSAs	R^2 R RMSE MAPE
Li et al. (2023) [58]	RSO-DNN	20,651 recordings from 401 earthquakes	Worldwide	PEER NGA-West2	T, M_w, SoF, Region, Z_{TOR}, Z_1, ln(R_{JB}), ln(Vs_{30})	PGA, PGV, and PSAs	R^2 RMSE MAE
Hu et al. (2022) [44]	SVR	15,960 recordings from 130 earthquakes	Japan	K-NET	M_w, R_{rup}, Vs_{30}	CAV, SD, and I_A	R K and K′ R^2_o, R'^2_o MAE MSE
Seo et al. (2022) [59]	ANN, RF, GB	1189 recordings from 77 earthquakes	Korea	NECIS	M_L, R_{epi}, Vs_{30}, FD, and slope angle	PSAs	RMSE MAE R^2
Qi et al. (2022) [60]	RF, AdaBoost, XGBoost	3106 recordings	Japan	Kik-NET	PBA, FP, site type, SFP, OBT, $Dv_{s=800}$, profile depth, Surface Vs, Bedrock Vs, Vse, Vs_{20}, and Vs_{30}	PGA	R^2 MAE MAPE MSE
Khosravikia & Clayton (2021) [61]	ANN, RF, SVM	4528 recordings from 374 earthquakes	Texas, Oklahoma, and Kansas	IRIS	M_w, R_{hyp}, and Vs_{30}	PGA and PSAs	R σ τ and ϕs ϕε

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters (cont.).

Literature	ML Technique	Data	Region	Source of Data	Estimator Parameters	Output	Evaluation metric(s)
Ji et al. (2021) [62]	RSO-DNN	20,900 recordings from 428 earthquakes	Worldwide	PEER NGA-West2	T, M_w, SoF, Region, Z_{TOR}, Z_1, R_{JB}, Vs_{30}	PGA, PGV, and PSAs	R^2 RMSE MAE
Ji et al. (2021) [45]	RSO-DNN	17,684 recordings from 466 earthquakes	Worldwide	PEER NGA-West2	M, SoF, δ, λ, Z_{TOR}, Z_{hyp}, R_{rup}, W, Vs_{30}, F_{HW}	CAV	R^2 MSE MAE
Ding et al. (2021) [63]	DNN, CGAN	503 sets of MS–AS recordings	Worldwide	PEER NGA-West2	M_{ms}, ∇M, R_{ms}, ∇M/R_{ms}, ∇R, HD, Vs_{30}, SoF, Sa_{ms}, and Sa_{as}	PSAs	PP R MAE MAPE RMSE
Junjie et al. (2021) [46]	RF and NN	6944 recordings from 32 earthquakes	Japan	K-NET and KiK-net	M_w, R_{ClstD}, H, and the dummy variable (Si)	PGV	R^2 MAE RMSE
Khosravikia et al. (2019) [64]	ANN	4528 recordings from 376 earthquakes	Kansas, Texas, and Oklahoma	IRIS	M_w, R_{hyp}, and Vs_{30}	PGA, PGV, and PSAs	R K, K' R^2_o, R'^2_o, m_{index}, n_{index}, R^2_m
Derakhshani & Foruzan (2019) [48]	DNN	12,556 recordings	Worldwide	PEER NGA-West2	M_w, λ, R_{ClstD}, and Vs_{30}	PGA, PGV, PGD	R MAE MAPE RMSE
Öztürk (2019) [65]	ANN	195 recordings from 18 earthquakes	Turkiye	AFAD and USGS	M_w, Site classification, R_{epi}, FD, SD, I_A, PGA, PGV, PGD	The macroseismic (or felt) intensity	MSE R
Dhanya & Raghukanth (2018) [49]	ANN	13,552 recordings from 288 earthquakes	Worldwide	PEER NGA-West2	SoF, M_w, Vs_{30}, and R_{rup},	PSAs, PGA, and PGV	R PP k, k' a, a' R^2_o, R'^2_o
Khosravikia et al. (2018) [47]	ANN (MLP)	4529 recordings from 374 earthquakes	Kansas, Texas, and Oklahoma	IRIS	M_w, R_{JB}, Vs_{30}	PGA, PGV	Residuals

Table 1.1. Literature review of ML techniques used in the prediction of strong ground motion parameters (cont.).

Literature	ML Technique	Data	Region	Source of Data	Estimator Parameters	Output	Evaluation metric(s)
Thomas et al. (2016) [50]	ANFIS (RANFIS)	2815 recordings	Worldwide	PEER-NGA Flatfile V 7.3	M_w, SoF, R_{ClstD}, and Vs_{30}.	PGA, PGV, PGD	RMSE MAE MAPE MSE R
Derras et al. (2014) [51]	ANN	1088 recordings from 320 earthquakes	Europe	RESORCE	M_w, R_{JB}, SoF, hypocentral depth, and Vs_{30}.	PSAs, PGA, and PGV	Residuals
Derras et al. (2012) [66]	ANN	3891 recordings from 335 earthquakes.	Japan	KiK-net	M_w, FD, R_{epi}, the site resonance frequency (f_0), and Vs_{30}	PGA	MSE S K
Tezcan & Cheng (2012) [67]	SVR	1482 recordings from 94 earthquakes	Worldwide	PEER-NGA Flatfile V 7.2	M_w, R_{ClstD}, Vs_{30}, and SoF	PSAs	MSE R^2
Alavi & Gandomi (2011) [52]	ANN, ANN/SA	2815 recordings	Worldwide	PEER-NGA Flatfile V 7.3	SoF, M_w, R_{ClstD}, and Vs_{30}	PGA, PGV, PGD	OBJ MAPE R MAE MSE
Ahmad et al. (2008) [53]	ANN	358 recordings from 42 shallow earthquakes	Europe	ESM	Ms, R, and broad categories of soil type	PGA, PGV, PGD	Residuals
Günaydın & Günaydın (2008) [68]	ANN (FFBP, RBF, GRNNs)	15 ground motions from 11 stations	Turkiye	COSMOSs-KOREI	M_w, R_{hyp}, R_{epi}, FD, Site conditions	PGA	RMSE MAE R
Güllü & Erçelebi (2007) [69]	ANN (developed)	3 datasets (These sets are 210, 47, and 221 recordings.)	Turkiye	GDDA-KOERI	M_w, R, and site conditions	PGA	R^2
B.-Y. Liu et al. (2006) [54]	Bayesian Regularization Back Propagation NN	472 recordings from 69 earthquakes	America	USGS	M, R_{epi}, site condition, and site intensity	PGV	Errors statistic
Kerh & Ting (2005) [70]	ANN (back-propagation)	Historical seismic recordings at 30 stations from 8 to 39 data sets for each station	Taiwan	CWBSN	R_{epi}, FD, and M	PGA	R^2

2. STRONG GROUND MOTION DATABASE

The state-of-the-art strong ground motion database—The New Turkish Strong Motion Database (N-TSMD) [43]—is utilized to both train and test processes of in this study. The acceleration time histories belong to 904 strong motion stations operated by the AFAD. A total of 743 earthquakes in the time span between 1983 and 2020 constitutes the preliminary version of N-TSMD (Figure 2.1). The N-TSMD consists of earthquakes with $3.5 \leq M_w \leq 7.6$ and hypocentral depths$\leq$30 km if at least five recordings exist within rupture distances 0 to 200 km. The earthquake's missing location, depth, M_w, and fault information were verified and set up from earthquake catalogues maintained by several national and international authorities. Among the recordings in N-TSMD, only the stations in which Vs_{30} information is measured were used in this study. 46% of the stations have a Vs_{30} range between 306 m/s and 760 m/s and 38% have a Vs_{30} range between 180 m/s and 360 m/s (Figure 2.2b). Recordings that were not visually checked for any non-standard errors [71] are eliminated in the dataset. A great majority of the N-TSMD is composed of SS (approx. 52%) and N (approx. 45%) earthquakes (Figure 2.2a). In the ML-based GMMs, the geometric mean of two horizontal components of PGV (varying between 2.397 cm/s and 105.527 cm/s in the N-TSMD) and CAV (varying between 0.155 cm/s and 1394.704 cm/s in the N-TSMD) was studied and evaluated. Also, the distance parameters, including R_{JB}, R_{rup}, and R_{epi}, are limited to 200 km (Figure 2.3 and Table 2.1). The number of $M_w \geq 7$ events in the database is only 4. They constitute 0.71% of the total database (23 recordings from the 1999 Düzce (M_w7.1), 2 recordings from the 2011 Van (M_w7.1), 60 recordings from the 2020 Samos (M_w7.0), and 22 recordings from the 1999 (M_w7.6) Kocaeli earthquakes).

In the final model development and comparison with the existing models the N-TSMD was complemented by a newly released an updated strong-motion database of Turkiye (SMD-TR) [72] to further train and test the best-performing ML algorithm with the N-TSMD. The SMD-TR consists of 95,890 recordings of 9244 earthquakes with $2.0 \leq M_w \leq 7.8$ that occurred between 1976 and 2023. A total of 148 earthquakes with $3.7 \leq M_w \leq 7.8$ occurred in Turkiye from 2020 to February 2023 were used from the SMD-TR.

It corresponds to 2746 recordings with 0.3 km≤R_{JB}≤199 km and 131 m/s≤Vs_{30}≤1822 m/s (Figures 2.2-2.3). This inclusion enables us to enrich the number of recordings especially for M_w≥5.5 earthquakes with 231 recordings including the 2023 Kahramanmaraş doublet events.

Besides the N-TSMD and SMD-TR, the NEar-Source Strong-motion flat-file (NESS) [73] containing global M_w≥5.5 earthquakes was also used to further evaluate the effectiveness of the trained model for predicting the large PGV and CAV. The NESS is a flat file of ground motion parameters and associated metadata from near-source recordings of moderate-to-large global events that have been manually analyzed. The NESS file contains 1189 recordings of 81 earthquakes with M_w≥5.5 and hypocentral depth≤40 km. Recordings with missing data and R faulting types were extracted from the dataset. Seven earthquakes in Turkiye were eliminated from the dataset since these earthquakes have already been included in the N-TSMD. Eventually, 156 recordings were utilized for the additional testing process in this study. Detailed statistical information of the six estimator parameters, including M_w, Vs_{30}, Dip Angle (δ), R_{JB}, R_{rup}, and R_{epi} along with PGV and CAV values are reported in Table 2.1. Furthermore, the distributions of distances parameters versus PGV and CAV value are presented in Figure 2.3. It is worth mentioning that since the SoF is a categorical estimator parameter, its information is not shown in Table 2.1.

Table 2.1. Minimum and maximum values of parameters in the N-TSMD and NESS (* the N-TSMD expanded with the SMD-TR).

Index	**N-TSMD & SMD-TR**			**NESS**	
	Min	Max	Max*	Min	Max
PGV (cm/s)	0.001	74.5	171.3	2.4	105.5
CAV (cm/s)	0.155	1394.7	4742.7	129.7	3042.8
M_w	3.5	7.6	7.8	5.5	7.3
R_{JB} (km)	0.292	199.4	199.4	0	37.7
R_{rup} (km)	1.21	200	200	0.63	37.76
R_{epi} (km)	0.634	200	200	1.7	88.676
δ (°)	10.59	90	90	40	90
Vs_{30} (m/s)	131	1862	1862	145	1018

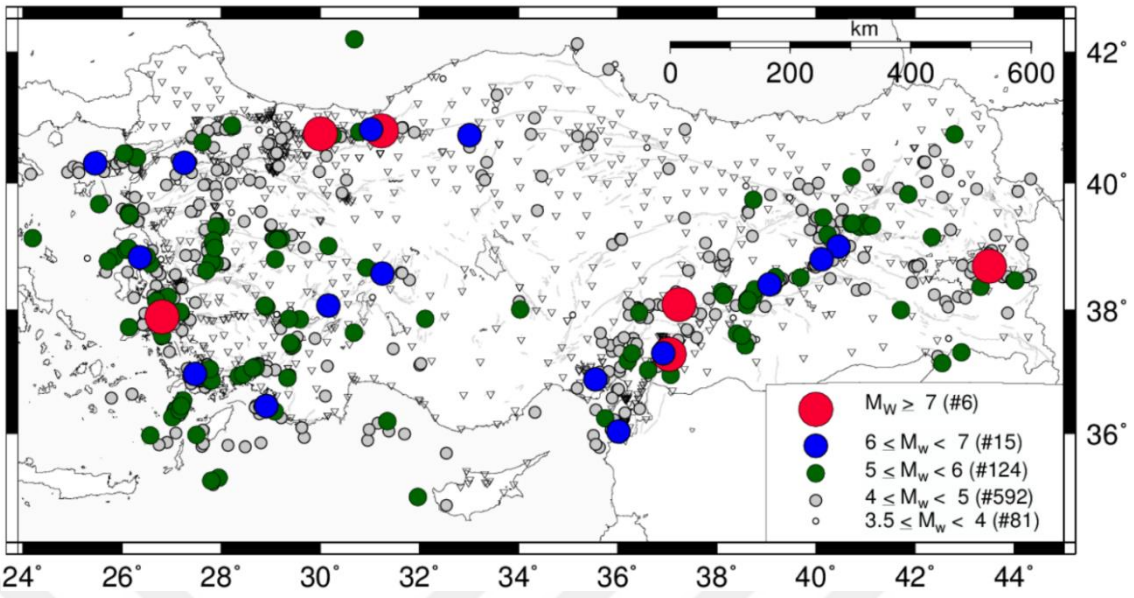


Figure 2.1. Epicentral distribution of earthquakes in the expanded N-TSMD. Grey lines are the fault traces [74].

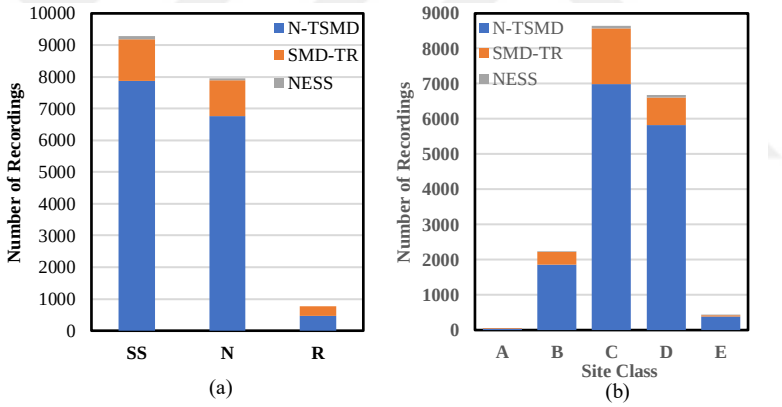


Figure 2.2. Style-of-faulting distribution of earthquakes (a) and site class of recordings according to NEHRP site classification (b) in the N-TSMD, SMD-TR, and NESS (A: Vs_{30}>1500 m/s, B: 760 m/s<Vs_{30}≤1500 m/s, C: 360 m/s<Vs_{30}≤760 m/s, D: 180 m/s<Vs_{30}≤360 m/s, and E: Vs_{30}<180 m/s).

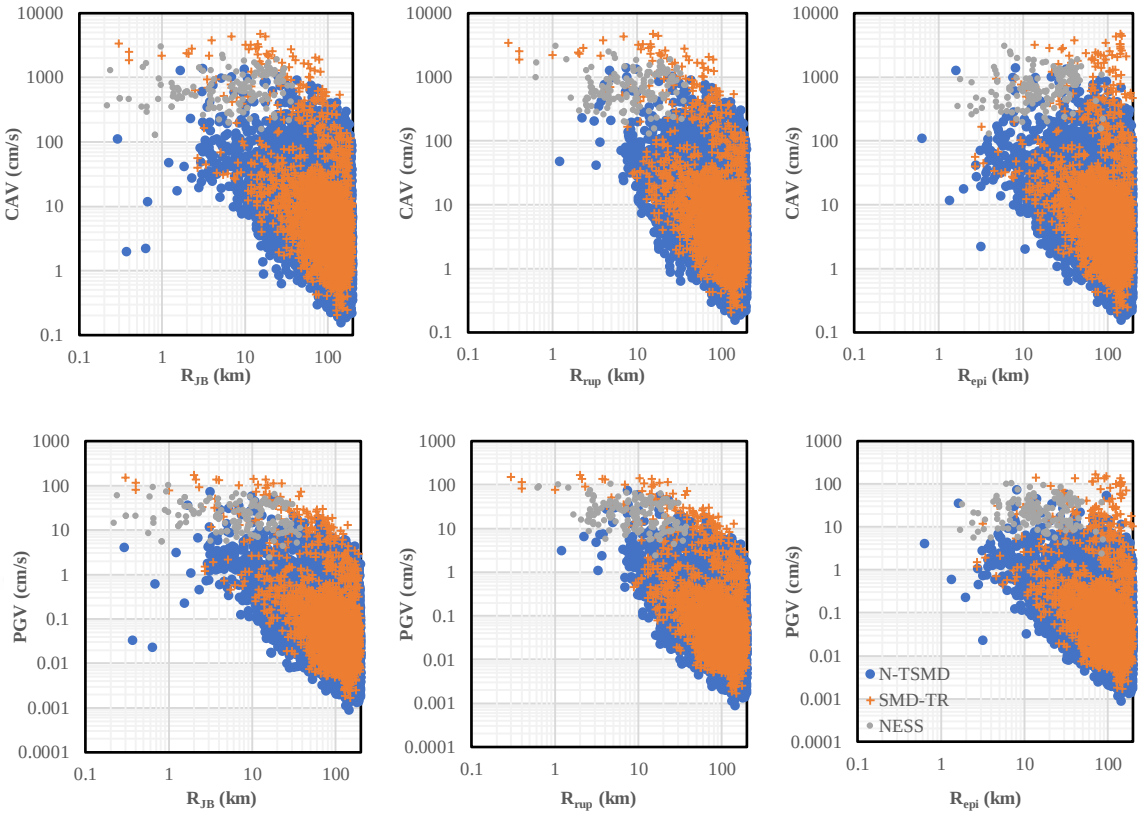


Figure 2.3. Distribution of CAVs/PGVs vs. distance parameters; R_{JB}, R_{rup}, and R_{epi}.

3. METHODOLOGY

This section provides the state-of-the-art methods and techniques applied to derive applicable non-parametric Turkiye-specific GMMs for PGV and CAV. Figure 3.1 conveys an overview of the process in general.

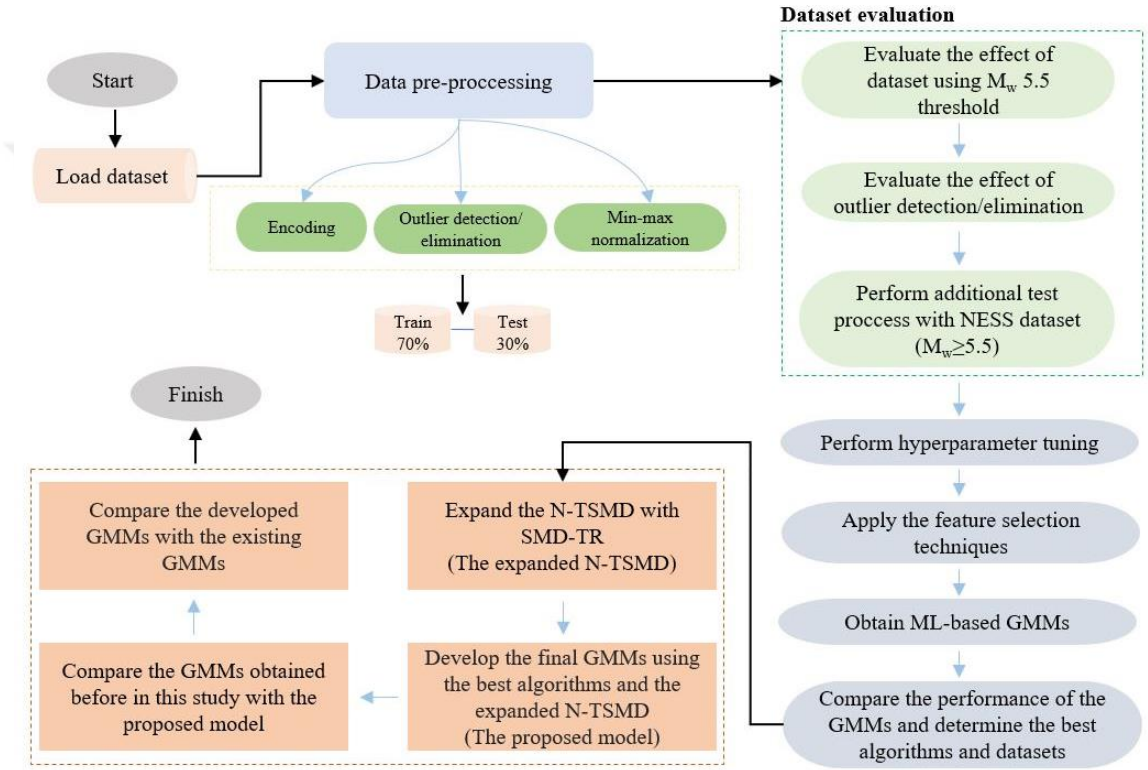


Figure 3.1. Flowchart followed in this study.

3.1. Data Preprocessing

Data preprocessing is critical for ML analysis since data quality has a large effect on the model performance. Three well-known data processing techniques, including encoding, min-max normalization, and outlier detection/elimination, are adopted in this study to obtain more accurate results.

Since various scaled variables do not contribute equally to the model fitting and model learning functions, which might lead to bias, all numeric data on the dataset were transformed into the range [0, 1] using the min-max normalization approach [75, 76].

$$x_{normalized} = \frac{x_{actual} - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

From seven estimator parameters in the N-TSMD, expanded N-TSMD, and NESS, the SoF feature is known as categorical data and is classified as SS, N, and R. Categorical data must be converted to a numerical form before applying ML methods, which was conducted on SoF by using the Frequency Encoding method. It helps the ML models to interpret and distribute weights in direct and inverse proportions [77].

Outlier detection and elimination, also known as anomaly detection, is a technique of detecting data points with extreme values relative to the rest of the distribution. Outliers have a remarkable effect on computations like mean and standard deviation, which can have a strong influence on ML. The Interquartile Range (IQR) method was applied to the N-TSMD to test whether the outlier elimination method is a plausible method for earthquake datasets. For this purpose, two different scenarios were examined: the original dataset with and without outlier elimination for building ML models. The IQR is employed by dividing a data set into quartiles which are represented as $Q25$ (the 25th percentile of the feature) and $Q75$ (the 75th percentile of the data), as seen in Figure 3.2. The IQR given below stands for the interval between the first and third quartiles. Data points both below $Q25 - 1.5IQR$ and above $Q75 + 1.5IQR$ are detected as outliers [78].

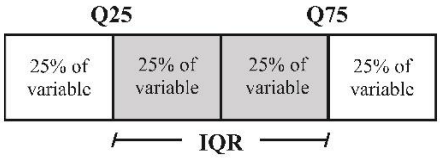


Figure 3.2. The representation of interquartile range (IQR).

The boxplot of the dataset for each estimator parameter obtained by the IQR method is shown in Figure 3.3. The red dots define the detected outliers. As expected, the method

automatically removes the recordings with M_w>5.8 and Vs_{30}>965 m/s from the N-TSMD since they consist of only 3.22% and 4.0% of the catalog, respectively. However, this operation contradicts the main goal of this study which is a better estimation of GMIMs of moderate-to-large earthquakes.

The threshold magnitude value was fixed to M_w5.5 in this study which aims to enhance recording numbers of moderate-to-large magnitude in the dataset, meanwhile, the NESS dataset also starts with M_w5.5. Thus, two new datasets were developed with 24,667 strong-motion recordings in the N-TSMD dataset to build ML models. The first dataset contains all recordings with $3.5 \leq M_w \leq 7.6$, which is called a "single dataset". The second dataset called the "separated dataset", was produced by splitting the original dataset into two parts on M_w=5.5 and contains recordings with $M_w \geq 5.5$. To understand whether a single or separated dataset would be more effective in the prediction of PGV and CAV, ML algorithms were trained and tested with both datasets. Figure 3.4 illustrates the schematic representation of datasets. The single dataset comprises 15,110 recordings with $3.5 \leq M_w \leq 7.6$, while two developed separated subsets contain small-magnitude recordings (14,102 recordings with $3.5 \leq M_w \leq 5.4$) and moderate-to-large magnitude recordings (1008 recordings with $5.5 \leq M_w \leq 7.6$). It is worth noting that an increase of approximately 86.0% is obtained in the number of recordings in the separated dataset containing moderate-to-large earthquakes when the outlier elimination approach is not applied to the data.

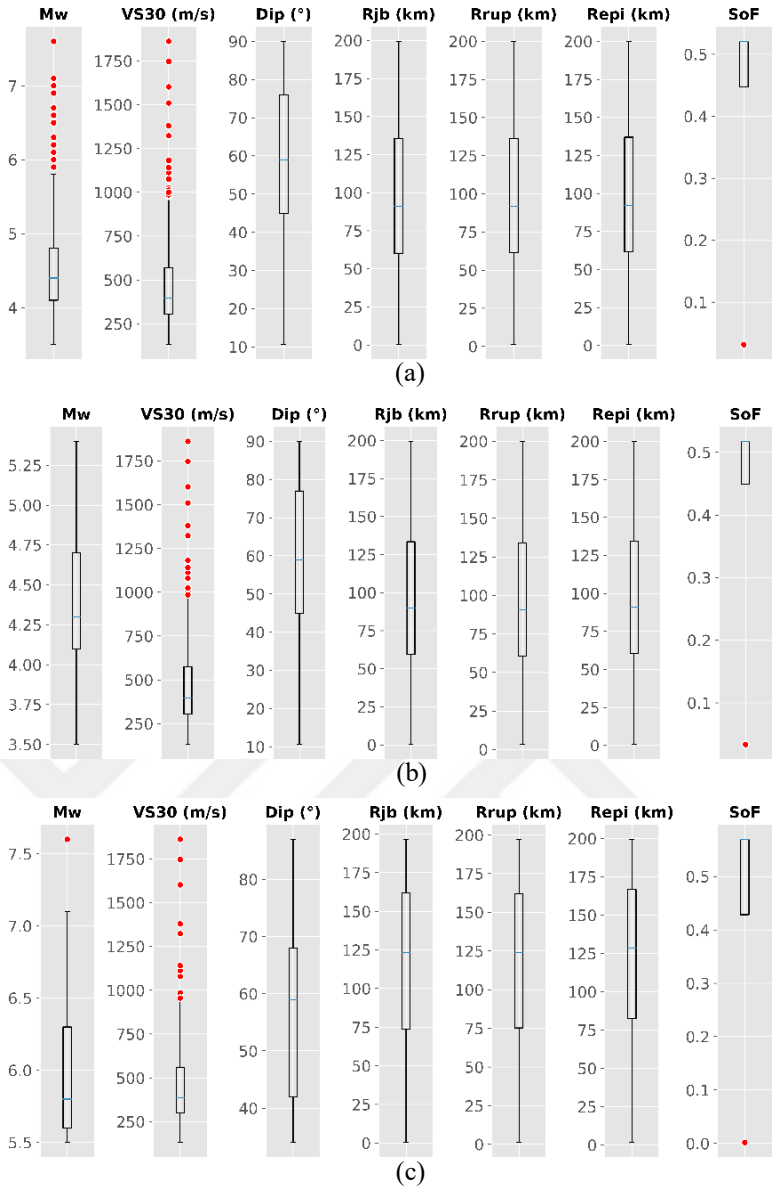


Figure 3.3. Outlier detection in the N-TSMD for (a) the single dataset (15,110 recordings), (b) the small-magnitude separated dataset (14,102 recordings), and (c) the moderate-to-large magnitude separated dataset (1008 recordings) (Red dots define outliers).

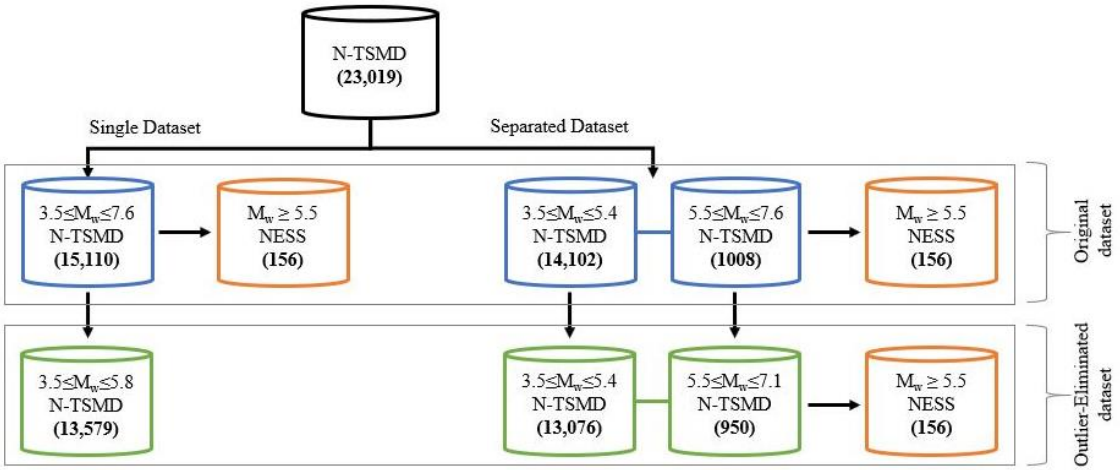


Figure 3.4. Schematic diagram of datasets used in the ML algorithms (The number of recordings indicated in parenthesis. Blue, orange, and green boxes define original, outlier-eliminated, and worldwide test datasets, respectively).

3.2. Feature Selection

Feature selection is a subset of data preparation, which is a fundamental and crucial step in the ML process. It is an approach to diminishing the number of input variables (here, the ground motion estimator parameters) when creating an ML model. It is preferable to limit the number of input variables in order to reduce modeling computational costs and, in some situations, to increase model performance. There are three types of feature selection techniques: filter, wrapper, and embedded methods. In this study, the performances of three well-known feature selection methods, including Pearson's Correlation as a filter approach, Recursive Feature Elimination (RFE) as a wrapper [79], and Least Absolute Shrinkage and Selection Operator (LASSO) as an embedded method [80] were investigated to enhance the efficiency of the ML models.

Pearson's correlation is a statistical test that assesses the linear correlation between two continuous variables.

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \tag{3.2}$$

where $\mathrm{cov}(x, y)$ indicates the covariance between x and y, and σ_x, and σ_y show the standard deviations of x and y, respectively.

It ranges from -1.0 to +1.0, while +1.0 indicates a positive linear correlation, 0 signifies no linear correlation, and -1.0 expresses a negative linear correlation. When the number of variables is large, filter-based methods are typically the faster and best option for feature selection.

RFE is a popular wrapper-based feature selection algorithm that is easy to configure and use. RFE attempts to determine a subset of features by systematically removing features one by one until the required number of features is left. It uses all of the features in the training dataset as a starting point, and the ultimate number of features to pick should be specified. Generally, wrapper approaches provide better performance than filter methods despite being computationally costly and prone to overfitting. On the other hand, embedded techniques incorporate the best features of both the filter and wrapper approaches, and lie in between the other two methods. It is less computationally intensive and less prone to overfitting than the wrapper technique. LASSO is one of the powerful embedded methods that perform both regularization and feature selection at the same time. If the feature is meaningless, LASSO penalizes its coefficient and sets it to 0. As a result, the features with a coefficient = 0 are deleted, while the others are retained.

3.3. A Brief Overview of ML Algorithms

Machine learning, a subset of artificial intelligence, has gained widespread popularity and frequent use in recent years. Its adaptation across various disciplines is attributed to its robust learning capabilities, versatility, and efficiency. In this study, ground motion models for PGV and CAV were developed using supervised ML algorithms, including ANN, BRR, GB, LR, RF, and SVM. These algorithms were chosen for their ease of use and ability to produce excellent performance results. The predictive performance of these models was systematically examined and compared to identify the most suitable Turkiye-specific GMMs for PGV and CAV. Below are brief explanations of these algorithms for comparison.

3.3.1. Linear Regression

Linear Regression is used as a statistical technique to model and analyze the relationships between a dependent variable and one or more independent variables, typically assuming a linear form but not necessarily. While it can suggest potential cause-and-effect relationships, proving causation requires careful consideration of underlying assumptions and, often, additional data and analysis. The case of one explanatory variable is called simple linear regression; for more than one variable, the process is called multiple linear regression [81]. Multiple linear regression was adopted due to the number of independent variables.

3.3.2. Random Forest

Random Forest is a powerful ML approach that combines the functionality of several Decision Tree (DT) techniques. RF contains several interesting characters and has various advantages over more established algorithms. For example, it is extremely stable to the point where fresh data does not affect it. Also, unlike decision tree regression, which employs a single tree, it is unaffected by noise. Therefore, RF has a very broad range of applications [82]. Each tree in a random forest depends on the values of a random vector that is sampled independently and with the same distribution for all the trees in the forest [83]. By introducing additional randomness into the tree-growing process, RFs assist in lessening tree correlation and can produce precise predictions [84].

3.3.3. Support Vector Machine

Support Vector Machine (SVM) is a widely used ML technique for classification tasks [85]. It works by mapping data points to a high-dimensional space and then determining the best hyperplane for classifying the data. SVR is built based on the concept of SVM which performs regression analysis on data. As a result, SVR can be used instead of SVM for working with continuous values. The core idea of SVR is locating the best-fit line, which is the hyperplane with the most points. SVR has an advantage over typical linear regression in that it can handle non-Independent and Identically Distributed (I.I.D) datasets and model non-linear relationships.

3.3.4. Bayesian Ridge Regression

One of the most effective Bayesian Regression approaches is Bayesian Ridge Regression (BRR), which uses probabilistic distribution to estimate the value of the dependent variable. BRR does not require any additional previous information about the dataset and makes predictions using all possible regression weights, weighted by their posterior probability. BRR uses Ridge Regularization which allows us to impose certain priors on dependent variables and solve the over-fitting problem. Another important reason to choose Bayesian approaches is that they allow integration uncertainty in parameter estimations, which is very valuable for predicting [86].

3.3.5. Artificial Neural Network

Artificial Neural Network is an ML modeling technology that has surpassed conventional regression and statistical models in terms of performance. ANN is widely employed due to its remarkable performance in self-learning, adaptivity, and nonlinearity. ANN can be used to solve a wide variety of problems, including those with complex solutions and others that do not have algorithmic solutions [47]. The multilayer perceptron (MLP) algorithm is the simplest type of ANN, which is composed of input, hidden, and output layers [61]. The MLP is made up of a network of straightforwardly coupled nodes or neurons [87]. The number of hidden layers and nodes in a hidden layer, the activation functions, and the input parameters that are pertinent to the outputs are all decisions that must be taken during the ANN model's development [66].

3.3.6. Gradient Boosting

Gradient Boosting is an ensemble of weak models such as DT. In particular, numerous weak trees are connected sequentially whereas they do not influence each other. Each tree seeks to minimize the mistakes of the preceding tree [59]. In other words, the gradient-boosting tree technique may improve prediction accuracy by strategically integrating additional trees and fixing issues in its preceding foundation models [88]. GB cultivates trees by measuring the differences in estimated and observed IMs, and the residual

is the output variable of each tree in the GB instead of the estimated IMs. It also provides considerable potential to build data-driven prediction models for many different types of engineering applications [89].

3.4. Performance Evaluation Metrics

The reliability of the ML prediction models can be assessed using a range of conventional metrics. Five evaluation metrics, which are mean square error (MSE), mean absolute error (MAE), correlation coefficient (R), root mean square error (RMSE), and mean absolute percentage error (MAPE), were employed to find the best-performing algorithm. A brief formulation of these performance evaluations can be given as

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i} - \mathrm{y_i})^2 \tag{3.3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum(\mathrm{x_i} - \mathrm{y_i})^2} \tag{3.4}$$

$$\mathrm{MAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}(|\mathrm{x_i} - \mathrm{y_i}|)}{\mathrm{N}} \tag{3.5}$$

$$\mathrm{MAPE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\left|\frac{\mathrm{x_i} - \mathrm{y_i}}{\mathrm{x_i}}\right|\right) \tag{3.6}$$

$$\mathrm{R} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sqrt{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2 \sum(\mathrm{y_i} - \bar{\mathrm{y}})^2}} \tag{3.7}$$

where $\mathrm{x_i}$, $\mathrm{y_i}$, $\bar{\mathrm{x}}_\mathrm{i}$, $\bar{\mathrm{y}}_\mathrm{i}$, and N are measured output, predicted output, the mean value of measured output, the mean value of the predicted output, and the total number of data, respectively.

3.5. Hyperparameter Tuning

An ML model is defined as a mathematical model with multiple parameters that must be optimized to increase the approach performance. When an ML algorithm is applied to a dataset, hyperparameters are the parameters that are purposefully designed to influence the learning process. These parameters describe crucial aspects of the model, such as its complexity or how quickly and precisely it should pick up new information.

Before the training process is activated, they are typically rectified, and their values cannot be readily obtained from the standard training procedure. Hyperparameter tuning is finding the best combination of the values for hyperparameters that gives the best performance. One of the most effective methods for hyperparameter tuning is Grid-search, which uses a grid of parameter values to search and obtain the optimum set of hyperparameters. The predefined ranges of hyperparameters for five different ML techniques are listed in Table 3.1. Also, the best sets of hyperparameters obtained using Grid-search are shown in bold format in Table 3.1.

Table 3.1. Hyperparameter tuning parameters of ML techniques for (a) CAV and (b) PGV output parameters.

(a) CAV	
Techniques	**Best parameters**
SVM	[{'C': [0.1, 1, 10], 'kernel': ['linear'], 'nu': [0.1, 0.5, 0.9]}, {'C': [0.1, 1, **10**], 'kernel': ['**rbf**'], 'gamma': [0.1, 0.3, 0.4, 0.6, **0.9**], 'nu': [0.1, 0.5, **0.9**]}]
RF	{'n_estimators': [50, 100, **500**], 'bootstrap': [**True**, False], 'max_depth': ['None', 80, **200**], 'max_features': [**1**, 3], 'min_samples_split': [1, **2**, 3, 8]}
ANN	{'alpha': [0.0001, **0.05**], 'solver': ['sgd', **'adam'**], 'hidden_layer_sizes': [(50,50,50), **(50,100,50)**, (100,)], 'learning_rate': ['constant',**'adaptive'**], 'activation': [**'relu'**,'tanh','logistic']}
BRR	{'alpha_1': [1e-4, 1e-6, **1e-12**], 'alpha_2': [**1e-4**, 1e-6, 1e-12], 'alpha_init': ['None', 1, 1.5, **1.9**], 'lambda_1': [**1e-4**, 1e-6, 1e-12], 'lambda_2': [1e-4, **1e-6**, 1e-12], 'lambda_init': [**'None'**,1e-1, 1e-5, 1e-9], 'n_iter': [**50**, 100, 300], 'tol': [1e-3, 1e-4, **1e-6**, 100]}
GB	{'n_estimators': [1, 5, 50, 100, **200**], 'max_depth': [1, 3, **5**], 'max_leaf_nodes': ['None', 5, **20**, 100], 'learning_rate': [0.01, **0.1**, 1.0]}
(b) PGV	
Techniques	**Best parameters**
SVM	[{'C': [0.1, 1, 10], 'kernel': ['linear'], 'nu': [0.1, 0.5, 0.9]}, {'C': [0.1, 1, **10**], 'kernel': **['rbf']**, 'gamma': [0.1, 0.3, 0.4, **0.6**, 0.9], 'nu': [0.1, 0.5, **0.9**]}]
RF	{'n_estimators': [50, **100**, 500], 'bootstrap': [**True**, False], 'max_depth': ['**None**', 80, 200], 'max_features': [**1**, 3], 'min_samples_split': [1, **2**, 3, 8]}
ANN	{'alpha': [**0.0001**, 0.05], 'solver': ['sgd', '**adam**'], 'hidden_layer_sizes': [(50,50,50), (50,100,50), **(100,)**], 'learning_rate': ['**constant**','adaptive'], 'activation': [**'relu'**,'tanh','logistic']}
BRR	{'alpha_1': [1e-4, 1e-6, **1e-12**], 'alpha_2': [**1e-4**, 1e-6, 1e-12], 'alpha_init': ['None', 1, 1.5, **1.9**], 'lambda_1': [**1e-4**, 1e-6, 1e-12], 'lambda_2': [1e-4, 1e-6, **1e-12**], 'lambda_init': ['None',**1e-1**, 1e-5, 1e-9], 'n_iter': [**50**, 100, 300], 'tol': [1e-3, 1e-4, **1e-6**, 100]}
GB	{'n_estimators': [1, 5, 50, 100, **200**], 'max_depth': [1, 3, **5**], 'max_leaf_nodes': ['None', 5, **20**, 100], 'learning_rate': [0.01, **0.1**, 1.0]}

4. ANALYSIS AND RESULT

This section reveals the findings obtained as a result of the effort spent to create the best ML-based Turkiye-specific ground motion models with the N-TSMD for PGV and CAV. In the first evaluation, the power of datasets and outlier elimination in process modeling are investigated (Section 4.1). The estimator parameter sets determined with feature selection methods are presented and evaluated in the second evaluation (Section 4.2). Also, the performances of TR-specific GMMs are obtained with six ML algorithms and different variations of the N-TSMD in section 4.2. In the final stages (Sections 4.3 and 4.4), state-of-the-art Turkiye-specific GMMs for PGV and CAV were developed with the expanded N-TSMD using the best algorithm determined in this study and compared the existing GMMs.

4.1. First Evaluation: Effect of Dataset and Outlier Elimination on Prediction

In this section, two questions are tried to answer: Should earthquake datasets be cleaned for outlier events in ground motion analyses? And whether or not to partition the original dataset based on the M_w value for better prediction performance.

For this purpose, seven estimator parameters, including M_w, δ, SoF, R_{JB}, R_{rup}, R_{epi}, and Vs_{30}, were considered as inputs of the dataset. These parameters represent source, faulting, path, and local site response. Variations of those and if available many more region-specific parameters are commonly utilized in all GMM studies. The six state-of-the-art ML methods were performed independently several times on the N-TSMD dataset since the different subsets of recordings can be obtained at each independent run. Also, Holdout Cross-Validation was used to measure errors. Holdout Cross-Validation divides the data into two groups. Generally, the split of test data is smaller than training data. The procedure is repeated with each iteration utilizing a new group as the test set. In this study, 70% of the data was used for training and 30% for the test process. The final estimation (test error) is based on the average error test. The generated results were assessed using five performance

metrics, which are the most popular evaluation criteria in ML regression studies. A single dataset ($3.5 \leq M_w \leq 7.6$) and the separated dataset consisting of small magnitude ($3.5 \leq M_w \leq 5.4$) and moderate-to-large magnitude ($5.5 \leq M_w \leq 7.6$) earthquakes were derived from the N-TSMD dataset and utilized for model training. The outlier elimination was performed on these datasets for both output variables (PGV and CAV). Subsections 4.1.1 and 4.1.2 display the results of the PGV and CAV prediction in terms of five performance metrics in details.

4.1.1. PGV Result of the First Evaluation

Figure 4.1 displays the average results of the PGV prediction in terms of five performance metrics, and detailed results are shown in Table 4.1 and Table 4.2. In all ML methods, the lowest MSE, RMSE, and MAE values are obtained from the single dataset (orange-straight line in Figure 4.1), while the lowest MAPE in all ML models (except for GB and RF) is observed on the small-magnitude dataset. The single dataset achieves remarkable success with the lowest MAE, MSE, and RMSE metrics. Also, the highest R values in four (GB, SVM, LR, RF) out of six ML methods are obtained with the single dataset (orange-straight line in Figure 4.1).

Figure 4.1 also shows the effect of outlier elimination on PGV prediction. Metric results for the outlier-eliminated dataset (dashed lines in Figure 4.1) also follow a similar trend to the original datasets. In original datasets, almost all ML methods trained by outlier-eliminated datasets result in a lower error and higher R metrics. The lowest MAE, MSE, and RMSE are observed with ML models trained with the outlier-eliminated single (orange-dashed line in Figure 4.1) datasets. The lowest MAPE is observed with the small-magnitude datasets (blue lines in Figure 4.1) in all ML models except RF. Looking at the R metric, results are generally close to each other, however, in two (GB and RF) of the six ML models, the algorithms trained with the outlier-eliminated single dataset (orange-dashed line in Figure 4.1) show the best testing performance. The difference in error metrics obtained with the original and outlier-eliminated dataset is insignificant for the separated ($3.5 \leq M_w \leq 5.4$ and $5.5 \leq M_w \leq 7.6$) datasets. Among all combinations, the highest correlation coefficients ($R=0.88$ and $R=0.89$) in the prediction of PGV are reached with the RF and GB models, respectively, using the outlier-eliminated single dataset.

Table 4.1. Performance metrics (in normalized form) of PGV prediction with six ML algorithms trained with the N-TSMD original datasets. Bold values represent the best results.

ML Methods	Magnitude Range	Number of Recordings	MAE ± σ (x10^{-2})	MAPE ± σ	MSE ± σ (x10^{-3})	RMSE ± σ (x10^{-2})	R ± σ
ANN	**3.5≤M_w≤7.6**	15110	0.4507±0.065	10.710±1.562	0.1723±0.107	1.2710±0.367	0.7114±0.031
	3.5≤M_w≤5.4	14102	0.8292±0.012	2.4366±0.338	0.4283±0.094	2.0593±0.229	0.7368±0.026
	5.5≤M_w≤7.6	1008	2.0317±0.138	4.2946±1.451	1.9561±1.000	4.3254±1.032	0.6173±0.074
GB	**3.5≤M_w≤7.6**	15110	0.1754±0.010	1.0679±0.106	**0.1150±0.021**	1.0689±0.102	**0.8876±0.024**
	3.5≤M_w≤5.4	14102	0.6419±0.028	0.9186±0.018	0.4619±0.139	2.1280±0.337	0.7789±0.018
	5.5≤M_w≤7.6	1008	1.2842±0.250	0.8043±0.108	1.4254±0.680	3.6384±1.126	0.8231±0.028
SVM	**3.5≤M_w≤7.6**	15110	**0.1314±0.070**	1.0247±0.199	0.420±0.6840	**0.9873±0.287**	0.7452±0.040
	3.5≤M_w≤5.4	14102	0.6318±0.007	**0.7730±0.030**	0.5294±0.092	1.8215±1.002	0.7416±0.014
	5.5≤M_w≤7.6	1008	1.1898±0.151	1.0931±0.555	1.3895±0.774	3.6192±0.998	0.7016±0.038
LR	**3.5≤M_w≤7.6**	15110	0.4356±0.007	13.243±0.868	0.1673±0.043	1.0212±0.586	0.5543±0.018
	3.5≤M_w≤5.4	14102	1.1324±0.016	4.2702±0.390	0.9147±0.090	3.0215±0.150	0.4542±0.007
	5.5≤M_w≤7.6	1008	2.3902±0.233	4.6886±1.622	3.3289±1.027	5.7080±0.940	0.5292±0.027
RF	**3.5≤M_w≤7.6**	15110	0.1735±0.012	0.9124±0.024	0.1696±0.045	1.2933±0.171	0.8557±0.009
	3.5≤M_w≤5.4	14102	0.6258±0.036	0.9343±0.015	0.4247±0.119	2.0453±0.283	0.7819±0.013
	5.5≤M_w≤7.6	1008	1.3185±0.166	1.0253±0.367	1.3989±0.331	3.7146±0.488	0.8240±0.030
BRR	**3.5≤M_w≤7.6**	15110	0.4475±0.008	13.373±0.970	0.2437±0.117	1.5204±0.396	0.4854±0.038
	3.5≤M_w≤5.4	14102	1.1121±0.021	4.1569±0.226	0.6777±0.232	2.1398±1.232	0.5105±0.018
	5.5≤M_w≤7.6	1008	2.4672±0.197	6.4179±1.931	2.2538±1.316	4.5849±1.376	0.5948±0.040

Table 4.2. Performance metrics (in normalized form) of PGV prediction with six ML algorithms trained with the N-TSMD outlier-eliminated datasets. Bold values represent the best results.

ML Methods	Magnitude Range	Number of Recordings	MAE ± σ (x10^{-2})	MAPE ± σ	MSE ± σ (x10^{-3})	RMSE ± σ (x10^{-2})	R ± σ
ANN	**3.5≤M_w≤5.8**	13579	0.3800±0.041	5.7261±1.272	0.1780±0.152	1.2259±0.588	0.5909±0.066
	3.5≤M_w≤5.4	13076	0.9162±0.059	2.4706±0.442	0.6435±0.165	2.5200±0.326	0.7028±0.008
	5.5≤M_w≤7.1	950	2.0620±0.292	6.1622±1.309	3.7739±2.429	5.8514±2.091	0.5757±0.020
GB	**3.5≤M_w≤5.8**	13579	0.1722±0.005	0.9273±0.032	0.0401±0.015	0.6258±0.110	**0.8973±0.018**
	3.5≤M_w≤5.4	13076	0.6402±0.020	0.8887±0.021	0.4511±0.061	2.1199±0.147	0.7862±0.023
	5.5≤M_w≤7.1	950	1.1497±0.227	0.9166±0.137	1.3069±1.151	3.3906±1.402	0.8392±0.030
SVM	**3.5≤M_w≤5.8**	13579	0.1716±0.011	**0.7612±0.036**	0.0601±0.040	0.7479±0.230	0.7289±0.028
	3.5≤M_w≤5.4	13076	0.6557±0.060	0.7667±0.021	0.6292±0.228	2.4720±0.476	0.7411±0.025
	5.5≤M_w≤7.1	950	0.9715±0.101	2.3071±0.969	0.7296±0.292	2.6576±0.540	0.7385±0.044
LR	**3.5≤M_w≤5.8**	13579	0.3495±0.007	6.0948±0.311	0.0560±0.015	0.7431±0.099	0.4706±0.009
	3.5≤M_w≤5.4	13076	1.1550±0.015	4.4516±0.371	0.8724±0.086	2.9507±0.146	0.4674±0.017
	5.5≤M_w≤7.1	950	1.9407±0.083	4.8064±3.217	0.9734±0.143	3.1133±0.229	0.6062±0.018
RF	**3.5≤M_w≤5.8**	13579	**0.1687±0.002**	0.9131±0.035	**0.0364±0.009**	**0.6000±0.076**	0.8832±0.016
	3.5≤M_w≤5.4	13076	0.6554±0.023	0.9246±0.029	0.5057±0.099	2.2402±0.220	0.7672±0.012
	5.5≤M_w≤7.1	950	1.1838±0.352	0.8151±0.134	1.8061±1.338	3.9242±1.824	0.8475±0.051
BRR	**3.5≤M_w≤5.8**	13579	0.3476±0.002	6.2687±0.369	0.0490±0.012	0.6960±0.089	0.5086±0.019
	3.5≤M_w≤5.4	13076	1.1785±0.033	3.7166±0.125	1.0064±0.269	3.1518±0.403	0.4941±0.010
	5.5≤M_w≤7.1	950	2.1231±0.128	4.2587±2.046	3.5317±2.638	5.5541±2.363	0.5272±0.026

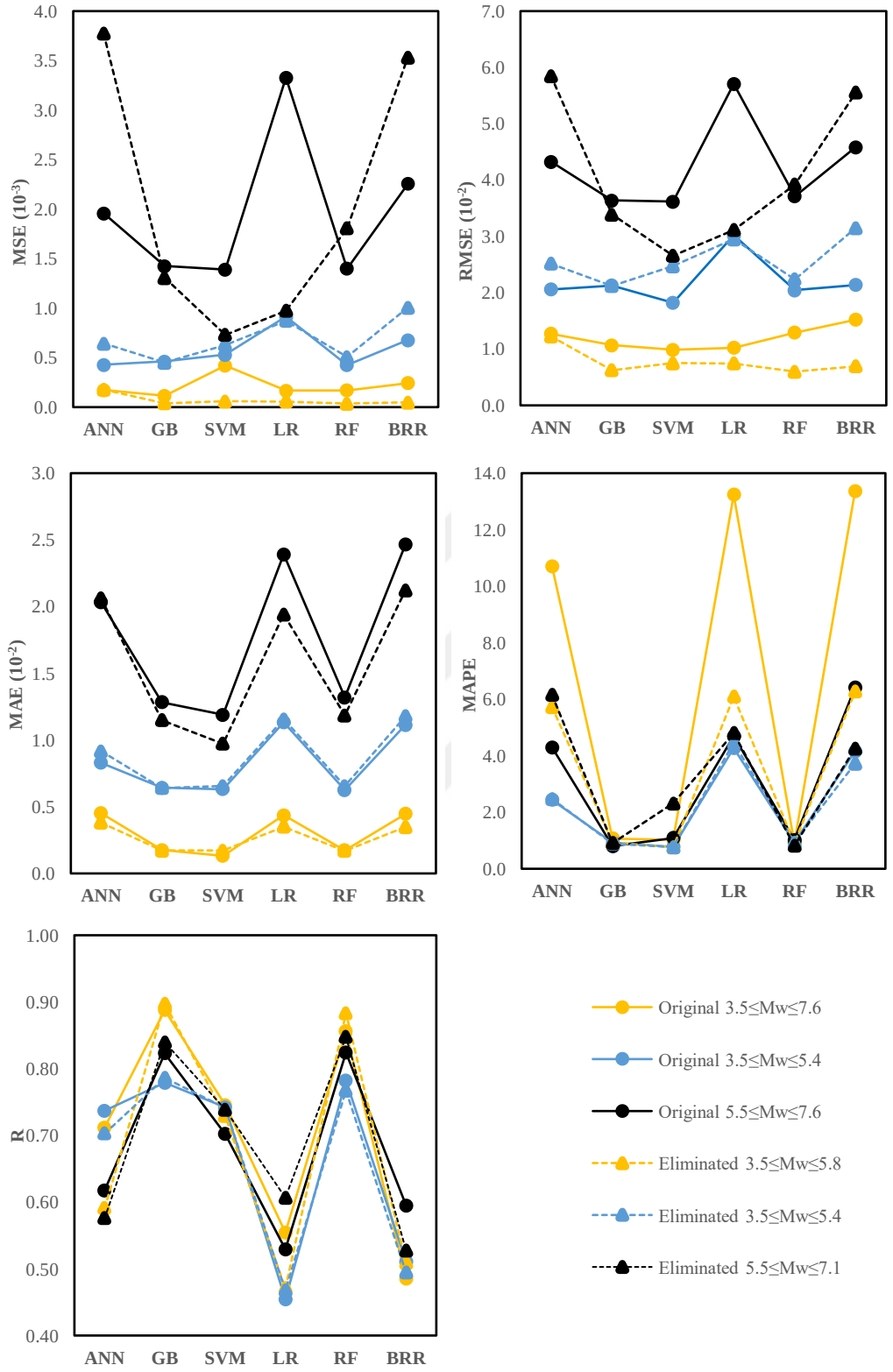


Figure 4.1. Performance metrics (in normalized form) of six ML algorithms trained with seven estimator parameters using the N-TSMD original and outlier-eliminated datasets for the prediction of PGV.

4.1.2. CAV Result of the First Evaluation

Figure 4.2 displays the average results of the CAV prediction in terms of five performance metrics, and detailed results are shown in Table 4.3 and Table 4.4. In all ML methods, the lowest MSE, RMSE, and MAE values are obtained from the single dataset (orange-straight line in Figure 4.2). While the lowest MAPE in all ML models is observed on the small-magnitude dataset. The single dataset achieves remarkable success with the lowest MAE, MSE, and RMSE metrics. It should be noted that the highest R values in four (GB, SVM, LR, BRR) out of six ML methods are obtained with the moderate-to-large magnitude dataset (black-straight line in Figure 4.2) which has the smallest number of recordings.

Figure 4.2 also shows the effect of outlier elimination on CAV prediction. Metrics results for the outlier-eliminated dataset (dashed lines in Figure 4.2) also follow a similar trend to the original datasets. In almost all datasets, ML methods trained by outlier-eliminated datasets result in a lower error and higher R metrics. The lowest MAE, MSE, and RMSE are observed with ML models trained with the outlier-eliminated single (orange-dashed line in Figure 4.2) datasets. The lowest MAPE is observed with the eliminated small-magnitude dataset (blue-dashed line in Figure 4.2) in all ML models except RF. Looking at the R metric, the results are generally close to each other, however, in three (GB, LR, BRR) of the six ML models, the algorithms trained with outlier-eliminated moderate-to-large magnitude dataset (black-dashed line in Figure 4.2) show the best testing performance. The difference in error metrics obtained with the original and outlier-eliminated dataset is insignificant for the small magnitude ($3.5 \leq M_w \leq 5.4$) dataset, except for the RF method. The RF algorithm trained with the original dataset has lower error terms; therefore, it can predict CAV more accurately than the RF algorithm trained by the outlier-eliminated dataset. The difference in error values obtained with the original and outlier-eliminated datasets is rather noteworthy for the moderate-to-large magnitude ($5.5 \leq M_w \leq 7.6$) dataset. Training of ML methods with the moderate-to-large magnitude dataset has the least success in terms of errors. However, among all combinations, the highest correlation coefficient ($R=0.86$) in the prediction of CAV is reached with the GB model using the outlier-eliminated moderate-to-large magnitude dataset.

Table 4.3. Performance metrics (in normalized form) of CAV prediction with six ML algorithms trained with the N-TSMD original datasets. Bold values represent the best results.

ML Methods	Magnitude Range	Number of recordings	MAE ± σ (x10^{-2})	MAPE ± σ	MSE ± σ (x10^{-3})	RMSE ± σ (x10^{-2})	R ± σ
ANN	**3.5≤M_w≤7.6**	15110	0.6705±0.089	2.8206±0.609	0.4621±0.062	2.1456±0.148	0.8008±0.024
	3.5≤M_w≤5.4	14102	1.1839±0.029	1.0345±0.081	0.6600±0.149	2.5550±0.300	0.7491±0.006
	5.5≤M_w≤7.6	1008	3.2148±0.341	2.4499±2.101	4.5356±1.228	6.6824±0.935	0.7478±0.027
GB	**3.5≤M_w≤7.6**	15110	**0.4223±0.031**	0.7945±0.065	**0.2716±0.098**	**1.6270±0.295**	0.8335±0.011
	3.5≤M_w≤5.4	14102	0.9531±0.023	**0.6788±0.012**	0.5095±0.054	2.2365±0.135	0.8176±0.011
	5.5≤M_w≤7.6	1008	2.6448±0.377	1.2503±0.633	3.5634±1.242	5.8961±1.042	0.8471±0.045
SVM	**3.5≤M_w≤7.6**	15110	0.4652±0.052	0.7392±0.057	0.5508±0.283	2.2829±0.617	0.7919±0.013
	3.5≤M_w≤5.4	14102	1.0630±0.044	0.7382±0.126	0.7045±0.123	2.6453±0.247	0.7572±0.004
	5.5≤M_w≤7.6	1008	2.6397±0.310	1.3802±1.221	4.1387±1.195	6.3673±1.027	0.8019±0.024
LR	**3.5≤M_w≤7.6**	15110	0.9563±0.012	5.7299±0.718	0.4170±0.040	2.0401±0.098	0.6651±0.030
	3.5≤M_w≤5.4	14102	1.5453±0.012	2.0353±0.090	0.9446±0.136	3.0673±0.218	0.5996±0.005
	5.5≤M_w≤7.6	1008	4.1823±0.131	3.2090±1.902	5.7563±1.281	7.5516±0.819	0.6788±0.018
RF	**3.5≤M_w≤7.6**	15110	0.4252±0.014	0.8232±0.049	0.3063±0.098	1.7333±0.273	**0.8471±0.033**
	3.5≤M_w≤5.4	14102	1.0024±0.025	0.7342±0.013	0.4817±0.055	2.1920±0.125	0.8331±0.014
	5.5≤M_w≤7.6	1008	2.9978±0.338	3.7916±2.146	4.5173±1.190	6.6756±0.872	0.7881±0.016
BRR	**3.5≤M_w≤7.6**	15110	0.9590±0.016	5.3207±0.391	0.6557±0.169	2.5440±0.326	0.5962±0.021
	3.5≤M_w≤5.4	14102	1.5729±0.046	2.0681±0.095	1.0747±0.230	3.2618±0.366	0.5913±0.003
	5.5≤M_w≤7.6	1008	4.2441±0.274	4.9620±1.486	6.3796±2.788	7.8399±1.707	0.6606±0.032

Table 4.4. Performance metrics (in normalized form) of CAV prediction with six ML algorithms trained with the N-TSMD outlier-eliminated datasets. Bold values represent the best results.

ML Methods	Magnitude Range	Number of recordings	MAE ± σ (x10^{-2})	MAPE ± σ	MSE ± σ (x10^{-3})	RMSE ± σ (x10^{-2})	R ± σ
ANN	**3.5≤M_w≤5.8**	13579	0.5883±0.044	1.7377±0.292	0.1140±0.013	1.0700±0.063	0.8172±0.009
	3.5≤M_w≤5.4	13076	1.1921±0.027	1.1303±0.168	0.5231±0.076	2.2826±0.160	0.8008±0.008
	5.5≤M_w≤7.1	950	2.9900±0.137	2.9082±1.100	3.9749±0.691	6.2849±0.557	0.7969±0.004
GB	**3.5≤M_w≤5.8**	13579	**0.3921±0.008**	0.6757±0.016	**0.0879±0.008**	0.9360±0.044	0.8501±0.015
	3.5≤M_w≤5.4	13076	0.9538±0.021	0.6531±0.022	0.5003±0.033	2.2358±0.075	0.8245±0.013
	5.5≤M_w≤7.1	950	2.2395±0.379	0.8505±0.214	2.3653±0.934	4.7926±1.076	**0.8634±0.006**
SVM	**3.5≤M_w≤5.8**	13579	0.4050±0.020	**0.6333±0.014**	0.0901±0.017	0.9450±0.091	0.8297±0.011
	3.5≤M_w≤5.4	13076	1.0379±0.030	0.6551±0.006	0.5866±0.088	2.4130±0.189	0.8099±0.020
	5.5≤M_w≤7.1	950	2.4567±0.066	1.3129±1.066	2.9016±0.529	5.3701±0.472	0.8237±0.023
LR	**3.5≤M_w≤5.8**	13579	0.6922±0.010	2.6465±0.161	0.1550±0.022	1.2400±0.089	0.6355±0.007
	3.5≤M_w≤5.4	13076	1.5830±0.024	2.0272±0.060	0.8781±0.146	2.9551±0.244	0.6192±0.009
	5.5≤M_w≤7.1	950	4.0064±0.351	3.1530±0.907	5.1233±2.170	7.0191±1.567	0.6746±0.016
RF	**3.5≤M_w≤5.8**	13579	0.4027±0.024	0.7422±0.015	0.0884±0.021	**0.9350±0.108**	0.8522±0.011
	3.5≤M_w≤5.4	13076	1.2713±0.594	1.0442±0.665	0.9125±0.880	2.8185±1.214	0.8331±0.006
	5.5≤M_w≤7.1	950	2.6556±0.185	1.2768±0.852	3.4986±0.855	5.8796±0.721	0.8205±0.035
BRR	**3.5≤M_w≤5.8**	13579	0.6969±0.012	2.5891±0.064	0.1600±0.021	1.2600±0.085	0.6277±0.009
	3.5≤M_w≤5.4	13076	1.6001±0.012	2.0013±0.120	0.8882±0.099	2.9766±0.163	0.6076±0.010
	5.5≤M_w≤7.1	950	3.8471±0.165	2.5336±0.521	4.7453±0.911	6.8619±0.677	0.6742±0.017

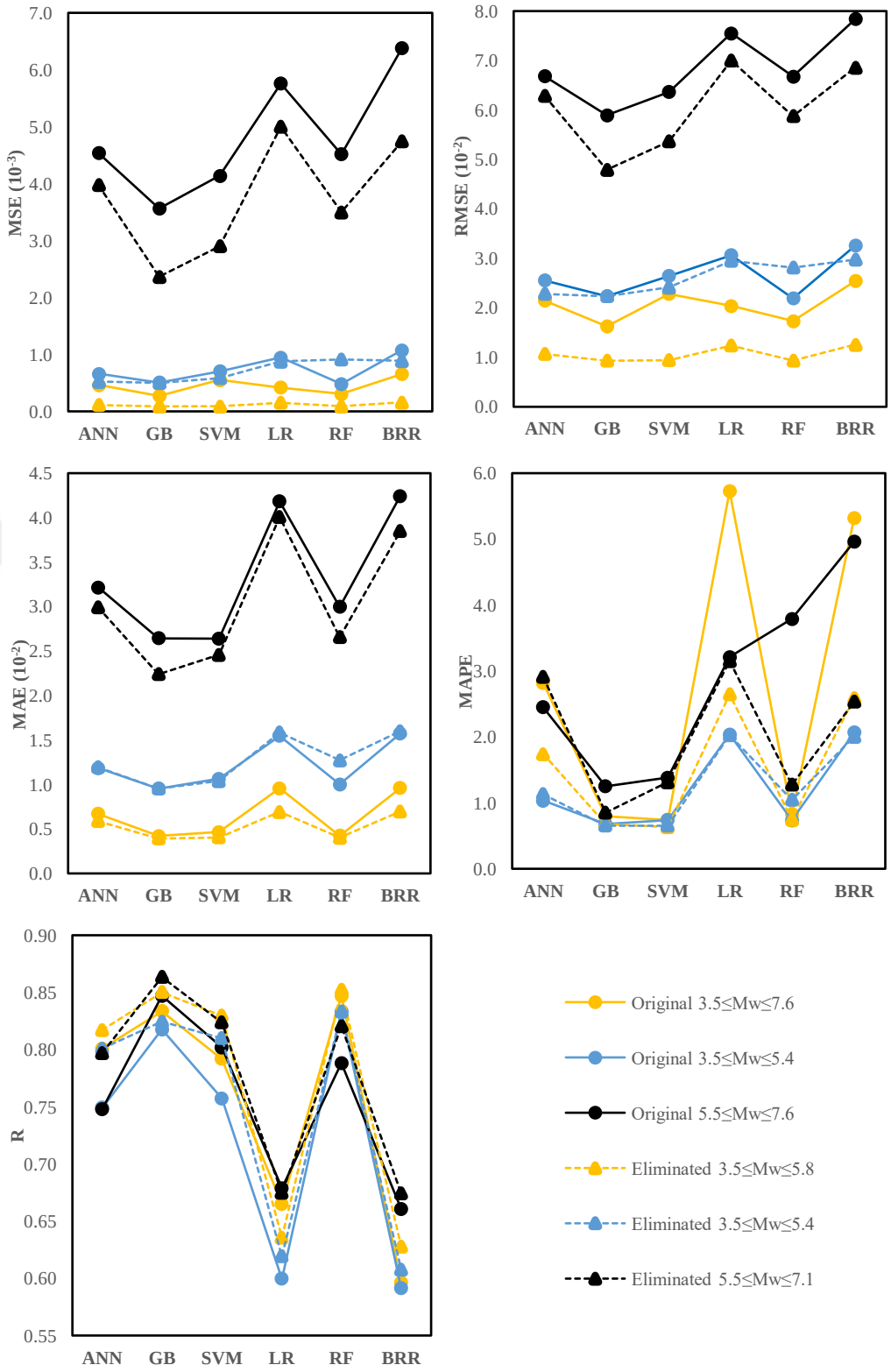


Figure 4.2. Performance metrics (in normalized form) of six ML algorithms trained with seven estimator parameters using the N-TSMD original and outlier-eliminated datasets for the prediction of CAV.

Analyses up to here indicate that the ML algorithms trained by the single dataset, which consists of the highest number of recordings, give better results based on performance metrics in PGV prediction as well as CAV prediction compared to the algorithms trained with the separated datasets. As expected, the performance of ML methods decreases due to the small size of training samples. Considering outlier elimination when the dataset is split, it observes that algorithms trained by the outlier-eliminated datasets give better prediction results than algorithms trained by the original datasets. Therefore, cleaning the data from outliers can affect the quality of ground motion analyses.

4.1.3. Verification Effort with the NESS Dataset

From Tables 4.1-4.4 and Figures 4.1-4.2, it can be seen that ML algorithms trained with outlier-eliminated datasets predict PGV and CAV better than the ML algorithms trained with original datasets. However, a great majority of moderate-to-large earthquakes (M_w>5.8) were eliminated as outliers in the associated dataset because they are not in the specified IQR. Due to the nature of the earthquake, large earthquakes occur less frequently than small and moderate earthquakes encountered frequently. However, the importance of large earthquakes is inversely proportional to the frequency of their recurrence. Also, an ML model must be generalizable for predicting unseen data, especially large earthquakes. Therefore, ML models trained with both the original and outlier-eliminated datasets were tested on additional unseen data, i.e. the NESS dataset. Unseen data is any form of data that a model has never learned before. By adding this additional step, the performance of the trained ML models on PGV and CAV prediction of moderate-to-large earthquakes was verified. In other words, the trained model on the NESS dataset is validated to ensure that it can still reliably predict large earthquakes. Figures 4.3-4.4 supported by Tables 4.5-4.6 show the test results for the NESS dataset in terms of all evaluation metrics.

In the prediction of CAV, GB and RF among all ML algorithms depict lower error trends regardless of dataset type. MSE, RMSE, and MAE have their lower values when the GB and RF algorithms are tested with the original ($5.5 \leq M_w \leq 7.6$) and separated outlier-eliminated ($5.5 \leq M_w \leq 7.1$) datasets, except that GB has the lowest MAE value with a single

original dataset ($3.5 \leq M_w \leq 7.6$). Moreover, the lowest MAPE values for the same two ML algorithm are obtained with the single original dataset ($3.5 \leq M_w \leq 7.6$). All ML algorithms trained with the original separated dataset ($5.5 \leq M_w \leq 7.6$) show the highest R metrics. The overall picture demonstrates that ML models trained with a single dataset do not present remarkable success in the prediction of CAV. Moreover, no obvious difference is observed between the algorithms trained with the original and separated datasets.

In the prediction of PGV, the lowest error terms for the LR, BRR, and SVM algorithms are obtained with the original separated dataset ($5.5 \leq M_w \leq 7.6$). However, the lowest error terms for the GB, RF, and ANN algorithms are obtained with the single original dataset ($3.5 \leq M_w \leq 7.6$). Moreover, ML models trained with separated outlier-eliminated ($5.5 \leq M_w \leq 7.1$) datasets do not present remarkable success in the prediction of PGV according to all performance metrics when tested with the NESS dataset.

It should be emphasized that MSE, RMSE, MAE, and MAPE terms shown in Tables 4.1-4.4 and Figures 4.1-4.2, are much smaller than the terms that show the test results obtained using the NESS dataset (seen Tables 4.5-4.5 and Figures 4.3-4.4). The reason is that all of the models used in this study were trained with the Turkiye dataset. Except for Tables 4.1-4.4 and Figures 4.1-4.2, all results are of the test dataset, which consists of randomly selected recordings from the Turkiye dataset (70% for training and 30% for the test process). On the other hand, the NESS dataset consists of worldwide earthquakes, but the model in this study is trained with the Turkiye dataset. This affects the performance results because the PGV and CAV ranges of Turkiye and worldwide earthquakes are so different from each other; for example, the N-TSMD's CAV range is between 0.155 cm/s and 1395 cm/s while the NESS's CAV values range between 133 cm/s and 3043 cm/s. Similarly, the NESS's PGV range is between 2.397 cm/s and 105.526 cm/s, while the N-TSMD's PGV range is between 0.001 cm/s and 74.531 cm/s (as seen in Table 2.1).

Further analyses to obtain a regional CAV ML-based GMM continued with using the separated dataset without applying outlier elimination due to the success of the algorithms trained with the separated dataset. This indicates that splitting the dataset improves the

effectiveness of ML algorithms for forecasting large earthquakes and eliminates the need for outlier removal.

Further analyses to obtain a regional PGV ML-based GMM, on the other hand, continued with using both the single and separated datasets without applying outlier elimination due to the performance results of the algorithms up to here leaving a dilemma. Therefore, analyses for PGV prediction will be continued with both datasets to make the final decision on whether the model trained with a single dataset or a separated dataset has a better success in the prediction of PGVs.

Table 4.5. Performance metrics (in normalized form) of six trained ML algorithms for PGV prediction of the NESS dataset.

ML Methods	Magnitude Range	Trained Dataset	MAE	MAPE	MSE	RMSE	R
ANN	$3.5 \leq M_w \leq 7.6$	Original	0.2365	0.5263	0.1117	0.3343	0.6374
	$5.5 \leq M_w \leq 7.6$	Original	0.2400	0.5520	0.1155	0.3399	0.5496
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2815	0.6770	0.1406	0.3749	0.3615
GB	$3.5 \leq M_w \leq 7.6$	Original	0.1902	0.5664	0.0717	0.2678	0.4073
	$5.5 \leq M_w \leq 7.6$	Original	0.2440	0.7871	0.0988	0.3144	0.2445
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2188	0.6340	0.0935	0.3058	0.2043
SVM	$3.5 \leq M_w \leq 7.6$	Original	0.3035	0.7458	0.1544	0.3929	0.5450
	$5.5 \leq M_w \leq 7.6$	Original	0.2498	0.5618	0.1196	0.3459	0.4205
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2869	0.6738	0.1451	0.3809	0.4084
LR	$3.5 \leq M_w \leq 7.6$	Original	0.3315	0.8528	0.1760	0.4195	0.2262
	$5.5 \leq M_w \leq 7.6$	Original	0.2890	0.6954	0.1469	0.3832	0.4005
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2943	0.7051	0.1508	0.3884	0.3900
RF	$3.5 \leq M_w \leq 7.6$	Original	0.1913	0.5327	0.0763	0.2762	0.4002
	$5.5 \leq M_w \leq 7.6$	Original	0.2094	0.6411	0.0798	0.2826	0.2706
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2142	0.6515	0.0859	0.2931	0.1688
BRR	$3.5 \leq M_w \leq 7.6$	Original	0.3313	0.8509	0.1760	0.4195	0.2306
	$5.5 \leq M_w \leq 7.6$	Original	0.2895	0.6897	0.1478	0.3845	0.4151
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2929	0.7002	0.1497	0.3869	0.2930

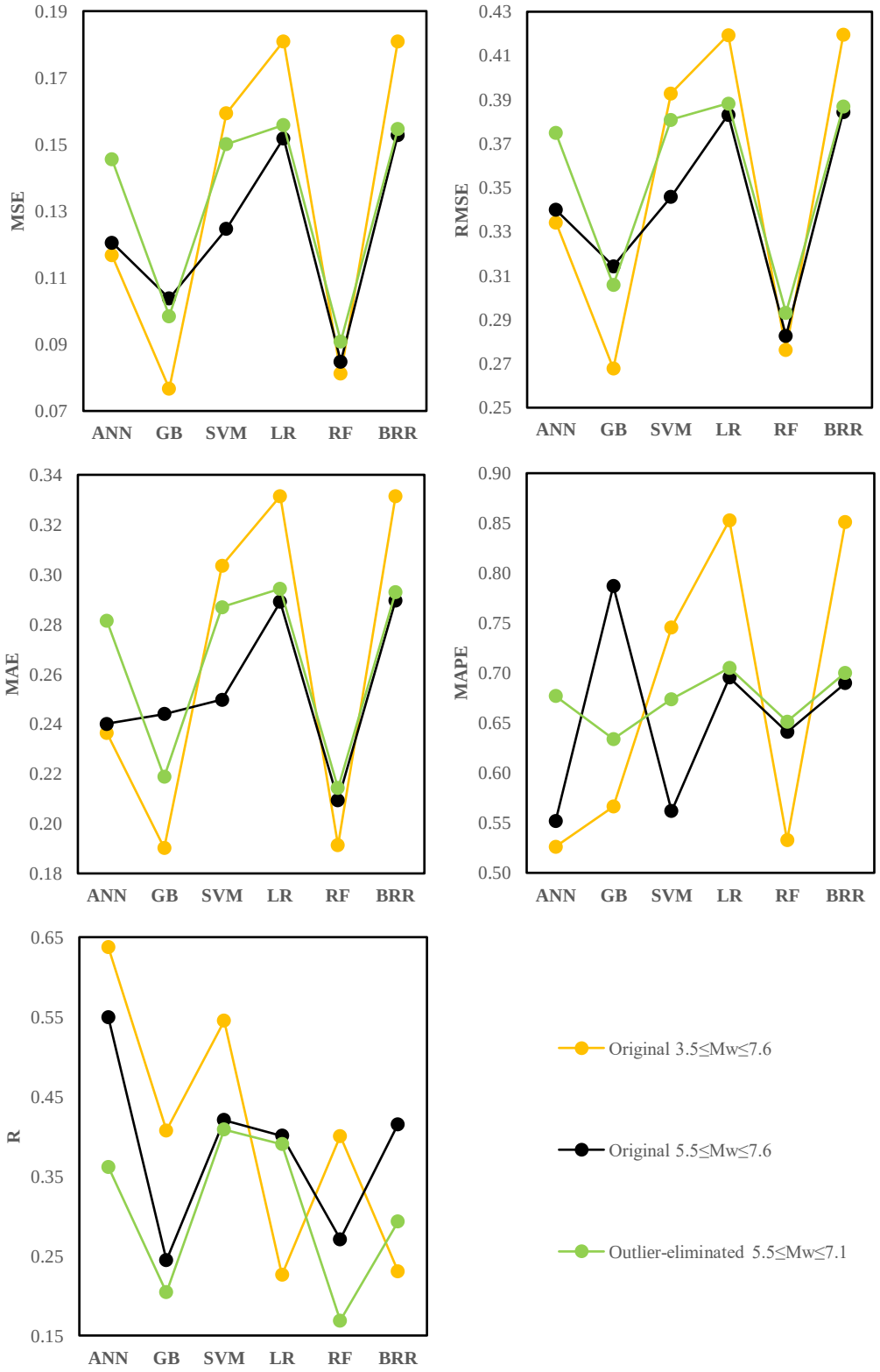


Figure 4.3. Distribution of performance metrics of six trained ML algorithms for PGV prediction of the NESS dataset.

Table 4.6. Performance metrics (in normalized form) of six trained ML algorithms for CAV prediction of the NESS dataset.

ML Methods	Magnitude Range	Trained Dataset	MAE	MAPE	MSE	RMSE	R
ANN	$3.5 \leq M_w \leq 7.6$	Original	0.3632	0.5640	0.2275	0.4770	0.6357
	$5.5 \leq M_w \leq 7.6$	Original	0.3378	0.5036	0.2109	0.4593	0.6558
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.3213	0.5018	0.1858	0.4310	0.6131
GB	$3.5 \leq M_w \leq 7.6$	Original	0.2919	0.4521	0.1762	0.4198	0.2204
	$5.5 \leq M_w \leq 7.6$	Original	0.3189	0.5447	0.1879	0.4334	0.2230
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.3046	0.5726	0.1632	0.4040	0.1649
SVM	$3.5 \leq M_w \leq 7.6$	Original	0.4178	0.6754	0.2776	0.5269	0.5986
	$5.5 \leq M_w \leq 7.6$	Original	0.3456	0.5424	0.2070	0.4550	0.6528
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.3486	0.5791	0.1994	0.4465	0.6510
LR	$3.5 \leq M_w \leq 7.6$	Original	0.4758	0.7947	0.3416	0.5844	0.2724
	$5.5 \leq M_w \leq 7.6$	Original	0.4083	0.6442	0.2740	0.5234	0.5824
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.3987	0.6296	0.2635	0.5133	0.5670
RF	$3.5 \leq M_w \leq 7.6$	Original	0.3040	0.4558	0.1881	0.4337	0.2200
	$5.5 \leq M_w \leq 7.6$	Original	0.2787	0.4751	0.1543	0.3929	0.2388
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.2746	0.5279	0.1387	0.3724	0.2339
BRR	$3.5 \leq M_w \leq 7.6$	Original	0.4728	0.7881	0.3384	0.5817	0.2813
	$5.5 \leq M_w \leq 7.6$	Original	0.4096	0.6438	0.2749	0.5243	0.6574
	$5.5 \leq M_w \leq 7.1$	Outlier-eliminated	0.4044	0.6479	0.2660	0.5158	0.5913

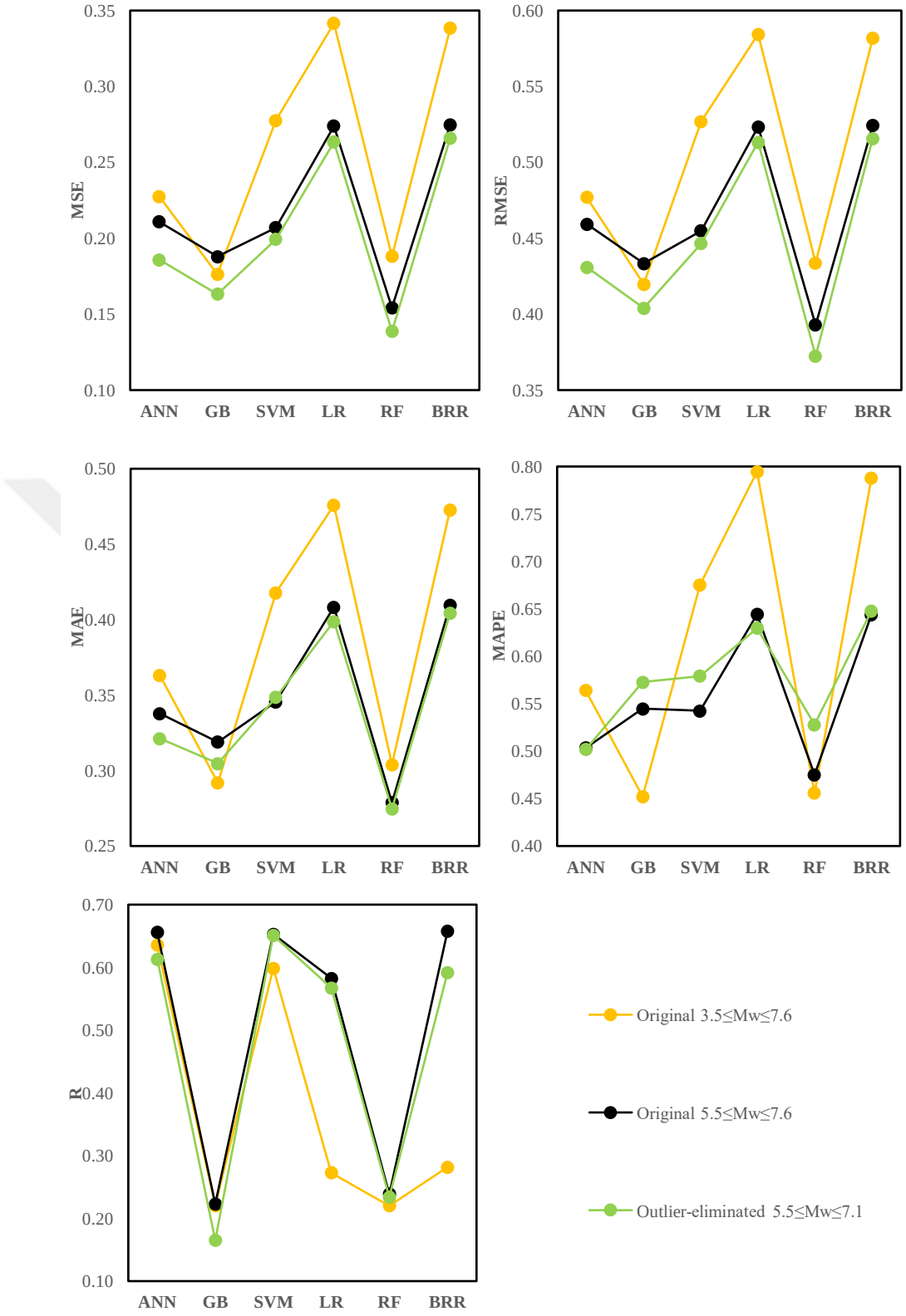


Figure 4.4. Performance metrics of six trained ML algorithms for CAV prediction of the NESS dataset.

4.2. Second Evaluation: Comparison of Feature Selection and ML Methods

To test the effect of feature selection methods on PGV and CAV prediction, three feature selection methods including Pearson, RFE, and LASSO were applied to separated datasets: ($3.5 \leq M_w \leq 5.4$) and ($5.5 \leq M_w \leq 7.6$) datasets. Also, the single dataset ($3.5 \leq M_w \leq 7.6$) was analyzed in the prediction of PGV due to the dilemma mentioned in section 4.1.

Table 4.7-4.8 display input features determined by different techniques. In Pearson's correlation, parameters with a weight value of 0.1 and above were selected as estimator parameters of the model. As seen in Figures 4.5-4.6, the importance coefficients of five out of seven parameters are greater than 0.1 in both separated datasets based on PGV and CAV parameters (except for the $5.5 \leq M_w \leq 7.6$ and $3.5 \leq M_w \leq 7.6$ datasets based on PGV). Vs_{30} estimator remains under the 0.1 threshold value for the $5.5 \leq M_w \leq 7.6$ and $3.5 \leq M_w \leq 7.6$ datasets based on PGV. Considering RFE, M_w, Vs_{30}, and R_{epi} were selected as major estimator parameters for the moderate-to-large magnitude ($5.5 \leq M_w \leq 7.6$) dataset based on PGV and CAV and the single ($3.5 \leq M_w \leq 7.6$) dataset based on PGV, while for the small-magnitude ($3.5 \leq M_w \leq 5.4$) dataset Vs_{30} is replaced by R_{rup}. Lastly, the LASSO method indicates three (M_w, Vs_{30}, and R_{epi}) and five (M_w, Vs_{30}, R_{JB}, R_{rup}, and R_{epi}) estimator parameters for moderate-to-large magnitude ($5.5 \leq M_w \leq 7.6$) and small-magnitude ($3.5 \leq M_w \leq 5.4$) datasets based on PGV and CAV, respectively (Figures 4.7-4.8). On the other hand, six (M_w, Vs_{30}, δ, R_{JB}, R_{rup}, and R_{epi}) estimator parameters were selected as major inputs for the single ($3.5 \leq M_w \leq 7.6$) dataset based on PGV by the LASSO method. In addition, SoF and δ are selected as major estimator parameters in the PGV prediction considering the moderate-to-large magnitude ($5.5 \leq M_w \leq 7.6$) dataset (Figure 4.8). The LASSO method shows that M_w and Vs_{30} estimators are not very sensitive in the prediction of PGV and CAV for small earthquakes. However, the M_w and Vs_{30} are fundamental parameters representing the seismic source and local site effects in all prediction studies. On the other hand, the fact that no method (except for LASSO in the PGV prediction considering the moderate-to-large magnitude dataset) considers SoF as an effective estimator. However, there is a clear relationship between SoF and both PGV and CAV. In light of these concerns, the authors proposed an extra feature set that includes M_w, R_{JB}, Vs_{30}, and SoF estimator parameters. Selection is based on the knowledge obtained from the review of related literature and the

study on the impact of several seismological parameters on ground motion. It may be inferred that the feature selection strategy does not affect model training or prediction performance considering the N-TSMD dataset.

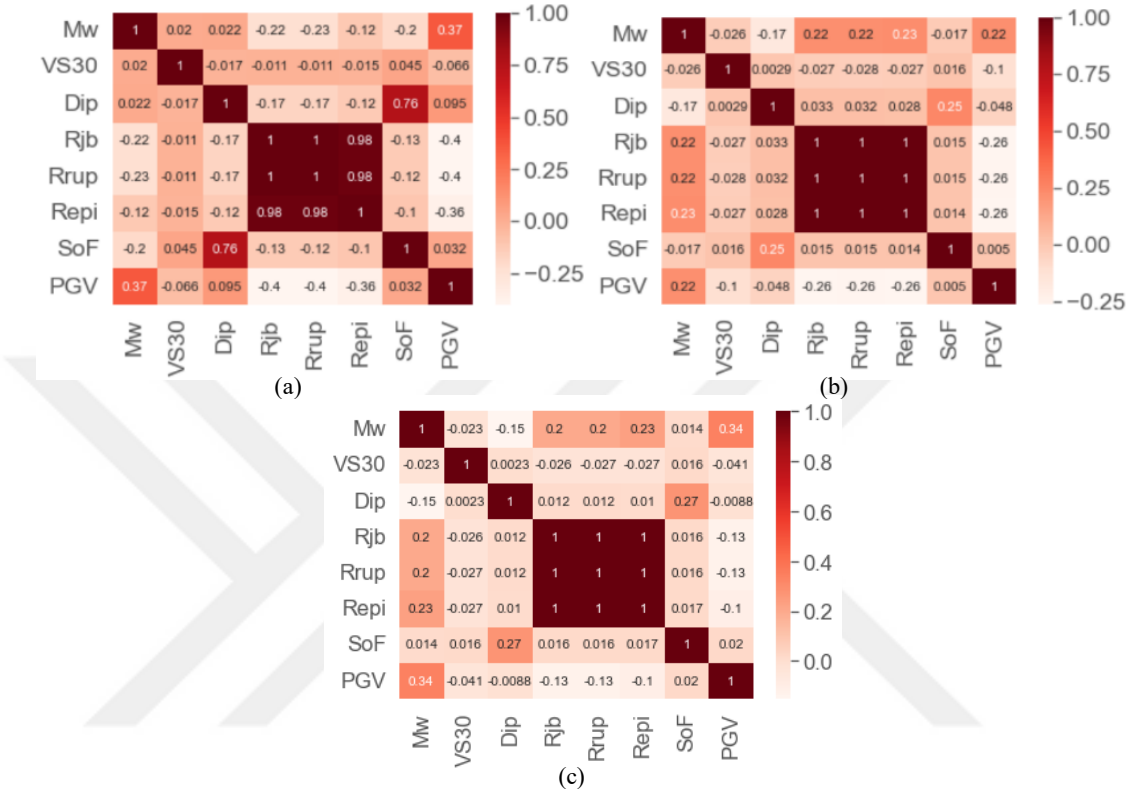


Figure 4.5. Pearson correlation coefficients of features for (a) the $5.5 \leq M_w \leq 7.6$, (b) the PGV $3.5 \leq M_w \leq 5.4$, and (c) the $3.5 \leq M_w \leq 7.6$ datasets.

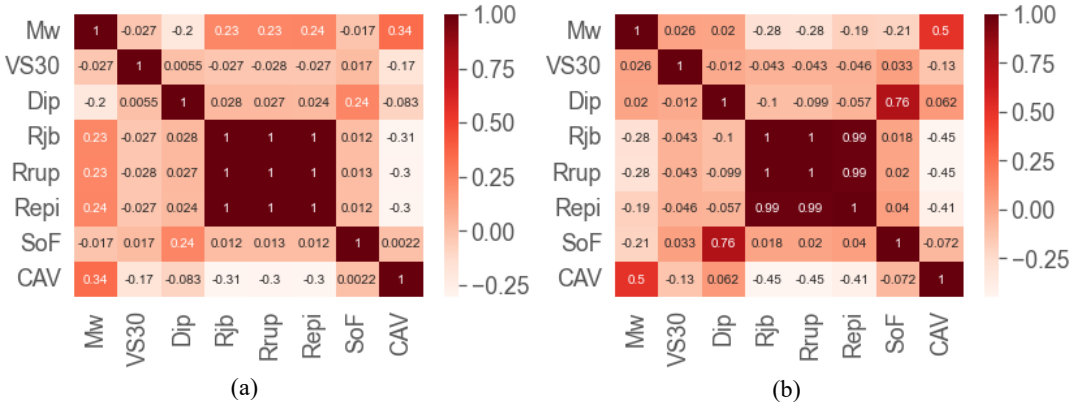


Figure 4.6. Pearson correlation coefficients of features for (a) the CAV $5.5 \leq M_w \leq 7.6$ and (b) the CAV $3.5 \leq M_w \leq 5.4$ datasets.

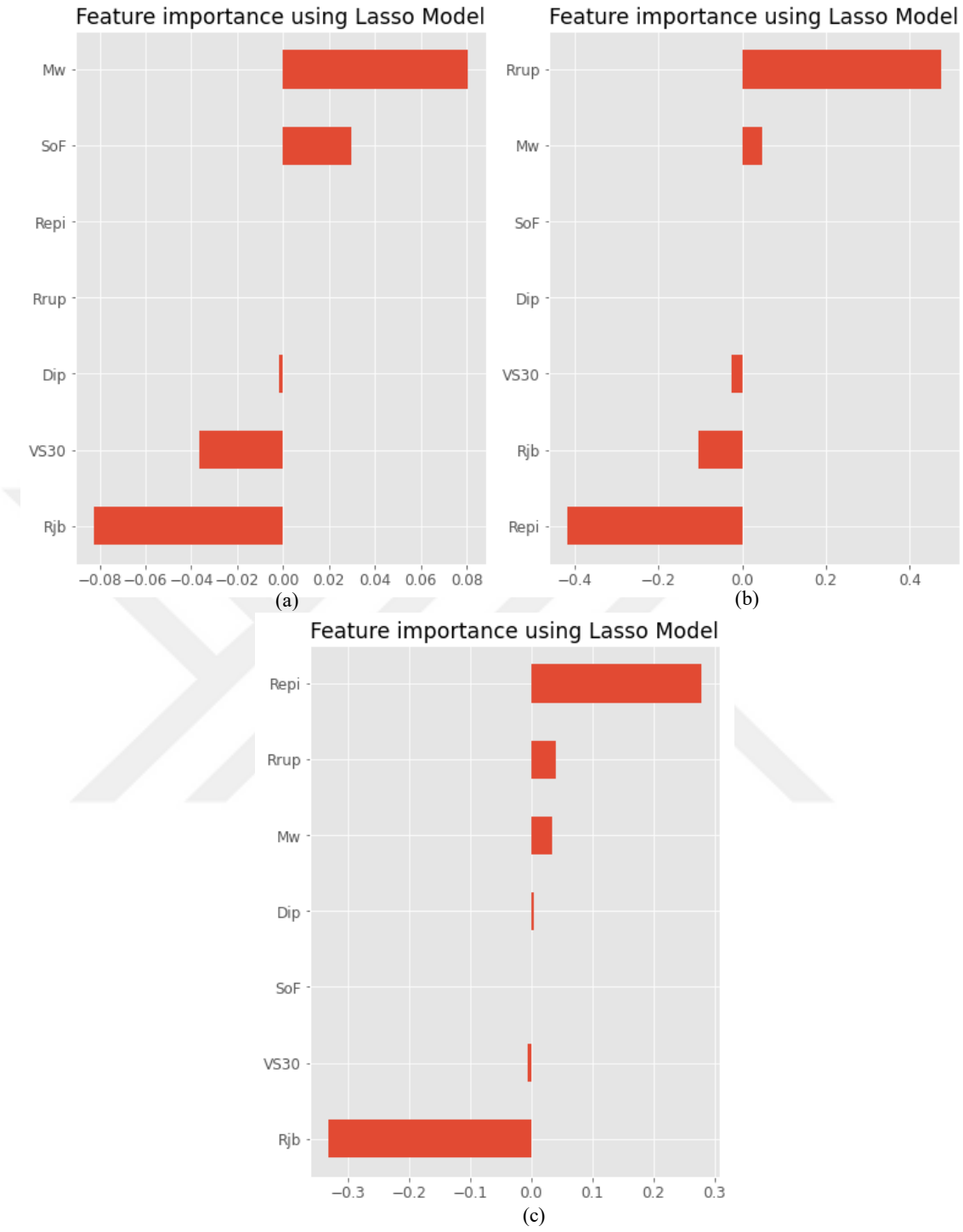


Figure 4.7. Feature importance results obtained by LASSO for (a) the PGV $5.5 \leq M_w \leq 7.6$, (b) the PGV $3.5 \leq M_w \leq 5.4$, and (c) the PGV $3.5 \leq M_w \leq 7.6$ datasets.

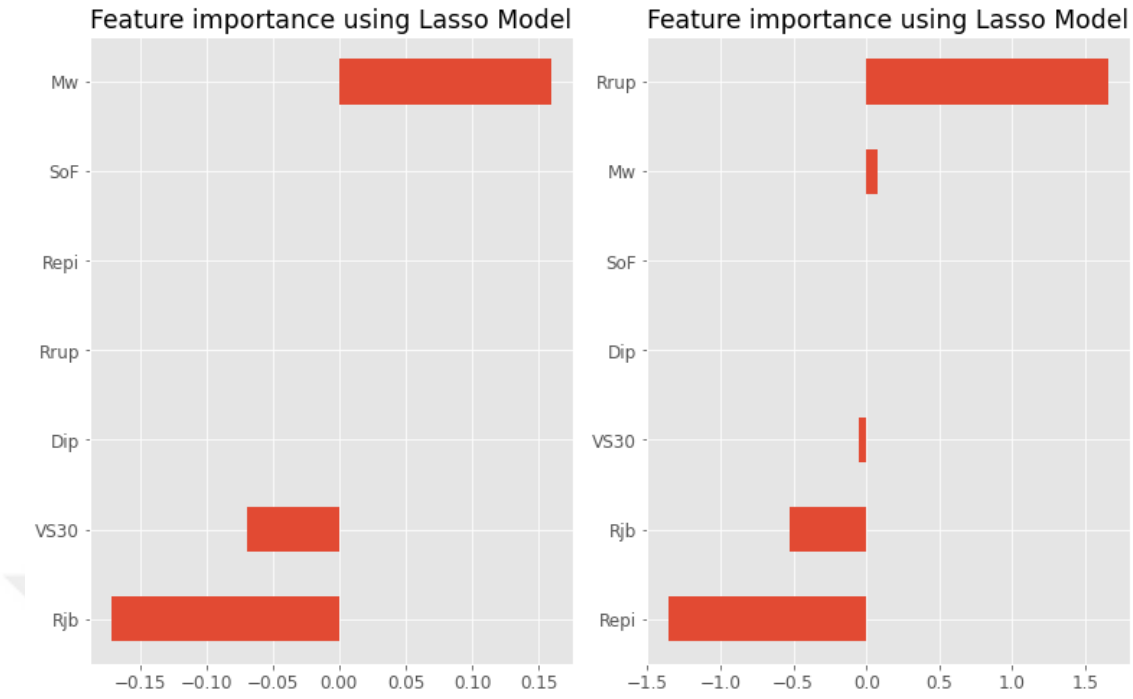


Figure 4.8. Feature importance results obtained by LASSO for (a) the CAV $5.5 \leq M_w \leq 7.6$ and (b) the CAV $3.5 \leq M_w \leq 5.4$ datasets.

Table 4.7. Summary of determining feature sets for (a) the PGV $3.5 \leq M_w \leq 5.4$, (b) the PGV $5.5 \leq M_w \leq 7.6$, and (c) the PGV $3.5 \leq M_w \leq 7.6$ datasets.

(a) $3.5 \leq M_w \leq 5.4$							
Feature Sets	**M_w**	**Vs_{30}**	**δ**	**R_{JB}**	**R_{rup}**	**R_{epi}**	**SoF**
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓		✓	✓	✓	
Pearson (Filter)	✓	✓		✓	✓	✓	
RFE (Wrapper)	✓			✓		✓	
User-selected	✓	✓		✓			✓
(b) $5.5 \leq M_w \leq 7.6$							
Feature Sets	**M_w**	**Vs_{30}**	**δ**	**R_{JB}**	**R_{rup}**	**R_{epi}**	**SoF**
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓	✓	✓			✓
Pearson (Filter)	✓			✓	✓	✓	
RFE (Wrapper)	✓	✓				✓	
User-selected	✓	✓		✓			✓
(c) $3.5 \leq M_w \leq 7.6$							
Feature Sets	**M_w**	**Vs_{30}**	**δ**	**R_{JB}**	**R_{rup}**	**R_{epi}**	**SoF**
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓	✓	✓	✓	✓	
Pearson (Filter)	✓			✓	✓	✓	
RFE (Wrapper)	✓	✓				✓	
User-selected	✓	✓		✓			✓

Table 4.8. Summary of determining feature sets for (a) the CAV $3.5 \leq M_w \leq 5.4$ and (b) the CAV $5.5 \leq M_w \leq 7.6$ separated datasets.

(a) $3.5 \leq M_w \leq 5.4$							
Feature Sets	**M_w**	**Vs_{30}**	**δ**	**R_{JB}**	**R_{rup}**	**R_{epi}**	**SoF**
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓		✓	✓	✓	
Pearson (Filter)	✓	✓		✓	✓	✓	
RFE (Wrapper)	✓			✓		✓	
User-selected	✓	✓		✓			✓
(b) $5.5 \leq M_w \leq 7.6$							
Feature Sets	**M_w**	**Vs_{30}**	**δ**	**R_{JB}**	**R_{rup}**	**R_{epi}**	**SoF**
All	✓	✓	✓	✓	✓	✓	✓
LASSO (Embedded)	✓	✓		✓			
Pearson (Filter)	✓	✓		✓	✓	✓	
RFE (Wrapper)	✓	✓				✓	
User-selected	✓	✓		✓			✓

Prediction of the PGV and CAV were performed using six different state-of-the-art ML techniques. The separated (($3.5 \leq M_w \leq 5.4$) and ($5.5 \leq M_w \leq 7.6$)) datasets without outlier elimination including the single ($3.5 \leq M_w \leq 7.6$) dataset for PGV were used for model training and evaluation using different feature selection methods. The obtained results are given in Subsections 4.2.1 and 4.2.2 by evaluating of different performance metrics to provide a greater depth of analysis.

4.2.1. PGV Result of the Second Evaluation

The original ($3.5 \leq M_w \leq 7.6$) dataset: Table 4.9 supported by Figure 4.9 shows the performance results of ML models on PGV prediction using the single ($3.5 \leq M_w \leq 7.6$) dataset. Namely, the SVM algorithm trained with all inputs provides better accuracy in PGV prediction according to nearly all performance metrics (except for MSE and MAPE). LR algorithm trained with LASSO inputs has the best prediction skills in the prediction of PGV among other input sets according to MSE, MAE, and R. According to MAE and MAPE, the RF algorithm trained with user-selected inputs shows better prediction skills in the prediction of PGV, while RF algorithm trained with Pearson/LASSO set gives more accurate prediction results according to MSE and RMSE. On the other hand, the ANN algorithm trained with user-selected inputs provides better accuracy in PGV prediction according to MSE and RMSE, while the ANN algorithm trained with Pearson inputs provides better accuracy

according to MAE and MAPE. The BRR algorithm trained with LASSO inputs provides better accuracy in PGV prediction according to nearly all performance metrics (except for MAPE). Lastly, the GB algorithm trained with all inputs shows better prediction skills in the prediction of PGV according to MSE and R, while the GB algorithm trained with the user-selected set gives more accurate prediction results according to MAE and RMSE.

Figure 4.9 also demonstrated that the GB algorithm trained with all inputs has the best prediction skill compared to the other five algorithms trained with the same input sets in terms of MSE and R, while the SVM algorithm trained with all inputs shows the best prediction skill according to RMSE and MAE. The ANN algorithm trained with user-selected inputs gives more accurate PGV prediction than the other five algorithms trained with the same input sets according to MSE and RMSE. In addition, the RF algorithm trained with Pearson inputs shows the best prediction skills compared to the other five algorithms trained with the same input sets in terms of all performance metrics (except for MAE and MAPE). The ANN algorithm trained with RFE inputs shows minimum MSE and RMSE as well as the highest R compared to other ML algorithms trained with the same inputs. Lastly, the SVM algorithm trained with the LASSO set shows the best prediction skill according to MAE and MAPE.

The original ($5.5 \leq M_w \leq 7.6$) dataset: As can be tracked from Figure 4.10, ANN and LR algorithms trained with Pearson estimator parameters, RF and BRR algorithms trained with RFE, SVM algorithm trained with all inputs, and GB algorithm trained with LASSO provide better accuracy in the prediction of PGV according to performance metrics. Numerical values of error terms are presented in Table 4.10.

Moreover, Figure 4.10 proves that the SVM algorithm trained with all inputs has the best prediction ability compared to the other five algorithms trained with the same input sets according to MSE, RMSE, and MAE. Also, the SVM algorithm trained with user-selected estimator parameters gives more approximate PGV according to all performance metrics (except R). In the training with Pearson estimator parameters, the ANN algorithm achieves the lowest MSE and RMSE values. The RF algorithm trained with RFE inputs has the best prediction ability compared to the other five algorithms trained with the same input sets

according to all performance metrics (except for MAPE). Furthermore, in the training with LASSO estimator parameters, the GB algorithm achieves remarkable prediction skills compared to the other five algorithms trained with the same input sets according to all performance metrics (except for MAPE).

The original ($3.5 \leq M_w \leq 5.4$) dataset: Table 4.11 supported by Figure 4.11 shows the performance results of ML models on PGV prediction using the ($3.5 \leq M_w \leq 5.4$) dataset. The SVM, BRR, and GB algorithms trained with all inputs provide better accuracy in PGV prediction according to nearly all performance metrics. LR algorithm trained with user-selected inputs has the best prediction skills in the prediction of PGV among other input sets according to MSE and RMSE. According to MAE and MAPE, the RF algorithm trained with all inputs shows better prediction skills in the prediction of PGV, while the RF algorithm trained with Pearson/LASSO set gives more accurate prediction results according to MSE and RMSE. Lastly, the ANN algorithm trained with user-selected inputs provides better accuracy in PGV prediction according to MAPE and R, while the ANN algorithm trained with all inputs provides better accuracy according to MSE and RMSE.

Figure 4.11 also demonstrated that the RF algorithm trained with all inputs has the best prediction skill compared to the other five algorithms trained with the same input sets in terms of MSE, MAE, and R. The ANN algorithm trained with user-selected inputs gives more accurate PGV prediction than the other five algorithms trained with the same input sets according to MSE and RMSE. In addition, the RF algorithm trained with Pearson/LASSO inputs shows the best prediction skills compared to the other five algorithms trained with the same input sets in terms of all performance metrics (except for MAE and MAPE). Lastly, the ANN algorithm trained with RFE inputs shows minimum MSE and RMSE compared to other ML algorithms trained with the same inputs, while the SVM algorithm trained with RFE inputs shows minimum MAE and MAPE compared to other ML algorithms trained with the input of the same approach.

Overall, the single ($3.5 \leq M_w \leq 7.6$) dataset expresses the best prediction skills for PGV compared to the separated ($5.5 \leq M_w \leq 7.6$) dataset according to all MSE, RMSE, and MAE. Indeed, the single dataset is not a loser against the separated dataset considering MAPE and

R terms. Therefore, the PGV GB model obtained using all input set displays the best prediction skill for PGV.

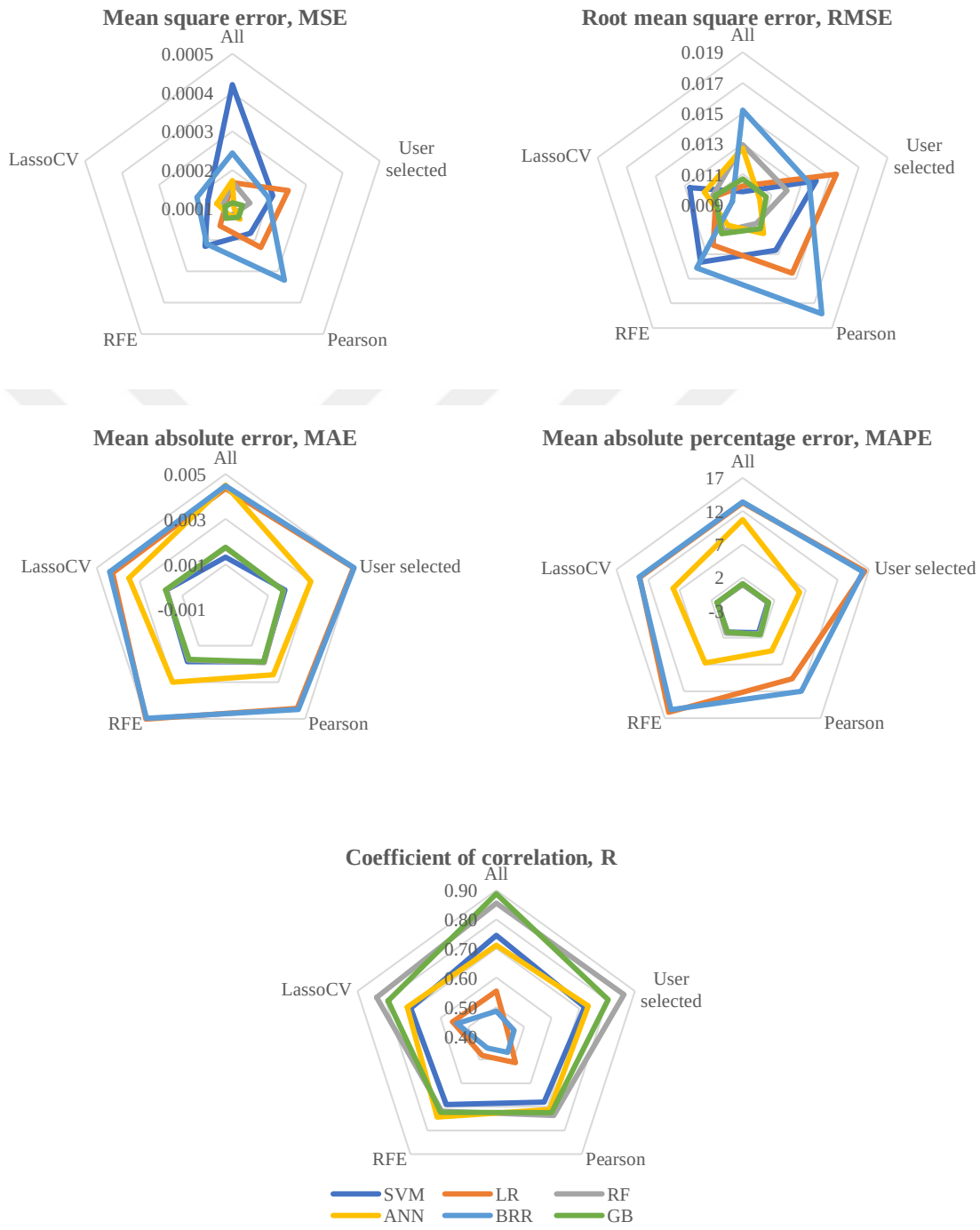


Figure 4.9. Results of PGV prediction analyses performed with the original $3.5 \leq M_w \leq 7.6$ dataset according to performance metrics.

Table 4.9. Summary of comparison of ML techniques for PGV on $3.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models.

(a) MSE					
Average and SD (σ) values of MSE ($x10^{-3}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	0.4201±0.684	0.2093±0.093	0.1791±0.131	0.2203±0.197	0.1659±0.065
LR	0.1673±0.043	0.2514±0.116	0.2238±0.117	0.1547±0.053	**0.1194±0.034**
RF	0.1696±0.045	0.1475±0.038	**0.1136±0.037**	0.1286±0.056	0.1260±0.048
ANN	0.1723±0.107	**0.1039±0.016**	0.1335±0.054	**0.1182±0.050**	0.1423±0.068
BRR	0.2437±0.117	0.1963±0.120	0.3283±0.127	0.2134±0.127	0.1961±0.132
GB	**0.1151±0.021**	0.1271±0.092	0.1266±0.070	0.1307±0.036	0.1196±0.026
(b) RMSE					
Average and SD (σ) values of RMSE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**0.9873±0.287**	1.4080±0.372	1.2674±0.481	1.3659±0.649	1.2641±0.277
LR	1.0212±0.586	1.5471±0.388	1.4548±0.389	1.2281±0.220	1.0838±0.155
RF	1.2933±0.171	1.2068±0.153	**1.0539±0.177**	1.1133±0.241	1.1032±0.233
ANN	1.2710±0.367	**1.0172±0.080**	1.1341±0.247	**1.0670±0.232**	1.1627±0.299
BRR	1.5204±0.396	1.3578±0.386	1.7842±0.353	1.4124±0.417	**0.9693±0.639**
GB	1.0689±0.102	1.0632±0.419	1.0953±0.289	1.1347±0.159	1.0885±0.121
(c) MAE					
Average and SD (σ) values of MAE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**0.1314±0.070**	0.1773±0.013	0.1878±0.026	0.1862±0.040	**0.1749±0.011**
LR	0.4356±0.007	0.4965±0.006	0.4418±0.014	0.5012±0.003	0.4261±0.006
RF	0.1735±0.012	0.1706±0.009	0.1919±0.017	0.1763±0.009	0.1752±0.017
ANN	0.4507±0.065	0.2981±0.026	0.2591±0.034	0.2991±0.030	0.3497±0.033
BRR	0.4475±0.008	0.4997±0.007	0.4490±0.010	0.4969±0.017	0.4390±0.016
GB	0.1754±0.010	**0.1692±0.018**	**0.1875±0.010**	**0.1727±0.010**	0.1788±0.007
(d) MAPE					
Average and SD (σ) values of MAPE					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	1.0247±0.199	**0.9732±0.121**	**1.0921±0.096**	1.0311±0.087	**1.0089±0.129**
LR	13.243±0.868	16.330±0.536	9.6547±5.423	15.898±0.804	13.318±1.055
RF	**0.9124±0.024**	1.0621±0.082	1.3300±0.102	1.1804±0.175	1.0167±0.114
ANN	10.710±1.562	6.0032±0.712	4.4316±1.008	6.7150±1.462	8.0219±1.172
BRR	13.373±0.970	16.031±0.749	12.013±1.078	15.421±1.070	13.386±0.722
GB	1.0679±0.106	1.0989±0.121	1.4563±0.181	**1.0243±0.058**	1.0546±0.081
(e) R					
Average and SD (σ) values of R					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	0.7452±0.040	0.7189±0.024	0.6784±0.042	0.6899±0.024	0.7080±0.023
LR	0.5543±0.018	0.4412±0.061	0.5125±0.020	0.4812±0.044	0.5568±0.014
RF	0.8557±0.009	**0.8604±0.008**	**0.7363±0.042**	0.7189±0.028	**0.8308±0.027**
ANN	0.7114±0.031	0.7320±0.012	0.7111±0.022	**0.7430±0.024**	0.7205±0.017
BRR	0.4854±0.038	0.4644±0.036	0.4687±0.052	0.4505±0.049	0.5410±0.043
GB	**0.8876±0.024**	0.8036±0.059	0.7243±0.055	0.7228±0.085	0.7895±0.029



Figure 4.10. Results of PGV prediction analyses performed with the original $5.5 \leq M_w \leq 7.6$ dataset according to performance metrics.

Table 4.10. Summary of comparison of ML techniques for PGV on $5.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models.

(a) MSE					
Average and SD (σ) values of MSE ($x10^{-3}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**1.3895±0.774**	**1.5503±0.683**	2.1301±1.544	3.5952±3.089	2.1165±1.662
LR	3.3289±1.027	2.0537±1.147	1.7575±0.364	2.1022±1.091	2.6607±1.167
RF	1.3989±0.331	1.9367±0.942	1.3152±0.261	**1.0846±0.272**	1.7766±1.047
ANN	1.9561±1.000	1.8007±1.515	**1.0975±0.342**	1.3630±0.597	2.6562±1.842
BRR	2.2538±1.316	2.6331±1.524	3.4665±2.396	1.6020±1.063	4.6442±2.284
GB	1.4254±0.680	1.9455±1.180	1.2446±0.561	2.7903±1.310	**0.9761±0.497**
(b) RMSE					
Average and SD (σ) values of RMSE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**3.6192±0.998**	**3.8568±0.886**	4.4189±1.489	5.4216±2.863	4.3556±1.656
LR	5.7080±0.940	4.4018±1.204	4.1755±0.417	4.4534±1.219	5.0619±1.108
RF	3.7146±0.488	4.2984±1.055	3.6124±0.358	**3.2703±0.435**	4.0482±1.312
ANN	4.3254±1.032	4.0201±1.518	**3.2741±0.565**	3.6152±0.836	4.9074±1.760
BRR	4.5849±1.376	4.9508±1.508	5.6088±2.002	3.8539±1.208	6.6140±1.836
GB	3.6384±1.126	4.1700±1.607	3.4550±0.798	5.1467±1.329	**3.0369±0.820**
(c) MAE					
Average and SD (σ) values of MAE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**1.1898±0.151**	**1.1325±0.132**	1.4071±0.280	1.5518±0.551	1.3242±0.245
LR	2.3902±0.233	2.3616±0.174	2.1667±0.111	2.2472±0.114	2.4484±0.217
RF	1.3185±0.166	1.3242±0.133	1.5282±0.105	**1.2661±0.133**	1.3097±0.212
ANN	2.0317±0.138	2.1298±0.312	1.6136±0.213	1.8727±0.183	2.0041±0.278
BRR	2.4672±0.197	2.3348±0.155	2.3775±0.258	2.2252±0.121	2.5165±0.138
GB	1.2842±0.250	1.3377±0.379	**1.2973±0.194**	1.5261±0.203	**1.1284±0.147**
(d) MAPE					
Average and SD (σ) values of MAPE					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	1.0931±0.555	**0.7004±0.042**	**0.6910±0.078**	**0.7104±0.090**	**1.8332±1.208**
LR	4.6886±1.622	5.9060±1.941	3.7377±0.597	3.9709±0.456	5.4346±1.310
RF	1.0253±0.367	0.7722±0.039	1.0196±0.070	1.4902±1.167	5.2681±6.033
ANN	4.2946±1.451	6.3314±3.099	1.4342±0.094	2.6359±0.626	3.4319±1.181
BRR	6.4179±1.931	6.5362±1.680	3.7547±1.132	4.8319±1.297	5.0236±1.276
GB	**0.8043±0.108**	0.8192±0.097	1.0486±0.131	1.2652±0.727	3.2423±1.972
(e) R					
Average and SD (σ) values of R					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	0.7016±0.038	0.7043±0.054	0.7035±0.028	0.7193±0.021	0.7063±0.044
LR	0.5292±0.027	0.5853±0.040	0.5930±0.040	0.5751±0.027	0.5574±0.024
RF	**0.8240±0.030**	0.8158±0.009	**0.8096±0.036**	**0.7438±0.046**	0.8023±0.046
ANN	0.6173±0.074	0.6431±0.042	0.7357±0.015	0.6905±0.011	0.6081±0.057
BRR	0.5948±0.040	0.5747±0.040	0.5701±0.059	0.6136±0.043	0.5074±0.016
GB	0.8231±0.028	**0.8670±0.011**	0.8084±0.024	0.6836±0.038	**0.8956±0.034**

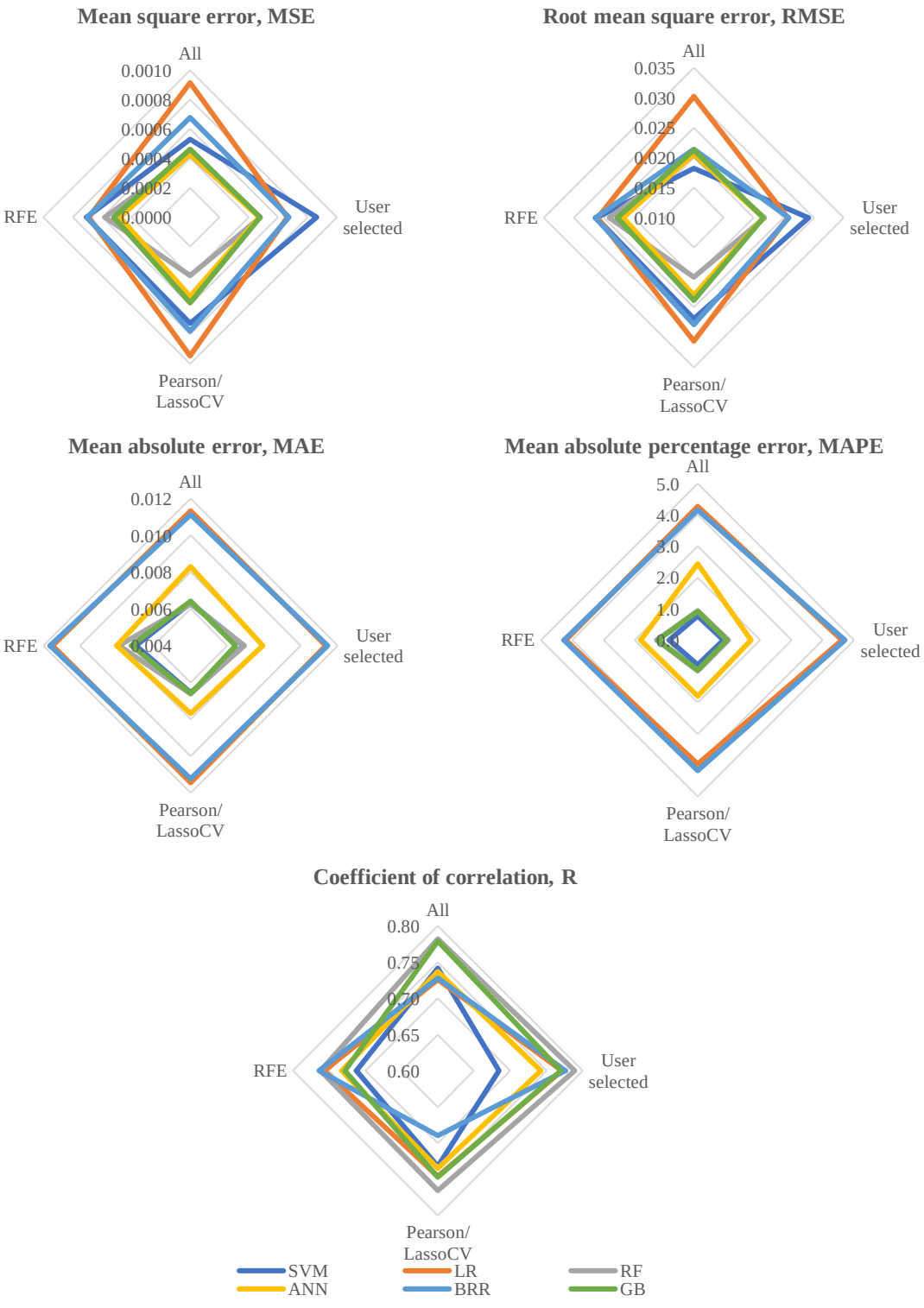


Figure 4.11. Results of PGV prediction analyses performed with $3.5 \leq M_w \leq 5.4$ dataset according to performance metrics.

Table 4.11. Summary of comparison of ML techniques for PGV on $3.5 \leq M_w \leq 5.4$ datasets using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models.

(a) MSE				
Average and SD (σ) values of MSE ($x10^{-3}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter)/ LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.5294±0.092	0.8617±0.225	0.7243±0.092	0.7051±0.266
LR	0.9147±0.090	0.6618±0.071	0.9456±0.202	0.6902±0.196
RF	**0.4247±0.119**	0.4782±0.082	**0.3988±0.053**	0.5796±0.044
ANN	0.4283±0.094	**0.4708±0.098**	0.5400±0.067	**0.4783±0.097**
BRR	0.6777±0.232	0.6724±0.103	0.7776±0.084	0.7007±0.110
GB	0.4619±0.139	0.4768±0.147	0.5840±0.222	0.5198±0.083
(b) RMSE				
Average and SD (σ) values of RMSE ($x10^{-2}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter)/ LASSO (Embedded)**	**RFE (Wrapper)**
SVM	**1.8215±1.002**	2.9140±0.395	2.6868±0.173	2.6207±0.478
LR	3.0215±0.150	2.5695±0.139	3.0610±0.328	2.6024±0.403
RF	2.0453±0.283	2.1804±0.187	**1.9936±0.131**	2.4061±0.091
ANN	2.0593±0.229	**2.1606±0.222**	2.2929±0.155	**2.1775±0.228**
BRR	2.1398±1.232	2.5870±0.197	2.7853±0.153	2.6403±0.211
GB	2.1280±0.337	2.1618±0.345	2.3810±0.461	2.2735±0.192
(c) MAE				
Average and SD (σ) values of MAE ($x10^{-2}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter)/ LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.6318±0.007	0.6888±0.044	**0.6560±0.031**	**0.6968±0.056**
LR	1.1324±0.016	1.1349±0.015	1.1450±0.041	1.1495±0.033
RF	**0.6258±0.036**	0.6925±0.012	0.6621±0.021	0.7885±0.008
ANN	0.8292±0.012	0.7924±0.056	0.7686±0.045	0.8011±0.091
BRR	1.1121±0.021	1.1452±0.021	1.1236±0.017	1.1636±0.012
GB	0.6419±0.028	**0.6449±0.038**	0.6609±0.051	0.7203±0.028
(d) MAPE				
Average and SD (σ) values of MAPE				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	**0.7730±0.030**	**0.8269±0.012**	**0.7910±0.013**	**0.8994±0.027**
LR	4.2702±0.390	4.6135±0.240	3.9580±0.097	4.2081±0.130
RF	0.9343±0.015	0.9839±0.036	0.9850±0.036	1.3014±0.022
ANN	2.4366±0.338	1.7077±0.157	1.7894±0.271	1.8165±0.351
BRR	4.1569±0.226	4.7071±0.337	4.1686±0.155	4.2647±0.258
GB	0.9186±0.018	0.9363±0.035	0.9568±0.021	1.2672±0.018
(e) R				
Average and SD (σ) values of R				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.7416±0.014	0.6843±0.010	0.7318±0.019	0.7124±0.009
LR	0.4542±0.007	0.4434±0.007	0.4740±0.010	0.4288±0.027
RF	0.7819±0.013	**0.7888±0.022**	**0.7653±0.017**	**0.7618±0.008**
ANN	0.7368±0.026	0.7422±0.021	0.7345±0.008	0.7326±0.007
BRR	0.5105±0.018	0.4360±0.009	0.4720±0.007	0.4238±0.020
GB	**0.7789±0.018**	0.7698±0.016	0.7467±0.012	0.7282±0.022

4.2.2. CAV Result of the Second Evaluation

The original ($5.5 \leq M_w \leq 7.6$) dataset: As can be tracked from Figure 4.12, SVM and RF algorithms trained with Pearson estimator parameters, LR and GB algorithms trained with user-selected estimator parameters, ANN algorithm trained with RFE, and BRR algorithm trained with both Pearson and RFE provide better accuracy in the prediction of CAV according to performance metrics. Numerical values of error terms are presented in Table 4.12.

Moreover, Figure 4.12 proves that the GB algorithm trained with all inputs has the best prediction ability compared to the other five algorithms trained with the same input sets according to all performance metrics (except MAE). The GB algorithm trained with user-selected estimator parameters gives more approximate CAV according to all performance metrics (except MAPE). In the training with LASSO estimator parameters, the GB algorithm achieves the lowest MSE, RMSE, and MAE values. Furthermore, Pearson and RFE estimator parameters employed in the training of RF and SVM algorithms, respectively show minimum error metrics compared to other algorithms.

The original ($3.5 \leq M_w \leq 5.4$) dataset: Table 4.13 supported by Figure 4.13 shows the performance results of ML models on CAV prediction using the ($3.5 \leq M_w \leq 5.4$) dataset. Namely, SVM, BRR, and ANN algorithms trained with Pearson/LASSO sets, LR and RF algorithms trained with all inputs, and GB algorithm trained with both all and user-selected inputs provide better accuracy in CAV prediction according to performance metrics.

Figure 4.13 also demonstrated that the RF algorithm trained with all inputs has the best prediction skill compared to the other five algorithms trained with the same input sets in terms of MSE, RMSE, and R. The GB algorithm trained with user-selected inputs gives more accurate CAV prediction than the other five algorithms trained with the same input sets according to all performance metrics (except RMSE). In addition, the GB algorithm trained with Pearson/LASSO inputs gives minimum MAE and maximum R value compared to other ML algorithms trained with Pearson/LASSO, while the RF algorithm trained with Pearson/LASSO selected inputs shows minimum MSE and RMSE compared to other ML

algorithms trained with the input of the same approach. In total, GB and RF approaches are better techniques for CAV prediction using small earthquake events.

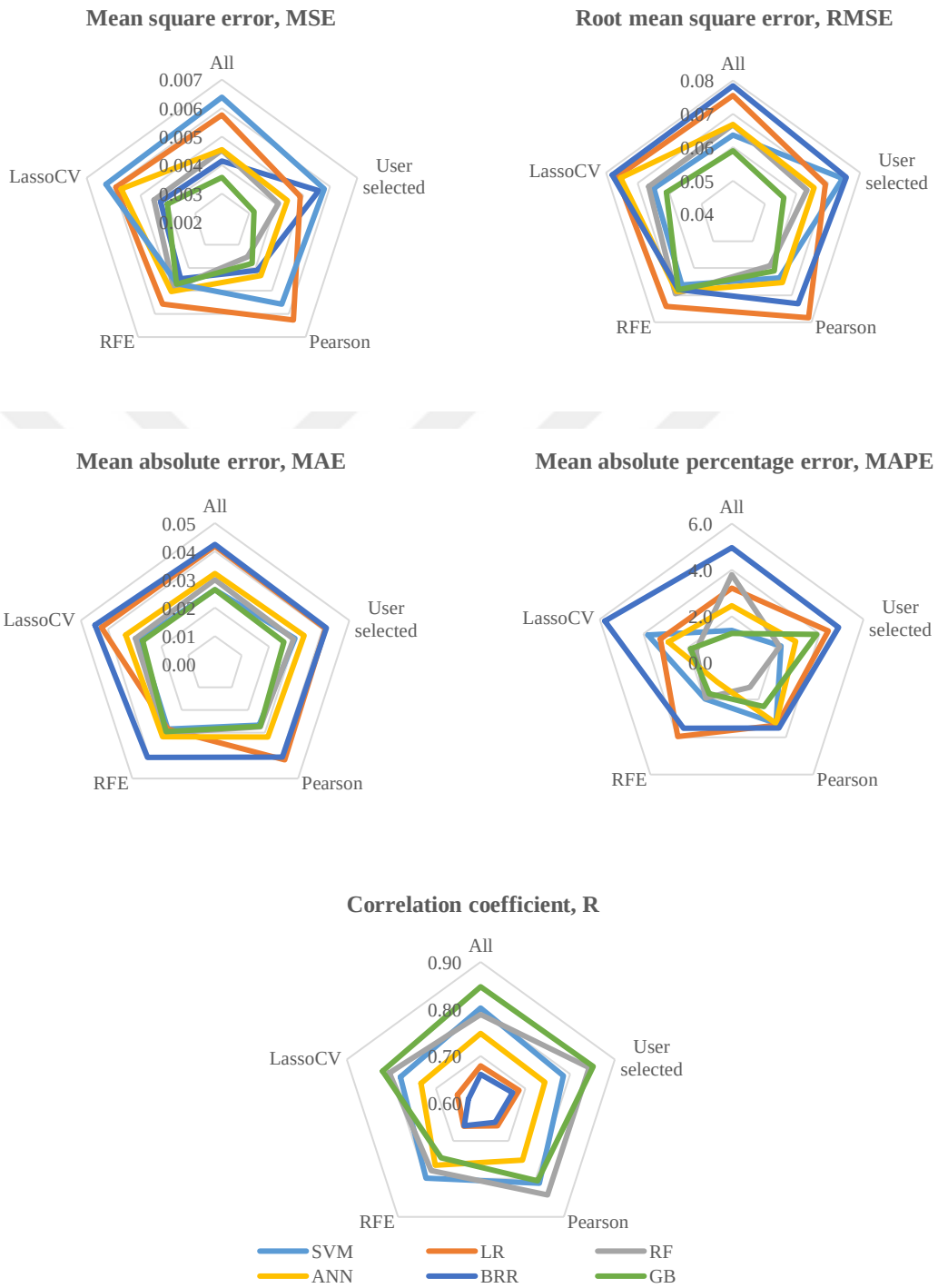


Figure 4.12. Results of CAV prediction analyses performed with the original $5.5 \leq M_w \leq 7.6$ dataset according to performance metrics.

Table 4.12. Summary of comparison of ML techniques for CAV on $5.5 \leq M_w \leq 7.6$ dataset using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models.

(a) MSE					
Average and SD (σ) values of MSE ($x10^{-3}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	4.1387±1.195	5.5787±0.873	4.1128±1.150	**4.4695±1.366**	4.2686±1.065
LR	5.7563±1.281	4.8904±1.724	6.2600±1.971	5.5662±1.478	5.9160±1.268
RF	4.5173±1.190	4.0857±1.272	**3.5333±0.907**	4.8733±1.080	4.5087±1.412
ANN	4.5356±1.228	4.4170±1.679	4.3230±1.229	5.0242±2.917	5.7244±1.092
BRR	6.3796±2.788	5.7795±1.273	5.5727±2.519	4.6739±1.373	6.2855±2.557
GB	**3.5634±1.242**	**3.1846±0.903**	3.7945±1.101	4.7045±1.349	**4.0157±1.572**
(b) RMSE					
Average and SD (σ) values of RMSE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	6.3673±1.027	7.4520±0.563	6.3597±0.923	**6.6208±1.036**	6.4959±0.781
LR	7.5516±0.819	6.9061±1.229	7.8347±1.233	7.4033±1.032	7.6529±0.860
RF	6.6756±0.872	6.3276±1.011	**5.9033±0.778**	6.9449±0.791	6.6555±0.994
ANN	6.6824±0.935	6.5508±1.285	6.5233±0.919	6.8635±1.979	7.5386±0.719
BRR	7.8399±1.707	7.5642±0.849	7.3153±1.663	6.7844±0.942	7.8038±1.563
GB	**5.8961±1.042**	**5.5971±0.805**	6.0992±0.964	6.7983±1.017	**6.0992±1.192**
(c) MAE					
Average and SD (σ) values of MAE ($x10^{-2}$)					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	**2.6397±0.310**	2.9753±0.076	**2.6624±0.231**	**2.8346±0.243**	2.7948±0.382
LR	4.1823±0.131	4.1111±0.318	4.1823±0.209	2.8346±0.255	4.2328±0.211
RF	2.9978±0.338	2.9236±0.312	2.7355±0.330	3.0074±0.255	2.9633±0.301
ANN	3.2148±0.341	3.3152±0.330	3.1842±0.197	3.1667±0.471	3.3281±0.263
BRR	4.2441±0.274	4.1451±0.196	4.0687±0.202	4.0744±0.214	4.4724±0.194
GB	2.6448±0.377	**2.5622±0.233**	2.7179±0.459	2.9457±0.186	**2.6953±0.231**
(d) MAPE					
Average and SD (σ) values of MAPE					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	1.3802±1.221	2.2474±2.006	3.2641±3.181	1.9590±1.817	3.8088±3.136
LR	3.2090±1.902	4.3971±1.730	3.3485±1.696	3.9689±1.477	3.2418±1.053
RF	3.7916±2.146	**2.1878±1.838**	**1.3278±1.140**	1.9030±1.347	**1.6218±0.700**
ANN	2.4499±2.101	2.9147±2.249	3.2627±2.370	**1.0497±0.261**	2.8820±1.911
BRR	4.9620±1.486	4.8669±2.881	3.5012±1.779	3.5250±1.677	5.7669±1.713
GB	**1.2503±0.633**	3.8711±3.497	2.3659±2.146	1.6659±0.847	1.8645±2.359
(e) R					
Average and SD (σ) values of R					
Techniques	**All**	**User-selected**	**Pearson (Filter)**	**RFE (Wrapper)**	**LASSO (Embedded)**
SVM	0.8019±0.024	0.7849±0.004	0.8114±0.029	**0.7976±0.008**	0.7799±0.025
LR	0.6788±0.018	0.6855±0.026	0.6597±0.014	0.6613±0.006	0.6523±0.004
RF	0.7881±0.016	0.8425±0.005	**0.8421±0.026**	0.7789±0.017	0.8047±0.012
ANN	0.7478±0.027	0.7434±0.009	0.7513±0.011	0.7641±0.026	0.7339±0.016
BRR	0.6606±0.032	0.6712±0.019	0.6511±0.024	0.6600±0.018	0.6280±0.038
GB	**0.8471±0.045**	**0.8516±0.007**	0.8051±0.016	0.7438±0.013	**0.8199±0.021**

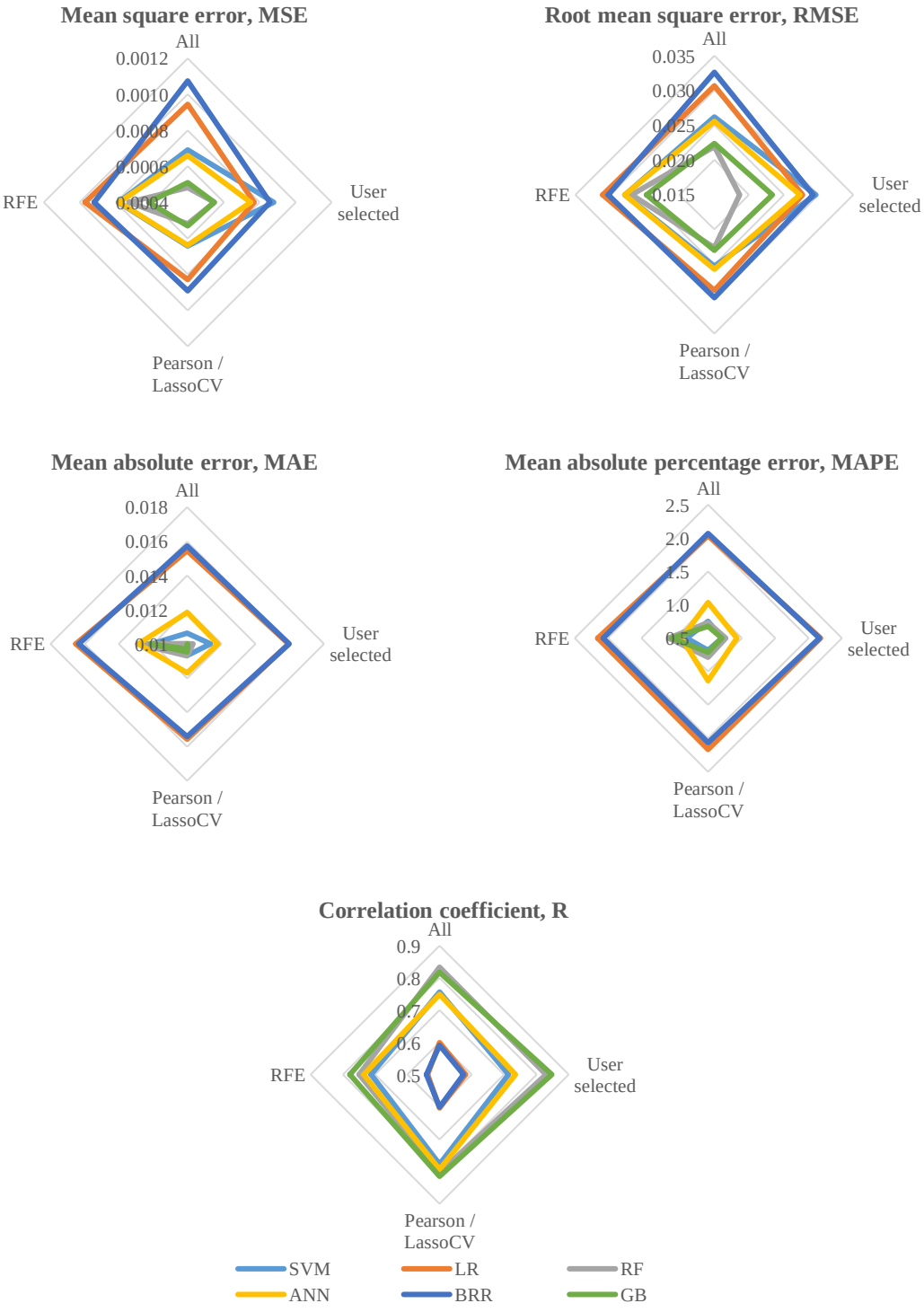


Figure 4.13. Results of CAV prediction analyses performed with $3.5 \leq M_w \leq 5.4$ dataset according to performance metrics.

Table 4.13. Summary of comparison of ML techniques for CAV on $3.5 \leq M_w \leq 5.4$ datasets using performance metrics: (a) MSE, (b) RMSE, (c) MAE, (d) MAPE, and (e) R metrics (in normalized form). Bold values represent the best results in terms of ML models.

(a) MSE				
Average and SD (σ) values of MSE ($x10^{-3}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.6914±0.138	0.8780±0.104	0.6427±0.092	0.7843±0.173
LR	0.9446±0.136	0.7693±0.078	0.8291±0.068	0.9704±0.162
RF	**0.4817±0.055**	0.5483±0.063	**0.5194±0.039**	0.7262±0.091
ANN	0.6600±0.149	0.7501±0.104	0.6384±0.023	0.7803±0.085
BRR	1.0747±0.230	0.8569±0.111	0.8915±0.078	0.9193±0.122
GB	0.5095±0.054	**0.5474±0.077**	0.5310±0.061	**0.6163±0.117**
(b) RMSE				
Average and SD (σ) values of RMSE ($x10^{-2}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	2.6184±0.277	2.9592±0.171	2.5298±0.186	2.7859±0.320
LR	3.0673±0.218	2.7708±0.139	2.8774±0.119	3.1060±0.266
RF	2.1920±0.125	**1.8645±1.017**	**2.2627±0.075**	2.6907±0.167
ANN	2.5550±0.300	2.7335±0.189	2.5735±0.117	2.7901±0.151
BRR	3.2618±0.366	2.9224±0.188	2.9837±0.129	3.0267±0.201
GB	**2.2365±0.135**	2.3351±0.164	2.3013±0.132	**2.4745±0.224**
(c) MAE				
Average and SD (σ) values of MAE ($x10^{-2}$)				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	1.0632±0.051	1.1362±0.035	1.0675±0.025	**1.2077±0.050**
LR	1.5453±0.012	1.5928±0.020	1.5571±0.017	1.6515±0.036
RF	1.0024±0.025	1.0339±0.033	1.0700±0.012	1.2887±0.026
ANN	1.1839±0.029	1.1793±0.066	1.1690±0.051	1.2857±0.030
BRR	1.5729±0.046	1.5958±0.024	1.5413±0.026	1.6344±0.034
GB	**0.9531±0.023**	**1.0006±0.031**	**1.0307±0.028**	1.2127±0.035
(d) MAPE				
Average and SD (σ) values of MAPE				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.7538±0.140	**0.7007±0.010**	**0.6813±0.017**	**0.8235±0.022**
LR	2.0353±0.090	2.1802±0.073	2.1648±0.142	2.1554±0.209
RF	0.7342±0.013	0.7612±0.008	0.7854±0.031	1.0928±0.036
ANN	1.0345±0.081	0.9265±0.148	1.1363±0.273	0.9007±0.090
BRR	2.0681±0.095	2.1631±0.069	2.0627±0.097	2.0738±0.064
GB	**0.6788±0.012**	0.7021±0.016	0.7105±0.012	1.0443±0.031
(e) R				
Average and SD (σ) values of R				
Techniques	**All**	**User-selected**	**Pearson (Filter) / LASSO (Embedded)**	**RFE (Wrapper)**
SVM	0.7564±0.004	0.7135±0.002	0.7772±0.015	0.7125±0.012
LR	0.5996±0.005	0.5805±0.014	0.6030±0.008	0.5375±0.004
RF	0.8334±0.014	0.8297±0.008	0.8031±0.009	0.7496±0.014
ANN	0.7491±0.006	0.7349±0.006	0.7933±0.007	0.7352±0.016
BRR	0.5913±0.002	0.5740±0.010	0.6002±0.008	0.5410±0.007
GB	**0.8176±0.011**	**0.8476±0.004**	**0.8149±0.017**	**0.7783±0.009**

4.3. Development of Turkiye-specific GMMs for PGV and CAV Using the Expanded N-TSMD

At the finalization phase of this study, the 2023 Kahramanmaraş doublet earthquakes (M_w7.8 and M_w7.7) and subsequently hundreds of aftershocks with magnitude as high as M_w6.8 occurred at southeastern Turkiye. Preparation of the PGV and CAV GMMs without these valuable data would narrow the applicability intervals in terms of magnitude and distance measures. Hence, the N-TSMD is enriched with the earthquakes occurred in between 2020 and 2023 from the recent SMD-TR. Up to here, the GB algorithm trained with the original single dataset ($3.5 \leq M_w \leq 7.6$) in the prediction of PGV and the original seperated dataset ($3.5 \leq M_w \leq 5.4$ and $5.5 \leq M_w \leq 7.6$) in the prediction of CAV perform superior success, as discussed at sections 4.1 and 4.2. Therefore, in the development of the most recent Turkiye-specific GMMs for PGV and CAV, the expanded N-TSMD is used to further train and test process. The expanded N-TSMD includes 17,856 recordings of 818 earthquakes with M_w ranging from 3.5 and 7.8, R_{JB} up to 200 km, and Vs_{30} ranging from 131 m/s and 1862 m/s, as seen section 2.

The final GB GMMs obtained with the expanded N-TSMD (proposed model) were assessed using the same performance metrics and compared with the model obtained the N-TSMD consisting of 15,110 recordings up to 2020. As expected, the proposed models indicate robust predictive abilities for PGV and CAV compared to the models developed with the N-TSMD according to all performance metrics (Tables 4.14-4.15).

Table 4.14. Comparison of the readvanced model and the model obtained with only the N-TSMD for PGV according to performance metrics (in normalized form).

Metrics	PGV proposed model	Model developed with only N-TSMD
MSE ($x10^{-3}$)	0.1034±0.019	0.1151±0.021
RMSE ($x10^{-2}$)	1.0346±0.076	1.0689±0.102
MAE ($x10^{-2}$)	0.1321±0.014	0.1754±0.010
MAPE	1.0263±0.142	1.0679±0.106
R	0.8979±0.018	0.8876±0.024

Table 4.15. Comparison of the readvanced model and the model obtained with only the N-TSMD for CAV according to performance metrics (in normalized form).

(a) 3.5≤M_w≤5.4		
Metrics	**CAV proposed model**	**Model developed with only N-TSMD**
MSE (x10^{-3})	0.5205±0.067	0.5474±0.077
RMSE (x10^{-2})	2.3003±0.179	2.3351±0.164
MAE (x10^{-2})	1.0004±0.027	1.0006±0.031
MAPE	0.6932±0.013	0.7021±0.016
R	0.8580±0.005	0.8476±0.004
(b) 5.5≤M_w≤7.8 (7.6 for the N-TSMD)		
MSE (x10-3)	2.3878±0.9324	3.1846±0.903
RMSE (x10^{-2})	4.8224±0.8822	5.5971±0.805
MAE (x10^{-2})	1.6970±0.2388	2.5622±0.233
MAPE	0.6422±0.8918	3.8711±3.497
R	0.8822±0.0156	0.8516±0.007

To check the reliability of the proposed model, residual analyses were performed following Abrahamson and Youngs [90]. The total residual is partitioned into two parts: inter-event and intra-event residuals.

$$\ln(PGV_{measured})_{ij} = \ln(PGV_{predicted})_{ij} + \eta_i + \varepsilon_{ij} \quad (4.1)$$

$$\ln(CAV_{measured})_{ij} = \ln(CAV_{predicted})_{ij} + \eta_i + \varepsilon_{ij} \quad (4.2)$$

where η_i indicates the random effect for the i^{th} event and η_i and ε_{ij} show the inter-event and intra-event residuals, respectively.

Inter-event and intra-event residuals of PGV are demonstrated in Figure 4.14. A general slightly overestimation is apparent in all magnitude ranges (approx. -0.24). No trends are detected in the intra-event residual of R_{JB}, R_{rup}, and R_{epi} plots shown in Figure 4.14. The performance of the model is high even at close distances. In addition, no trends are detected in the intra-event residual of Vs_{30}. The proposed model is consistent with (residual approx. -0.08) the observed PGV in the range of 131 m/s≤Vs_{30}≤400 m/s while slightly underestimating (residual approx. 0.1) it for higher Vs_{30}. Such a result is mainly related to

the fact that there are a few recordings on stiff/rock sites in the dataset. In general, inter- and intra-residual behaviors show consistency with the data.

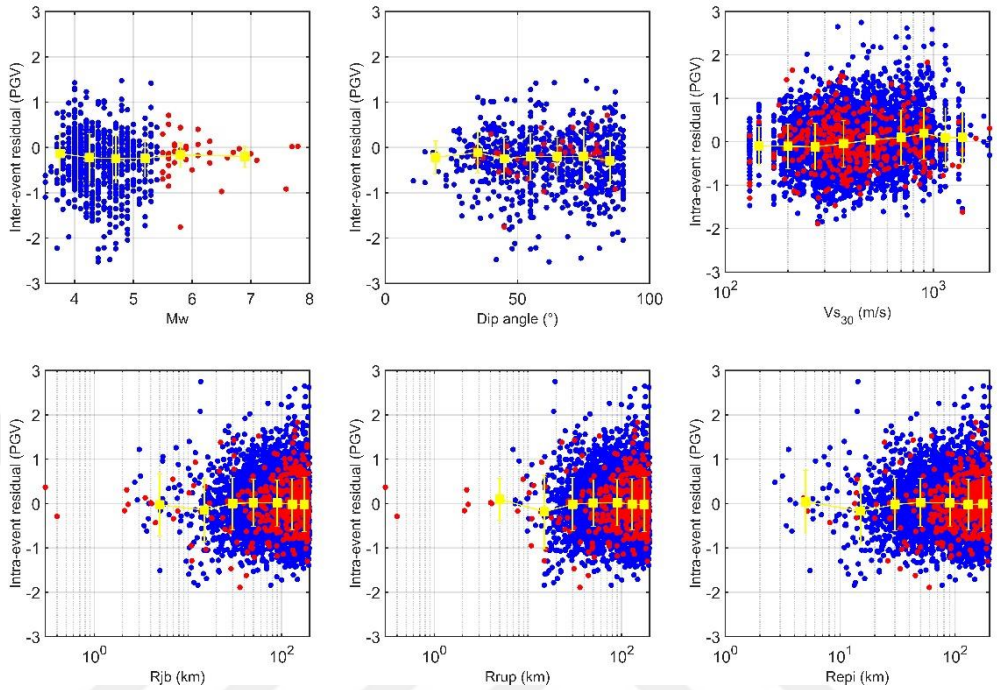


Figure 4.14. Residuals of PGVs predicted with the proposed model (Yellow squares show the mean residuals and yellow error lines are their standard deviations. Blue and red dots represent $M_w \leq 5.4$ and $M_w \geq 5.5$ recordings, respectively).

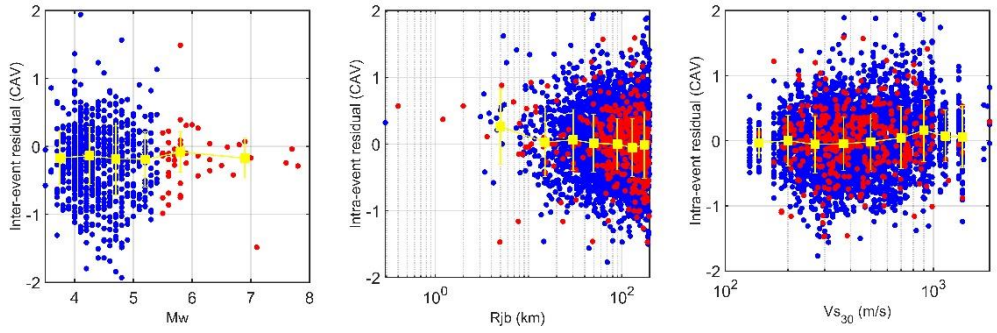


Figure 4.15. Residuals of CAVs predicted with the proposed model (Yellow squares show the mean residuals and yellow error lines are their standard deviations. Blue and red dots represent $M_w \leq 5.4$ and $M_w \geq 5.5$ recordings, respectively).

Inter-event and intra-event residuals of CAV are demonstrated in Figure 4.15. Overall, there is almost no trend in the distribution of the intra-event residuals. A very slight negative inter-event residual throughout the magnitude range of (3.5-7.8) depicts an overestimation (<-0.17) in the prediction of CAV. Hence model predictions can be considered unbiased in all ranges.

As seen in Figures 4.15, in nearly all magnitude ($3.5 \leq M_w \leq 7.8$) bins the average inter-event residuals of CAV are closer to zero (residual approx. -0.15). In contrast to PGV residuals, a better trend is also observed for intra-event CAV residuals in all V_{S30} (approx. 0.02) and R_{JB} (approx. 0.04) ranges. Therefore, the GB model shows superior success in the prediction of CAV than PGV prediction. It should also be noted that CAV is the cumulative total energy parameter, while PGV shows the maximum absolute pulse in the recordings. Therefore, PGVs are more difficult to predict than CAVs since the dataset includes a quiet number of pulse-type recordings with very high amplitude. The results of this study also confirm this difference.

4.4. Comparison of the Proposed Model with the Existing GMMs

To demonstrate the effectiveness of the proposed Turkiye-specific GB model for PGV prediction, two comprehensive GMMs (CY14 and BSSA14) and one ANN-based GMM (KMHCAL23) were implemented and compared in this section. Also, two well-known GMMs adjusted for Turkiye, (TR-SA17 and TR-CB19) were used and compared in this part to show the efficacy of the suggested GB-based Turkiye-specific model for CAV prediction.

The expanded N-TSMD dataset with 1239 ($5.5 \leq M_w \leq 7.8$) and 16,617 ($3.5 \leq M_w \leq 5.4$) strong-motion recordings for CAV and with 17,856 ($3.5 \leq M_w \leq 7.8$) for PGV were utilized for model development. Since only 4% of the recordings in the database belong to R faulting earthquakes, the ML-based models are only approved for SS and N faulting types.

The predicted PGV and CAV values for scenario cases were calculated with the GB models as well as the existing GMMs for N and SS earthquakes with M_w (5.0, 6.0, 7.0) and

R_{JB} (1:200) parameter variations shown in Figures 4.16-4.17. Vs_{30} constant parameter for calculation was assumed as 550 m/s. Estimators used in each GMM for PGV and CAV are presented in Tables 4.16-4.17.

Table 4.16. Summary of estimator parameters used in GMMs for CAV.

Estimators	CAV proposed model	TR-CB19	TR-SA17
M_w	✓	✓	✓
Vs_{30}	✓	✓	✓
$Z_{1.0}$		✓	
$Z_{2.5}$		✓	
Z_{hyp}		✓	
Z_{TOR}		✓	
R_{JB}	✓	✓	✓
R_{rup}			
R_{epi}			
R_x		✓	
HW			
SoF	✓	✓	✓
δ		✓	
λ		✓	

Table 4.17. Summary of estimator parameters used in GMMs for PGV.

Estimators	PGV proposed model	BSSA14	CY14	KMHCAL23
M_w	✓	✓	✓	✓
Vs_{30}	✓	✓	✓	✓
$Z_{1.0}$		✓	✓	
Z_{TOR}			✓	
R_{JB}	✓	✓	✓	✓
R_{rup}	✓		✓	
R_{epi}	✓			
R_x			✓	
HW			✓	
SoF	✓	✓		✓
δ	✓		✓	
λ			✓	
FD				✓

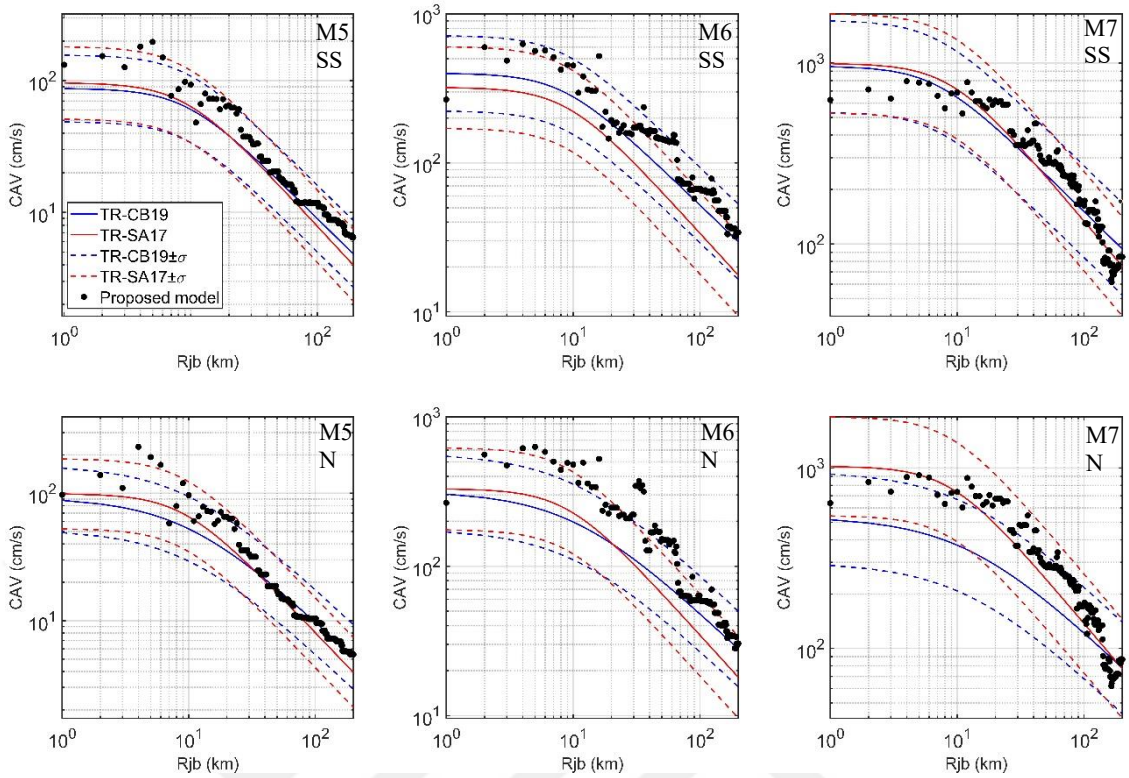


Figure 4.16. CAVs predicted with the TR-CB19, TR-SA17, and proposed model for SoF= SS (δ=90° and λ=0°) and N (δ=60° and λ=-90°) with R_{JB}=[1:200], M_w=[5.0, 6.0, 7.0], and Vs_{30}=[550 m/s].

The CAV proposed model produces fairly good results even if CAV values are roughly estimated in the ML-based method just using four estimator parameters: M_w, Vs_{30}, R_{JB}, and SoF (seen in Table 4.16). Just a small number of predicted CAV values in the near-field exceed the standard deviation limits of the TR-CB19 and TR-SA17, implying that, with the exception of near-field (R_{JB}<10 km) recordings, practically all CAV predicted with the regional proposed model fall within the standard deviation limits of the TR-CB19 and TR-SA17 GMMs shown in Figure 4.16.

Notably, the CAV proposed model predicts higher CAV than the TR-CB19 and TR-SA17 GMMs in the near-field for M_w5-6 earthquakes as shown in Figure 4.16. The illustration from Figure 4.17 represents the distribution of the expanded N-TSMD recordings in the near-field (R_{JB}<10 km). It can be seen that there are 111 strike-slip, 42 normal, and 4 reverse faults among 17,856 recordings in the near-field. The majority of the near-fault

recordings fall into the ($3.5 \leq M_w \leq 5.0$) range. There are only 9 recordings in the $5.0 \leq M_w \leq 6.0$ ranges. Therefore, ML algorithms were not well-trained due to the lack of near-field data for CAV. The CAV of the GB model for $M_w \leq 6$ earthquakes at $R_{JB} > 8$ km follows a similar trend of median estimates of TR-CB19 and TR-SA17 GMMs. Although, there are just 24 recordings of ($M_w > 7.0$) earthquakes, including the recordings of the 1999 Düzce (M_w7.1), the 1999 Kocaeli (M_w7.6), and the Kahramanmaraş (M_w7.8) events, the proposed model for CAV follows a similar trend of median estimates of especially TR-SA17 model in both near-field and far-field, as seen Figure 4.16.

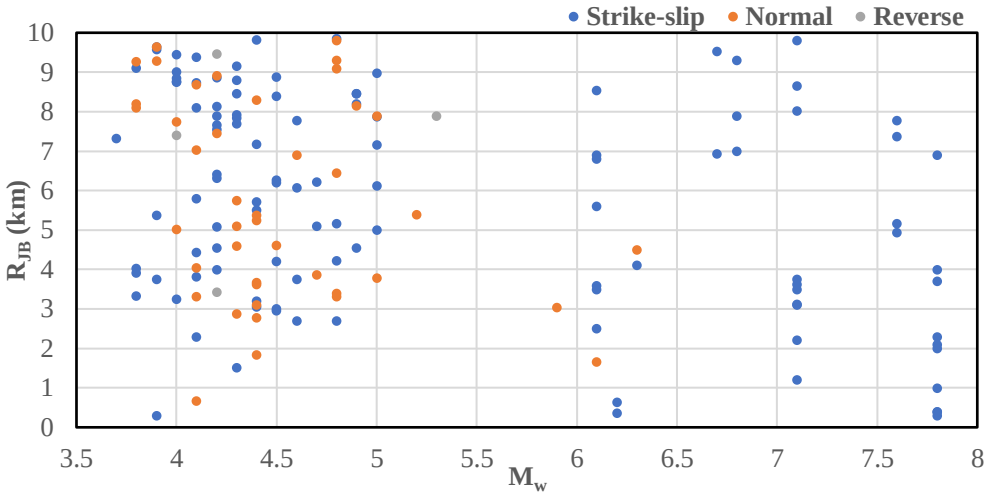


Figure 4.17. Distribution of the near-field earthquakes (10 km<R_{JB}) in the expanded N-TSMD.

On the other hand, PGV predictions were calculated for six scenario cases with the CY14, BSSA14, KMHCAL23 and proposed model. Those scenarios are the combination of M_w5.0, 6.0, 7.0, N and SS earthquakes predicted at R_{JB}=1:200 distances at a site with Vs_{30}=550 m/s. Estimator parameters used in each GMM for PGV are presented in Table 4.17. The Turkiye-specific PGV proposed model, obtained by training seven estimators, is compared with CY14, BSSA14, and KMHCAL23 models, as seen in Figure 4.18.

Nearly all PGVs predicted with the proposed model are within the ±1 standard deviation (σ) boundary of the CY14 and BSSA14 models for M_w5-6 earthquakes, while an underestimation is observed in the far-field for M_w7 earthquakes for both SS and N. It might

be attributed that the majority of far-field recordings are in the range of $3.5 \leq M_w \leq 6.0$. Such a fast attenuation of PGV is also observable both at the proposed model and KMHCAL23 for the M_w7, SS events in the far-field. Near-field (up to 10 km) PGV prediction with the proposed model for M_w7 SS earthquake is in line with estimation of other models but, not for N events. Such result could be mainly due to the fact that recordings of near-field N earthquakes are about three times less than those for SS earthquakes in the database. In the same graph, PGV values (M_w7 N) predicted by KMHCAL23 for the first 10 km are in between median prediction of BSSA14 and the +1σ prediction of CY14. Such a result also depicts complementary effect of the simulated dataset on the near-field PGV prediction for M_w7 N events.

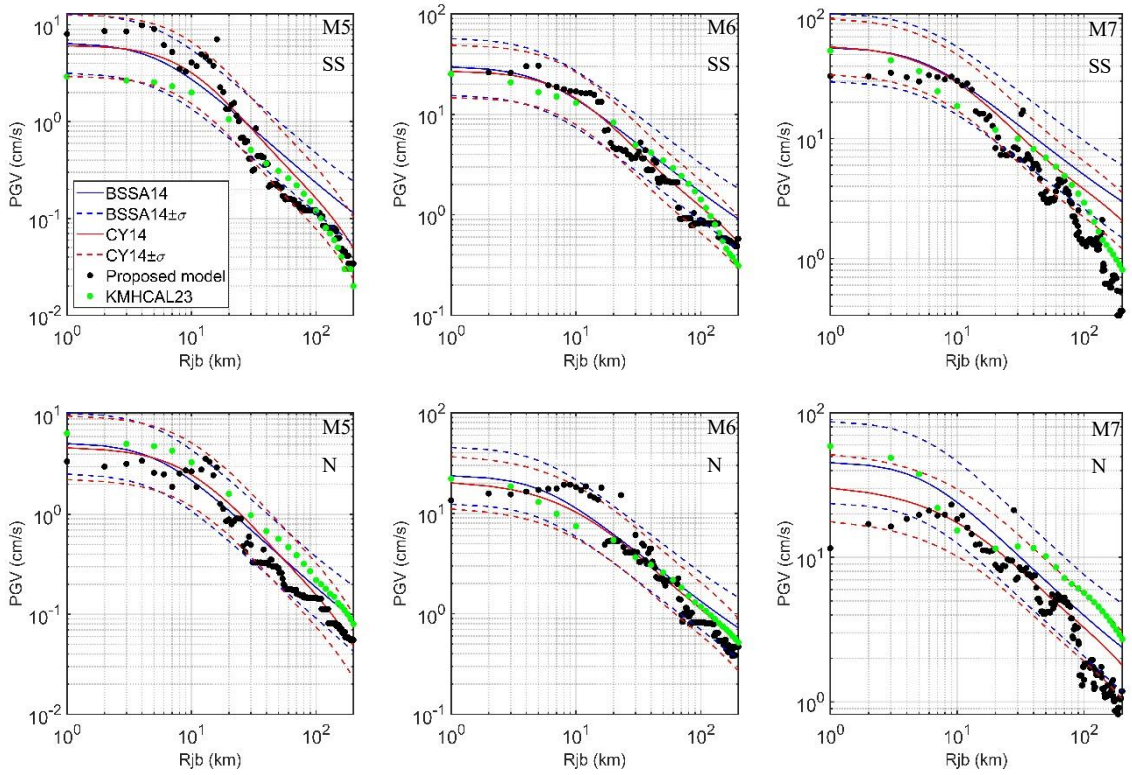


Figure 4.18. PGVs predicted with the BSSA14, CY14, KMHCAL23, and proposed model for SoF= SS (δ=90° and λ=0°) and N (δ=60° and λ=-90°) with R_{JB}=[1:200], M_w=[5.0, 6.0, 7.0], Vs_{30}=[550 m/s], and FD=[7 km].

5. CONCLUSION

In this work, various ML models were systematically tested to identify the optimal model for predicting peak ground velocity (PGV) and cumulative absolute velocity (CAV) for Turkiye. To that end, well-known and widely used ML techniques, including SVM, GB, BRR, ANN, LR, and RF, were employed with the extensive dataset consisting of over 15,000 recordings of real earthquakes that occurred in Turkiye. A comparative investigation of the earthquake database and estimator parameters was done to obtain the best adaptable models for PGVs and CAVs. Seven estimator parameters, M_w, R_{JB}, R_{rup}, R_{epi}, δ, SoF, and Vs_{30}, were evaluated while building data driven GMMs. Different input sets were determined according to feature selection methods, including LASSO, Pearson, and RFE, and earthquake engineering background to see the effect of estimator parameters. In addition, the outlier elimination method and separation process were applied to the dataset, and it was examined how the dataset should be used to obtain the most accurate PGVs and CAVs. MAE, MAPE, MSE, RMSE, and R evaluation metrics were employed to assess the models. PGV and CAV prediction model were then updated with the expanded N-TSMD consisting of 17,856 recordings of 818 earthquakes occurred up to 2023 in Turkiye. Residual analyses were conducted to validate the effectiveness of the proposed GB models. The best PGV and CAV models developed in this study were compared with conventional and non-conventional models (TR-SA17, TR-CB19, BSSA14, CY14, and KMHCAL23) to see the efficiency of the models. The following items are the key contributions of this work:

- The trained algorithms with the original-separated ($3.5 \leq M_w \leq 5.4$) and ($5.5 \leq M_w \leq 7.6$) datasets give better results in CAV prediction compared to the model trained with the original-single ($3.5 \leq M_w \leq 7.6$) dataset considering overall results and analyses in this study. However, models obtained with the original single ($3.5 \leq M_w \leq 7.6$) dataset are more successful in the prediction of PGV. This conclusion is drawn based on a comprehensive analysis and comparison process carried out in this study.

- The ML models trained by the outlier-eliminated datasets perform well for the prediction of PGV and CAV according to the 30% unseen test recordings chosen within the N-TSMD. However, the models trained by the original dataset without an outlier detection process predict PGV and CAV more reliably on the ($M_w \geq 5.5$) NESS test data. Moreover, recordings of large-magnitude earthquakes are needed in the dataset, even if they are relatively few, when constructing the model to predict the PGV and CAV of large earthquakes. Therefore, the outlier-elimination method should not be applied to a dataset consisting of real earthquakes.

- The impact of feature selection on the building of robust prediction models was investigated using different feature selection methods. For this purpose, five datasets were created with seven estimator parameters, user-selected estimators, and estimators determined by applying three feature selection methods, and they were subjected to the model training and evaluation process. It has been concluded that the models trained using user-selected parameters (i.e., M_w, Vs_{30}, R_{JB}, and SoF) give the best test result in the CAV prediction compared to the ML algorithms trained with other feature subset data. The model trained with all parameters (i.e., M_w, Vs_{30}, δ, R_{JB}, R_{rup}, R_{epi}, and SoF), however, shows superior success in the prediction of PGV.

- GB and later RF approaches demonstrate the best performance in the learning model and the prediction of PGV and CAV. These are followed by ANN and SVM, respectively. In other words, the GB model is found to be superior to other models in the prediction of PGV and CAV. Conversely, the models developed with BRR and LR algorithms show less success in predicting PGV and CAV compared to the other models according to performance evaluation metrics.

- The proposed models developed with the GB algorithm and the expanded N-TSMD show strong prediction skills for PGVs and CAVs compared to their peer ground motion models.

- These models are made available to the user on the Internet platform, as seen in Addenda A and B.

For the further extension of this work, novel ML and deep learning algorithms can be developed for the prediction of strong ground motion parameters to get more compatible and enhanced results. GMMs can be improved for both PGV and CAV predictions by developing especially boosting algorithms. Furthermore, the correlation between PGV/CAV and spectral accelerations is of importance in ground motion selection and vector-valued probabilistic seismic hazard assessment. It would be useful to derive correlations using the expanded N-TSMD and the proposed model and compare the results with the existing correlation models.

REFERENCES

1. Wu, M. H., J. P. Wang, and P. E. Chiang, "Cumulative absolute velocity (CAV) seismic hazard assessment for Taiwan", *Journal of Earthquake Engineering*, Vol. 26, No. 7, pp. 3440–3460, 2020.

2. Omine, H., T. Hayashi, H. Yashiro, and I. Fukushima, "Seismic risk analysis method using both PGA and PGV", *The 14 World Conference on Earthquake Engineering*, Beijing/China, 2008.

3. Bray, J. D. and J. Macedo, "6th Ishihara lecture: Simplified procedure for estimating liquefaction-induced building settlement", *Soil Dynamics and Earthquake Engineering*, Vol. 102, pp. 215–231, 2017.

4. Zhang, X., L. Tang, X. Ling, A. H. C. Chan, and J. Lu, "Using peak ground velocity to characterize the response of soil-pile system in liquefying ground", *Eng Geol*, Vol. 240, pp. 62–73, 2018.

5. Orense, R. P., "Assessment of liquefaction potential based on peak ground motion parameters", *Soil Dynamics and Earthquake Engineering*, Vol. 25, No. 3, pp. 225–240, 2005.

6. Kostadinov, M. V and I. Towhata, "Assessment of liquefaction-inducing peak ground velocity and frequency of horizontal ground shaking at onset of phenomenon".

7. O'Rourke, T. D., H. E. Stewart, and S.-S. Jeon, "Geotechnical aspects of lifeline engineering", *Proceedings of the Institution of Civil Engineers-Geotechnical Engineering*, Vol. 149, No. 1, pp. 13–26, 2001.

8. Davis, C. A. and J. P. Bardet, "Responses of buried corrugated metal pipes to earthquakes", *Journal of Geotechnical and Geoenvironmental Engineering*, Vol. 126, No. 1, pp. 28–39, 2000.

9. Razin, T., K. Khatimah, Y. Annisa, A. Hamzah, and M. F. I. Massinai, "Peak ground acceleration (PGA) and peak ground velocity (PGV) analyze for microzonation of

earthquake hazard area: Case study in West Nusa Tenggara", *The 3rd Southeast Asian Conference on Geophysics*, 2021.

10. Alcik, H., O. Ozel, N. Apaydin, and M. Erdik, "A study on warning algorithms for Istanbul earthquake early warning system", *Geophys Res Lett*, Vol. 36, No. 5, 2009.

11. Fahjan, Y. M., H. Alcik, and A. Sari, "Applications of cumulative absolute velocity to urban earthquake early warning systems", *J Seismol*, Vol. 15, No. 2, pp. 355–373, 2011.

12. Fang, R., J. Zheng, J. Geng, Y. Shu, C. Shi, and J. Liu, "Earthquake magnitude scaling using peak ground velocity derived from high-rate gnss observations", *Seismological Research Letters*, Vol. 92, No. 1, pp. 227–237, 2020.

13. Picozzi, M., A. Emolo, C. Martino, A. Zollo, N. Miranda, G. Verderame, T. Boxberger, and the REAKT Working Group, "Earthquake early warning system for schools: A feasibility study in southern Italy", *Seismological Research Letters*, Vol. 86, No. 2A, pp. 398–412, 2015.

14. Carranza, M., E. Buforn, S. Colombelli, and A. Zollo, "Earthquake early warning for southern Iberia: A P wave threshold-based approach", *Geophys Res Lett*, Vol. 40, No. 17, pp. 4588–4593, 2013.

15. Colombelli, S., A. Caruso, A. Zollo, G. Festa, and H. Kanamori, "A P wave-based, on-site method for earthquake early warning", *Geophys Res Lett*, Vol. 42, No. 5, pp. 1390–1398, 2015.

16. Atkinson, G. M. and S. L. I. Kaka, "Relationships between felt intensity and instrumental ground motion in the Central United States and California", *Bulletin of the Seismological Society of America*, Vol. 97, No. 2, pp. 497–510, 2007.

17. Muin, S. and K. M. Mosalam, "Cumulative absolute velocity as a local damage indicator of instrumented structures", *Earthquake Spectra*, Vol. 33, No. 2, pp. 641–664, 2017.

18. Kramer, S. L. and R. A. Mitchell, "Ground motion intensity measures for liquefaction hazard evaluation", *Earthquake Spectra*, Vol. 22, No. 2, pp. 413–438, 2006.

19. Reed, J. W. and R. P. Kassawara, "A criterion for determining exceedance of the operating basis earthquake", *Nuclear Engineering and Design*, Vol. 123, pp. 387–396, 1990.

20. Rayes, A., J. Adinata, and M. Taiebat, "Liquefaction hazard evaluation under bidirectional seismic shearing: Optimal ground motion intensity measures", *Earthquake geotechnical engineering for protection and development of environment and constructions: proceedings of the 7th International Conference on Earthquake Geotechnical Engineering (ICEGE 2019)*, Rome/Italy, 2019.

21. Campbell, K. W. and Y. Bozorgnia, "Cumulative absolute velocity (CAV) and seismic intensity based on the PEER-NGA database", *Earthquake Spectra*, Vol. 28, No. 2, pp. 457–485, 2012.

22. Arias, A., "A measure of earthquake intensity", *Cambridge, Massachusetts*, pp. 438–483, 1970.

23. Zhang, B., Y. Yu, X. Li, and Y. Wang, "Ground motion prediction equation for the average horizontal component of PGA, PGV, and 5% damped acceleration response spectra at periods ranging from 0.033 to 8.0s in southwest China", *Soil Dynamics and Earthquake Engineering*, Vol. 159, 2022.

24. Montalva, G. A., N. Bastías, and F. Leyton, "Strong ground motion prediction model for PGV and spectral velocity for the Chilean subduction zone", *Bulletin of the Seismological Society of America*, Vol. 112, No. 1, pp. 348–360, 2022.

25. Bindi, D., M. Massa, L. Luzi, G. Ameri, F. Pacor, R. Puglia, and P. Augliera, "Pan-European ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods up to 3.0 s using the RESORCE dataset", *Bulletin of Earthquake Engineering*, Vol. 12, No. 1, pp. 391–430, 2014.

26. Boore, D. M., J. P. Stewart, E. Seyhan, and G. M. Atkinson, "NGA-West 2 equations for predicting PGA, PGV, and 5%-damped PSA for shallow crustal earthquakes", *Earthquake Spectra*, Vol. 30, No. 3, pp. 1057–1085, 2014.

27. Chiou, B. S. J. and R. R. Youngs, “Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra”, *Earthquake Spectra*, Vol. 30, No. 3, pp. 1117–1153, 2014.

28. Boore, D. M. and G. M. Atkinson, “Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s”, *Earthquake Spectra*, Vol. 24, No. 1, pp. 99–138, 2008.

29. Akkar, S. and J. J. Bommer, “Empirical prediction equations for peak ground velocity derived from strong-motion records from Europe and the Middle East”, *Bulletin of the Seismological Society of America*, Vol. 97, No. 2, pp. 511–530, 2007.

30. Bullock, Z., A. B. Liel, S. Dashti, and K. A. Porter, “A suite of ground motion prediction equations for cumulative absolute velocity in shallow crustal earthquakes including epistemic uncertainty”, *Earthquake Spectra*, Vol. 37, No. 2, pp. 937–958, 2021.

31. Campbell, K. W. and Y. Bozorgnia, “A ground motion prediction equation for the horizontal component of cumulative absolute velocity (CAV) based on the PEER-NGA strong motion database”, *Earthquake Spectra*, Vol. 26, No. 3, pp. 635–650, 2010.

32. Campbell, K. W. and Y. Bozorgnia, “Campbell-Bozorgnia NGA-West2 horizontal ground motion model for active tectonic domains”, *NCEE 2014-10th U.S. National Conference on Earthquake Engineering: Frontiers of Earthquake Engineering*, 2014.

33. Campbell, K. W. and Y. Bozorgnia, “Ground motion models for the horizontal components of Arias intensity (AI) and cumulative absolute velocity (CAV) using the NGA-West2 database”, *Earthquake Spectra*, Vol. 35, No. 3, pp. 1289–1310, 2019.

34. Danciu, L. and G.-A. Tselentis, “Engineering ground-motion parameters attenuation relationships for Greece”, *Bulletin of the Seismological Society of America*, Vol. 97, pp. 162–183, 2007.

35. Du, W. and G. Wang, “A simple ground-motion prediction model for cumulative absolute velocity and model validation”, *Earthq Eng Struct Dyn*, Vol. 42, No. 8, pp. 1189–1202, 2013.

36. Foulser-Piggott, R. and K. Goda, “New prediction equations of Arias intensity and cumulative absolute velocity for Japanese earthquakes”, *2nd ECEES*, p. 1277–1287, Istanbul/Turkey.

37. Foulser-Piggott R. and K. Goda, “Ground-motion prediction models for Arias intensity and cumulative absolute velocity for Japanese earthquakes considering single- station sigma and within-event spatial correlation”, *Bulletin of the Seismological Society of America*, Vol. 105, No. 4, pp. 1903–1918.

38. Sandıkkaya, M. A. and S. Akkar, “Cumulative absolute velocity, Arias intensity and significant duration predictive models from a pan-European strong-motion dataset”, *Bulletin of Earthquake Engineering*, Vol. 15, No. 5, pp. 1881–1898, 2017.

39. Tetik T., B. Akbaş, G. Tanırcan, and Z. Gülerce, “Performance evaluation of cumulative absolute velocity (CAV) prediction models for the New Turkish Strong Motion Database”, *3rd European Conference on Earthquake Engıneerıng & Seısmology*, Romania, p. 3674–3683, 2022.

40. Ancheta, T. D., R. B. Darragh, J. P. Stewart, E. Seyhan, W. J. Silva, B. S.-J. Chiou, K. E. Wooddell, R. W. Graves, A. R. Kottke, D. M. Boore, T. Kishida, and J. L. Donahue, “NGA-West2 database”, *Earthquake Spectra*, Vol. 30, No. 3, pp. 989–1005, 2014.

41. Chiou, B. S. J. and R. R. Youngs, “An NGA model for the average horizontal component of peak ground motion and response spectra”, *Earthquake Spectra*, Vol. 24, No. 1, pp. 173–215, 2008.

42. Akkar, S., M. A. Sandıkkaya, M. Şenyurt, A. Azari Sisi, B. Ö. Ay, P. Traversa, J. Douglas, F. Cotton, L. Luzi, B. Hernandez, and S. Godey, “Reference database for seismic ground-motion in Europe (RESORCE)”, *Bulletin of Earthquake Engineering*, Vol. 12, No. 1, pp. 311–339, 2014.

43. Tanırcan, G., Z. Gülerce, T. Tetik, B. Akbaş, “Adapting a Global Fourier Amplitude Spectrum Estimation Model to the Earthquake Database of Turkiye”, TUBİTAK 1002 Project Report No: 120M799, 2022 (in Turkish).

44. Hu, J., C. Jin, H. Zhang, L. Hu, and Z. Wang, “Support vector regression for developing ground-motion models for Arias intensity, cumulative absolute velocity,

and significant duration for the Kanto region, Japan", *Seismological Research Letters*, Vol. 93, No. 3, pp. 1619–1635, 2022.

45. Ji, D., J. Liu, W. Wen, C. Zhai, W. Wang, and E. I. Katsanos, "Prediction of cumulative absolute velocity based on refined second-order deep neural network", *Journal of Earthquake Engineering*, Vol. 26, No. 15, pp. 8021–8040, 2021.

46. Junjie, W., Y. Maruyama, and W. Liu, "Construction of attenuation relationship of peak ground velocity using machine learning and examination of station correction factor", 2021.

47. Khosravikia, F., Y. Zeinali, Z. Nagy, P. Clayton, and E. Rathje, "Neural network-based equations for predicting PGA and PGV in Texas, Oklahoma, and Kansas", 2018.

48. Derakhshani, A. and A. H. Foruzan, "Predicting the principal strong ground motion parameters: A deep learning approach", *Applied Soft Computing Journal*, Vol. 80, pp. 192–201, 2019.

49. Dhanya, J. and S. T. G. Raghukanth, "Ground motion prediction model using artificial neural network", *Pure Appl Geophys*, Vol. 175, No. 3, pp. 1035–1064, 2018.

50. Thomas, S., G. N. Pillai, K. Pal, and P. Jagtap, "Prediction of ground motion parameters using randomized ANFIS (RANFIS)", *Applied Soft Computing Journal*, Vol. 40, pp. 624–634, 2016.

51. Derras, B., P. Y. Bard, and F. Cotton, "Towards fully data driven ground-motion prediction models for Europe", *Bulletin of Earthquake Engineering*, Vol. 12, No. 1, pp. 495–516, 2014.

52. Alavi, A. H. and A. H. Gandomi, "Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing", *Comput Struct*, Vol. 89, No. 23–24, pp. 2176–2194, 2011.

53. Ahmad, I., M. H. el Naggar, and A. N. Khan, "Neural network based attenuation of strong motion peaks in Europe", *Journal of Earthquake Engineering*, Vol. 12, No. 5, pp. 663–680, 2008.

54. Liu, B.-Y., L.-Y. Ye, M.-L. Xiao, and S. Miao, “LNCS 4233-Peak ground velocity evaluation by artificial neural network for West America Region”, *LNCS*, Berlin, Heidelberg, p. 942–951, 2006.

55. Karimzadeh, S., A. Mohammadi, U. Salahuddin, A. Carvalho, and P. B. Lourenço, “Backbone ground motion model through simulated records and XGBoost machine learning algorithm: An application for the Azores plateau (Portugal)”, *Earthq Eng Struct Dyn*, 2023.

56. Mohammadi, A., S. Karimzadeh, S. A. Banimahd, V. Ozsarac, and P. B. Lourenço, “The potential of region-specific machine-learning-based ground motion models: Application to Turkey”, *Soil Dynamics and Earthquake Engineering*, Vol. 172, 2023.

57. Karimzadeh, S., A. Mohammadi, S. M. S. Hussaini, D. Caicedo, A. Askan, and P. B. Lourenço, “ANN-based ground motion model for Turkey using stochastic simulation of earthquakes”, *Geophys J Int*, Vol. 236, No. 1, pp. 413–429, 2023.

58. Li, C., D. Ji, C. Zhai, Y. Ma, and L. Xie, “Vertical ground motion model for the NGA-West2 database using deep learning method”, *Soil Dynamics and Earthquake Engineering*, Vol. 165, 2023.

59. Seo, H., J. Kim, and B. Kim, “Machine-learning-based surface ground-motion prediction models for South Korea with low-to-moderate seismicity”, *Bulletin of the Seismological Society of America*, Vol. 112, No. 3, pp. 1549–1564, 2022.

60. Qi, W., R. Sun, and J. Qi, “PGA prediction by three machine learning models”, *Proceedings of the Third European Conference on Earthquake Engineering and Seismology – 3ECEES*, Romania, p. 186–191, 2022.

61. Khosravikia, F. and P. Clayton, “Machine learning in ground motion prediction”, *Comput Geosci*, Vol. 148, 2021.

62. Ji, D., C. Li, C. Zhai, Y. Dong, E. I. Katsanos, and W. Wang, “Prediction of ground-motion parameters for the NGA-west2 database using refined second-order deep neural networks”, *Bulletin of the Seismological Society of America*, Vol. 111, No. 6, pp. 3278–3296, 2021.

63. Ding, Y., J. Chen, and J. Shen, “Prediction of spectral accelerations of aftershock ground motion with deep learning method”, *Soil Dynamics and Earthquake Engineering*, Vol. 150, 2021.

64. Khosravikia, F., P. Clayton, and Z. Nagy, “Artificial neural network-based framework for developing ground-motion models for natural and induced earthquakes in Oklahoma, Kansas, and Texas”, *Seismological Research Letters*, Vol. 90, No. 2 A, pp. 604–613, 2019.

65. Öztürk, S., *Prediction of felt intensity from ground motion parameters using artificial neural network method a thesis submitted to the graduate school of natural and applied sciences of Middle East Technical University*, M.S. Thesis, Middle East Technical University, 2019.

66. Derras, B., P. Y. Bard, F. Cotton, and A. Bekkouche, “Adapting the neural network approach to PGA prediction: An example based on the KiK-net data”, *Bulletin of the Seismological Society of America*, Vol. 102, No. 4, pp. 1446–1461, 2012.

67. Tezcan, J. and Q. Cheng, “Support vector regression for estimating earthquake response spectra”, *Bulletin of Earthquake Engineering*, Vol. 10, No. 4, pp. 1205–1219, 2012.

68. Günaydın, K. and A. Günaydın, “Peak ground acceleration prediction by artificial neural networks for northwestern Turkey”, *Math Probl Eng*, Vol. 2008, 2008.

69. Güllü, H. and E. Erçelebi, “A neural network approach for attenuation relationships: An application using strong ground motion data from Turkey”, *Eng Geol*, Vol. 93, No. 3–4, pp. 65–81, 2007.

70. Kerh, T. and S. B. Ting, “Neural network estimation of ground peak acceleration at stations along Taiwan high-speed rail system”, *Eng Appl Artif Intell*, Vol. 18, No. 7, pp. 857–866, 2005.

71. Douglas, J., “What is a poor quality strong-motion record?”, *Bulletin of Earthquake Engineering*, Vol. 1, pp. 141–156, 2003.

72. Sandıkkaya, M. A., B. Güryuva, Ö. Kale, O. Okçu, A. İçen, E. Yenier, and S. Akkar, "An updated strong-motion database of Türkiye (SMD-TR)", *Earthquake Spectra*, 2023.

73. Sgobba, S., F. Pacor, C. Felicetta, G. Lanzano, M. C. D'Amico, E. Russo, and L. Luzi, "NEar-Source Strong-motion flatfile (NESS), version 2.0 [Data set]", *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*. 2021.

74. Emre, Ö., T. Y. Duman, S. Özalp, F. Şaroğlu, Ş. Olgun, H. Elmacı, and T. Çan, "Active fault database of Turkey", *Bulletin of Earthquake Engineering*, Vol. 16, No. 8, pp. 3229–3275, 2018.

75. Han, J. and M. Kamber, *Data mining: concepts and techniques*, Vol. 31. Morgan Kaufmann Publishers, 2001.

76. Wang, P., Q. Zhao, Y. Huang, J. Nie, X. Man, and Z. Mingxue, "Comprehensive evaluation of road risk in earthquake disaster area based on normalization method", *International Conference on Computer, Information and Telecommunication Systems (CITS)*, p. 1–5, 2020.

77. Potdar, K., T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for Neural Network Classifiers", *Int J Comput Appl*, Vol. 175, pp. 7–9, 2017.

78. Tabatabaei, M., R. Kimiaefar, A. Hajian, and A. Akbari, "Robust outlier detection in geo-spatial data based on LOLIMOT and KNN search", *Earth Sci Inform*, Vol. 14, pp. 1065–1072, 2021.

79. Guyon, I., J. Weston, and S. Barnhill, "Gene Selection for Cancer Classification using Support Vector Machines", 2002.

80. Tibshirani, R., "Regression Shrinkage and Selection Via the Lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.

81. Maulud, D. and A. M. Abdulazeez, “A review on linear regression comprehensive in machine learning”, *Journal of Applied Science and Technology Trends*, Vol. 1, No. 4, pp. 140–147, 2020.

82. Liu, Y., Y. Wang, and J. Zhang, “LNCS 7473-new machine learning algorithm: Random Forest”, *LNCS*, p. 246–252, 2012.

83. Breiman, L., “Random Forests”, *Machine Learning*, Vol. 45. pp. 5–32, 2001.

84. Segal, M. R., “Machine learning benchmarks and Random Forest regression”, UCSF: Center for Bioinformatics and Molecular Biostatistics, 2023.

85. Chang, C.-C. and C.-J. Lin, “LIBSVM: A library for Support Vector Machines”, *ACM Trans Intell Syst Technol*, Vol. 2, No. 3, pp. 1–27, 2001.

86. Flavin, T., T. Steiner, B. Mitra, and V. nagaraju, “Bayesian Ridge regression based model to predict fault location in HVdc network”, 2022.

87. Gardner, M. W. and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences”, *Atmos Environ*, Vol. 32, No. 14, pp. 2627–2636, 1998.

88. Zhang, Y. and A. Haghani, “A gradient boosting method to improve travel time prediction”, *Transp Res Part C Emerg Technol*, Vol. 58, pp. 308–324, 2015.

89. Wang, M.-X., D. Huang, G. Wang, and D.-Q. Li, “SS-XGBoost: a machine learning framework for predicting Newmark sliding displacements of slopes”, *Journal of Geotechnical and Geoenvironmental Engineering*, Vol. 146, No. 9, 2020.

90. Abrahamson, N. A. and R. R. Youngs, “A stable algorithm for regression Analyses using the random effects model”, 1992.

APPENDIX A: THE DEVELOPMENT OF THE WEBSITE FOR THE APPLICATION OF THE PGV PROPOSED MODEL

The best prediction model obtained as a result of this study, the Turkiye-specific GB ground motion model for peak ground velocity (PGV), is presented to the user as seen in Figure A.1. The software tool can be accessed by clicking on the link: https://ktp24-turkiyespecificgmmforpgv.streamlit.app/.

The prediction of the PGV with the proposed model can be easily performed by defining the M_w, R_{JB}, R_{rup}, R_{epi}, δ, SoF, and Vs_{30} estimator parameters within applicable ranges.



Figure A.1. The software interface of Turkiye-specific ground motion model for peak ground velocity.

APPENDIX B: THE DEVELOPMENT OF THE WEBSITE FOR THE APPLICATION OF THE CAV PROPOSED MODEL

The best prediction model obtained as a result of this study, the Turkiye-specific GB ground motion model for cumulative absolute velocity (CAV), is presented to the user as seen in Figure B.1. The software tool can be accessed by clicking on the link: https://ktp23updated-turkiyespecificgmmforcav.streamlit.app/.

The prediction of the CAV with the proposed model can be easily performed by defining the M_w, R_{JB}, SoF, and Vs_{30} estimator parameters within applicable ranges.

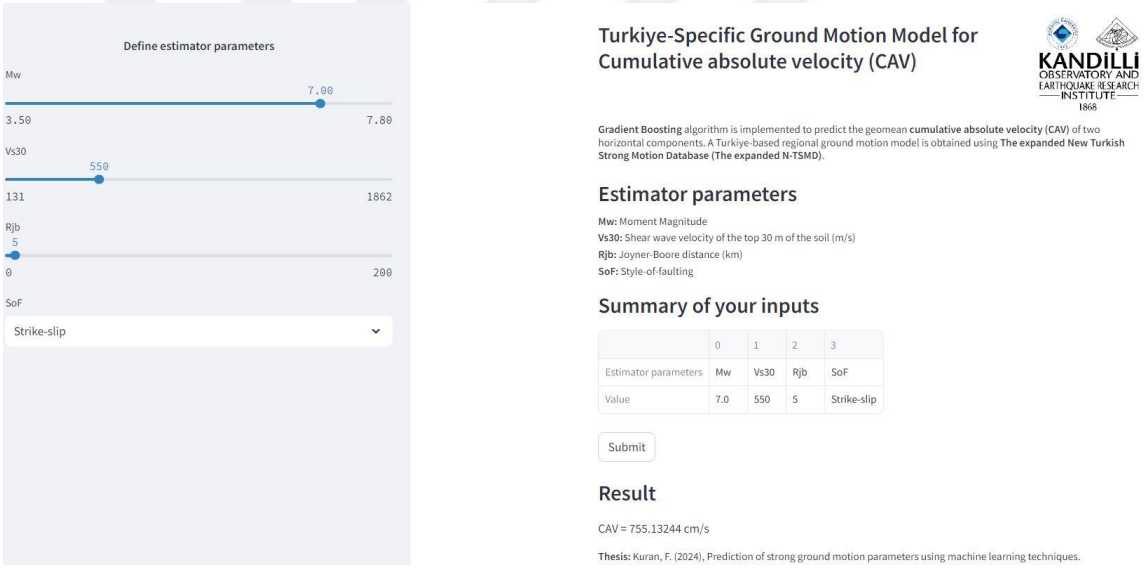


Figure B.1. The software interface of Turkiye-specific ground motion model for cumulative absolute velocity.