Sinopfbd, 2024; 9(2): 266-286 https://doi.org/10.33484/sinopfbd.1460421

Sinop Üniversitesi Fen Bilimleri Dergisi
Sinop Uni J Nat Sci

E-ISSN: 2564-7873

https://dergipark.org.tr/tr/pub/sinopfbd

Regional Analysis of Earthquakes and Earthquake Magnitude Estimation with Machine Learning Techniques

Gül Cihan HABEK[1] and Humar KAHRAMANLI ÖRNEK[2]

How to cite: Habek, G. C., & Kahramanli Örnek, H. (2024). Regional analysis of earthquakes and earthquake magnitude estimation with machine learning techniques. *Sinop Üniversitesi Fen Bilimleri Dergisi, 9*(2), 266-286. https://doi.org/10.33484/sinopfbd.1460421

Research Article

Corresponding Author
Gül Cihan HABEK
gulhabek@kmu.edu.tr

ORCID of the Authors
G.C.H: 0000-0003-1748-3486
H.K.Ö: 0000-0003-2336-7924

Received: 28.03.2024
Accepted: 10.06.2024

Abstract

Natural disasters, which have been increasing in recent years due to the impact of climate change, pose a significant threat worldwide. Natural disasters, which can cause a large number of human losses and material damages due to their uncertain nature and sudden effects, vary depending on the location and natural environment of the countries. Türkiye located in the Alpine-Himalayan Earthquake Zone, is one of the countries most exposed to earthquake disasters. Although timely prediction of earthquakes is of vital importance in minimizing the destructive effects that may occur during the disaster and increasing resistance to the destructive effects of the disaster, it cannot yet be predicted successfully due to its non-linear chaotic behavior. However, many researchers continue to work on the subject, and earthquake prediction models are actively used in some countries where earthquake disasters occur frequently and cause great destruction. In this study, the magnitudes of future earthquakes were predicted using various machine learning models: Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Random Forests (RF), Gradient Boosting Algorithm (GB), Extreme Gradient Boosting Algorithm (XGBoost), 2-hidden-layer Artificial Neural Networks (ANN), and an ANN-KNN hybrid learning model. The performances of the established models were evaluated with MSE, MAE, RMSE, and R^2 metrics; and the ANN-KNN model showed that it was more effective than other models by exhibiting the highest performance with 0.0418 MSE, 0.0030 MAE, 0.0552 RMSE, and 0.7138 R^2 values. Additionally, unlike other studies, seven regions of Türkiye were considered separately and earthquakes were analyzed in detail according to their geography. The analysis results aim to add a new perspective to the literature.

Keywords: Earthquake prediction, machine learning, regression, earthquake analysis, regional analysis

Makine Öğrenmesi Teknikleriyle Depremlerin Bölgesel Analizi ve Deprem Büyüklüğü Tahmini

[1]Karamanoğlu Mehmetbey University, Faculty of Engineering, Institute of Science, Department of Computer Engineering, Karaman, Türkiye

Öz

İklim değişikliğinin etkisiyle son yıllarda artan doğal afetler dünya çapında önemli bir tehdit oluşturuyor. Belirsiz doğası ve ani etkileri nedeniyle çok sayıda insan kaybına ve maddi hasara neden olabilen doğal afetler, ülkelerin bulunduğu konuma ve doğal ortamlarına göre değişiklik göstermektedir. Alp-Himalaya Deprem Bölgesi'nde yer alan Türkiye, deprem felaketlerine en fazla maruz kalan ülkelerden biridir. Depremlerin

[2]Selçuk University, Faculty of Technology, Institute of Science, Department of Computer Engineering, Konya Turkiye

zamanında tahmini, afet sırasında oluşabilecek yıkıcı etkilerin en aza indirilmesi ve afetin yıkıcı etkilerine karşı direncin arttırılması açısından hayati öneme sahip olmasına rağmen, doğrusal olmayan kaotik davranışı nedeniyle henüz başarılı bir şekilde tahmin edilememektedir. Ancak pek çok araştırmacı konu üzerinde çalışmaya devam etmekte ve deprem felaketlerinin sıklıkla yaşandığı ve büyük yıkımlara neden olduğu bazı ülkelerde deprem tahmin modelleri aktif olarak kullanılmaktadır. Bu çalışmada gelecekte meydana gelebilecek depremlerin büyüklükleri Uzun Kısa Süreli Bellek (LSTM), Tekrarlayan Sinir Ağı (RNN), Rastgele Ormanlar (RF) ve Gradient Boosting Algoritması (GB) kullanılarak ölçülmektedir. Extreme Gradient Boosting Algoritması (XGBoost), 2 gizli katmanlı Yapay Sinir Ağları (ANN) ve ANN-KNN hibrit öğrenme modeli kullanılarak tahmin edilmeye çalışıldı. Kurulan modellerin performansları MSE, MAE, RMSE ve R^2 metrikleri ile değerlendirilmiş; ANN-KNN modeli ise 0.0418 MSE, 0.0030 MAE, 0.0552 RMSE ve 0.7138 R^2 değerleri ile en yüksek performansı sergileyerek diğer modellere göre daha etkili olduğunu göstermiştir. Ayrıca diğer çalışmalardan farklı olarak Türkiye'nin yedi bölgesi ayrı ayrı ele alınmış ve depremler coğrafyalarına göre detaylı bir şekilde analiz edilmiştir. Elde edilen analiz sonuçlarının literatüre yeni bir bakış açısı kazandırması amaçlanmaktadır.

Anahtar Kelimeler: Deprem tahmini, makine öğrenmesi, regresyon, deprem analizi, bölgesel analiz

Introduction

Natural disasters, which have increased in frequency due to climate changes in recent years, pose a significant threat in Türkiye and the world; they cause inevitable losses and destruction due to their unpredictable nature and sudden effects. According to the definition of Disaster made by the Disaster and Emergency Management Presidency, natural disasters; are events that may arise from various reasons such as nature, technological events, or human factors, affecting a part or all of the society, causing physical, economic and social losses, disrupting normal life or stopping it completely [1]. EM-DAT (Emergency Events Database), an international disaster database, examines disasters by dividing them into two groups: natural and technological [2]. When we look at the Türkiye data of the EM-DAT database, 216 of the 389 disasters that occurred between 1900 and 2023 occurred due to nature and 173 due to technological reasons. As seen in Figure 1, the most common disaster type among natural disasters is 53%. There have been earthquakes.

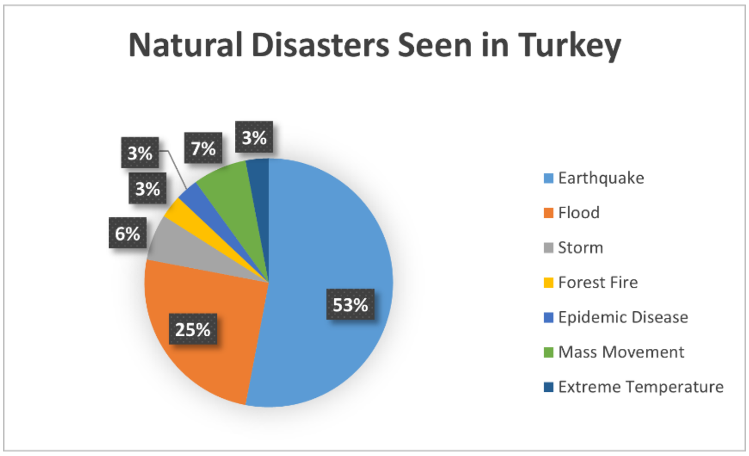


Figure 1. *Distribution of natural disasters that occurred between 1900-2023*

The types and frequencies of natural disasters vary depending on the geographical location and natural environment of countries. Floods, which can be seen all over the world and can easily turn into disasters due to different reasons, are seen in China due to extreme rainstorms, Volcanic Eruptions, and Earthquakes frequently occur in Pacific Ring of Fire countries such as Japan, Indonesia, the Philippines, and New Zealand [3, 4]. Due to frequent and destructive earthquakes, a large number of human and property losses occur in Türkiye, which is located in the Alpine-Himalayan Earthquake Zone. [5]. When we look at the natural disaster statistics of the last 20 years in our country, which is a 1st-degree earthquake zone, it is seen that a total of 50,999 people lost their lives and 2,827,859 dollars of damage were caused in the earthquakes that occurred [2]. Being able to predict earthquakes that societies are likely to encounter in the future may enable them to be prepared for disasters and develop coping strategies [6]. Timely and acceptable predictions are vital to be able to take strategic measures to determine and minimize the devastating effects that may occur during a disaster and to increase resilience to disasters [7]. Although no successful results regarding earthquake prediction can be achieved due to the non-linear chaotic behavior of the earthquake, studies continue to be carried out and discussed by many researchers [8]. When we look at the studies in the literature on the prediction of earthquakes, Çam and Duman [9] study four different regions with intense seismic activity in the west of Türkiye: Gölhisar Çameli Region, Burdur Fault Region, Büyük and Küçük Menderes Graben Region, and Gediz and Alaşehir Grabens. They aimed to predict future earthquakes. In this direction, he developed a feed-forward back-propagation ANN using the "b" value used in earthquake predictions based on the Gutenberg-Richter relationship; They trained the developed network using earthquake data. While the trained network gave successful results in predicting earthquakes that will not occur for 4 regions, it gave unsuccessful results in predicting earthquakes that will occur. In another study, Mallouhy et al. [10] aimed to predict earthquakes using a single time series data set obtained from an earthquake data center in Northern California. Each data point in the considered data set represents hourly average readings taken between 1967 and 2003 and only data from a single time series of a particular earthquake center were used in each forecast. In the analysis, earthquakes were classified as negative magnitude and positive magnitude earthquakes. Aftermath earthquakes that occurred in the same region after major earthquakes and were below 5 on the Rictor scale were ignored, and earthquakes that exceeded 5 on the scale were taken into account. This study aims to classify earthquakes as negative magnitude, which are less than 4, and are followed by at least 20 earthquakes that are not 0 in the last 512 hours~~;~~. Major earthquakes that occur 512 hours or more after another major earthquake are classified as positive major earthquakes. Random Forests (RF), Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Multilayer Perceptron, When the classification accuracy results obtained from the models established with eight machine learning algorithms, including AdaBoost and Classification and Regression Trees (CART), were compared, the highest result was obtained as 76.97% with the RF algorithm. Wang et al. [11] used machine learning techniques in their study to investigate whether earthquakes with magnitudes less than 4.0 can be used to predict earthquakes with magnitudes greater than 6.0. The data set used in the study consists of earthquakes greater than 3.0 that occurred in the Sichuan-Yunnan region between 1970 and 2021, taken from CEDC (China Earthquake Data Center). During the testing phase, the Chuandian region of Southwest China was preferred because there were many earthquakes in the region. There are two questions that the study focuses on. The first of these is whether an earthquake greater than 6.0 will occur next year, and four types of traditional machine learning algorithms were used to classify this: RF, Decision Tree (DT), SVM, and LR. In this part, the highest success was achieved at 97.5% with the RF algorithm. The second question that the study focuses on is what the maximum magnitude of the earthquake that will occur next year will be, and the LSTM network was used to predict this. At the end of the study, it was concluded that the network used showed sufficient success in predicting the earthquake magnitude and that small earthquakes can be used to predict large earthquakes. Demirelli et al. [12] aimed to predict earthquakes by combining geological and geodetic data in their studies. The data set consists of earthquakes that occurred between 1970 and 2021, recorded by the Kandilli Observatory and Earthquake Research Institute (KRDAE). Data on fault lines were taken from the Mineral Research and Exploration (MTA) and Active Fault Database of Türkiye [13]. When the prediction models created using RF, XGBoost, DT, and KNN regression algorithms were evaluated using MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) performance metrics, the highest success was achieved when the KNN algorithm was used. In the last study discussed, Karcı and Şahin [14] developed two types of prediction models: earthquake magnitude prediction and prediction of earthquakes that will occur within a certain date range. The data set used in the study is earthquakes with a magnitude of 3.5 and above that occurred between 1970 and 2021 in the Kandilli Observatory and Earthquake Research Institute Regional Earthquake-Tsunami Monitoring and Evaluation Center (BDTİM) Earthquake Inquiry System. A deep learning architecture consisting of 3 hidden layers and 1 input layer was developed using the Keras library for earthquake magnitude estimation. For time estimation, the LSTM model consisting of 3 UKSB layers, 1 Dropout layer, and 1 output layer was designed. It has been concluded that the proposed models provide more accurate results in earthquake magnitude estimation when compared to traditional machine learning methods such as Multiple Linear Regression, Polynomial Regression, DT, and RF. In this study, a dataset was created using the main earthquakes that occurred in Türkiye between 2000 and 2023, obtained from the Kandilli Observatory and Earthquake Research Institute. The dataset was divided into geographical regions of Türkiye based on the location of occurrence, and the tables for each region were analyzed according to the total number of earthquakes and average magnitudes every month; the depth-magnitude relationship of earthquakes was examined. In the second stage, a hybrid learning model was employed, integrating LSTM, RNN, RF, GB, XGBoost, 2 hidden-layer ANN, and ANN with KNN algorithms, to attempt to predict the magnitudes of future earthquakes. The performance of the models was compared using MSE, MAE, RMSE, and R^2 evaluation metrics. In the continuation of the study, the created dataset and preprocessing steps are described in detail in the Materials and Methods section. Additionally, general information about the machine learning algorithms and evaluation metrics used in earthquake prediction is provided. The comparative results obtained from regional analysis and earthquake prediction models are discussed in the Results section. The comparison of the results with other studies and their contributions to science are discussed in detail in the Conclusion section.

Material and Method

In this section, after describing the data set created for the study in detail, general information about the machine learning algorithms and evaluation metrics used in prediction is included.

Dataset

The data set to be used in the study was created by taking the earthquakes between 2020 and 2023 in the earthquake catalog of Kandilli Observatory and Earthquake Research Institute (KRDAE) [15]. The captured data includes 94935 earthquake data that occurred at 35.00-42.00 latitude and 26.00-45.00 longitude. “earthquake_ID”, “code”, “occurrence date”, “occurrence time”, “latitude”, “longitude”, “der(km)”, “xM”, “MD”, “ML”, “Mw”, “ From the data set consisting of "Ms", "Mb", "type", "location" information, unnecessary columns were deleted and "occurrence date", "occurrence time", "latitude", "longitude", "depth (km)", "xM" and “location” information continued to work. A new column named "time" was created by combining the "occurrence date" columns, which contain the date information when the earthquake occurred, and the "occurrence time" columns, which contain the time it occurred. To be processed later, the time column was converted to the DateTime object of the pandas library and made ready for regression by converting it into a Unix timestamp, which is a numerical representation of seconds. The data includes main earthquakes with magnitudes ranging from 0 to 9 and minor earthquakes, that is, aftershocks, that occur after the main earthquakes. Aftershocks do not have a specific duration and can continue for a period ranging from 1 month to 2 years after the main earthquake [16]. Since it is thought to negatively affect learning, earthquakes that occurred in the same region within 3 months after the main earthquake and whose magnitude was smaller than the main earthquake were considered aftershocks and were removed from the data set. In addition, major earthquakes with magnitudes below 3 and considered very mild earthquakes by the USGS [17] were also removed from the data set, and the final data set was created with the remaining 7043 data, the first five examples of which are given in Table 1.

Table 1. *The first five rows in the Data Set*

Earthquake id	Time	Latitude	Longitude	Depth (km)	xm
0	1609518981	35.9273	27.8658	26.1	3.9
1	1609353912	36.4918	28.7092	5.0	3.9
2	1609335612	37.796	26.4165	14.4	4.4
3	1609196128	36.4522	26.7882	120.2	3.3
4	1609151401	35.8905	32.5063	5.6	3.3

For the models used to perform better, the values in the data set were scaled to a certain range using the MinMaxScaler class in the SciKit-Learn library; The data set was divided into 80% training and 20% test data, making it ready for the models to learn on the training data and then evaluate their generalization ability on the test data.

Machine Learning Techniques

Machine learning is a method that improves the ability of computer systems to make automatic decisions when faced with similar situations based on their previous experiences [18]. With the rapid advancement of technology, various machine learning methods have been developed and started to be widely used in classification and regression studies, thanks to the ability of computer systems to analyze and learn complex data sets [19]. In the study, earthquake magnitude predictions were made using a hybrid model created using machine learning methods from the ensemble learning category such as RF, GB, XGBoost algorithms, deep learning methods including LSTM and RNN, 2 hidden-layer ANN, and ANN with the K-nearest neighbor algorithm.

Gradient Boosting Algorithm (GB): It is an ensemble learning technique that aims to create a stronger learner by combining weak learners, that is, models with generally low accuracy [20]. The algorithm works by adding a new model that tries to correct the errors of the previous model at each step until it reaches a predetermined goal [21]. The success of the algorithm often depends on the hyperparameter settings and the type of weak learner used.

Extreme Gradient Boosting Algorithm (XGBoost): The XGBoost algorithm, published by Chen and Guestring [22], is one of the widely used applications of GB. Similar to GB, the model is trained iteratively by adding new trees, focusing on correcting errors produced by previous trees. As a base classifier, it focuses only on decision trees and thus creates a powerful classification model by controlling the structure of the trees using a special loss function and regularization terms [23].

Random Forests Algorithm (RF): The method introduced by Leo Breiman in 1997 was developed as an alternative to the Boosting method [24]. Unlike Boosting, which aims to create a strong learner by combining weak learners, it brings together different decision trees by creating random sub-feature subsets. Each tree is trained on randomly selected subsets of data points, thus adding diversity to the model. These trees are then combined to make an overall prediction. This multi-tree approach provides the advantages of being resistant to overfitting, obtaining a model with high performance and increased generalization.

Artificial Neural Network (ANN): These are mathematical models inspired by biological neural networks [25]. It is designed to consist of an input layer, an output layer, and one or more hidden layers. ANNs, which lack the correlation approach, produce the output by multiplying the input data with weights through the neurons in their layers and passing them through an activation function [26]. The disadvantage of this network is that it does not effectively handle time dependencies and historical information between input data and output data, but performs direct mapping. RNNs are used to eliminate this disadvantageous situation.

Recurrent Neural Network (RNN): RNNs are designed to handle connections in a time series and dependencies over time [26]. The output of each neuron is fed by the input from the previous steps and its previous output. During the training of RNNs, which are extremely effective in modeling complex dependencies over time, difficulties such as sudden growth (exploding gradients) or decrease (vanishing gradients) of backpropagating gradients over time are experienced [27, 28]. Such situations can limit the ability of RNNs to effectively learn long-term dependencies and negatively impact their performance. To cope with such difficulties, LSTMs are used that contain modifications designed to learn long-term dependencies more effectively.

Long Short-Term Memory (LSTM): A special type of RNN designed to cope with the challenges of RNNs. It uses memory units with input, output, and forget gates [29]. These gates control how much information the cell retains and how much it forgets. LSTMs can generally handle long-term dependencies better and learn more effectively than RNNs.

ANN-KNN Hybrid Model: The hybrid model is the integration of two models to create a more comprehensive and powerful prediction model by combining ANN's deep learning capabilities and KNN's example-based similarity measurement. The k-Nearest Neighbors (KNN) algorithm, which was introduced to the literature in the book "Pattern Classification and Scene Analysis" [30], is an easy-to-use and popular supervised learning algorithm. It makes a prediction based on the relationship between the sample to be predicted and its closest neighbors in the training set. Neighborhood relationships are usually determined using the Euclidean distance, which directly measures the distance between two points. Alternatively, Manhattan, Minkowski, and Chebyshev functions can also be used [31]. Once the distance is calculated, the nearest neighbors are ranked and then the sample to be predicted is appropriately assigned to the classes of its nearest neighbors based on this ranking [32].

Evaluation Metrics

In the study, MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and R^2 metrics, which are frequently preferred in literature studies, were used to compare the performances of the models. MSE is a metric that measures how far the predictions of regression models are from the true values [33]. As seen in the formula given in Equation 1, the error at each data point is found by squaring the difference between the actual value and the prediction, and MSE is calculated by adding the squares of the obtained values and taking their average. The closer the MSE value is to zero, the better the model performs. The advantage of MSE is that it eliminates the differences between negative and positive errors by squaring the errors. However, it may cause major errors to be highlighted more.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \tag{1}$$

MAE is the average of the absolute differences between the real values and the values produced by the models, as seen in Equation 2 [34]. Unlike the MSE metric, since errors are not squared, it is resistant to outliers and the best-case scenario is that the MAE value is zero.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \tag{2}$$

RMSE is the square root of MSE, and the closer RMSE is to zero, the better the performance of the model [33]. As seen in Equation 3, the squares of the errors are summed, their average is calculated and the square root is taken. Similar to MSE, this metric emphasizes large errors, but the values become smaller as they get closer to correct predictions.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{3}$$

R^2 measures how close the actual values are to the values predicted by the model. The R^2 value varies between 0 and 1, and the best-case scenario is an R^2 value of 1. As given in Equation 4, the R^2 value is obtained by subtracting the Residual Sum of Squares (SSres) and Total Sum of Squares (SStot) ratio from 1. SSres is the sum of the squares of the differences between the values predicted by the model and the actual values, and SStot is the sum of the squares of the deviations of the actual values from the mean.

$$\mathrm{R2} = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sqrt{\sum_{i=1}^{n}(y_i - y_i')^2}} \tag{4}$$

In the formulas given in Equation 1, Equation 2, Equation 3, and Equation 4, n represents the number of data points, y_i represents the actual values, and y_i' represents the predicted values of the model.

Results

In this section, regional analyses of earthquakes that occurred in the seven geographical regions of Türkiye are discussed under the heading "Earthquake Analysis", while the results obtained from earthquake prediction models are addressed under the title "Earthquake Prediction Model".

Earthquake Analysis

The data obtained by the Kandilli Observatory provides an important resource for evaluating Türkiye's earthquake activity. When we look at the earthquake map of Türkiye in Figure 2, which was created using the earthquake data in the Kandilli Observatory database; It is seen that earthquake disasters have been intense in the Marmara Region and its surroundings, where the North Anatolian Fault line passes, in the Eastern Anatolia region, where the Eastern Anatolian Fault line passes, and in the regions where the Western Anatolian Fault line extending from the Aegean region to the Southeast passes, in the last 23 years. These regions represent the main tectonic structures at earthquake risk in Türkiye.

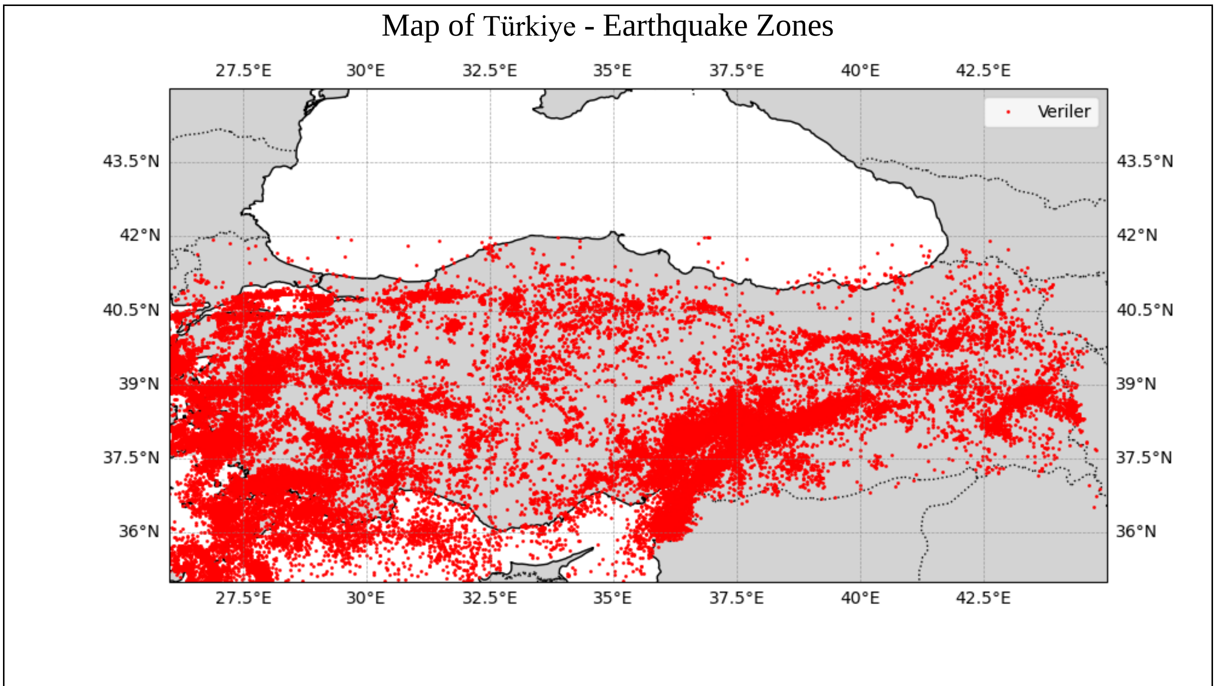


Figure 2. *Map of earthquakes occurring between 2020-2023*

To analyze the number and magnitude of earthquakes that occurred in the 7th region of Türkiye every month by region, the data to obtain the location information in the location column is based on 7 sub-tables. Table 2 shows an example of regular subtables for the Mediterranean Region. Other tables consist of earthquake data with the same columns, and the "occurrence date" column in the tables has been converted to date format and made ready for analysis.

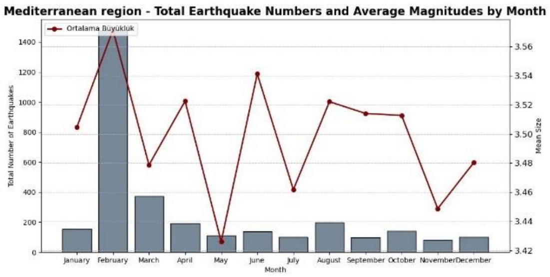
Table 2. *Earthquakes occurring in the Mediterranean Region*

Earthquake id	Date	Time	Latitude	Longitude	Depth (km)	xm	Location	Region
0	2023-02-06 00:00:00	10:24:47	38.0818	37.1773	5	7.6	Kahramanmaras	Akdeniz
1	2021-01-01 00:00:00	16:36:21	35.9273	27.8658	26.1	3.9	Akdeniz	Akdeniz
2	2020-12-25 00:00:00	22:35:24	37.6125	30.6683	5	3.3	Burdur	Akdeniz
3	2020-12-18 00:00:00	9:24:10	37.1175	31.0832	104.9	3.7	Antalya	Akdeniz
4	2020-01-28 00:00:00	21:44:57	35.1688	27.9028	6.4	4.5	Akdeniz	Akdeniz

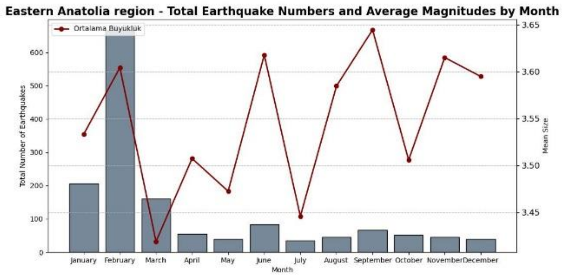
For the analysis, the monthly total number of earthquakes and average earthquake magnitude of each region were calculated and the values obtained are shown in the graphs given in Figure 3.

Upon examining the graphs provided in Figure 3, it can be observed that moderate-sized earthquakes have generally occurred in the seven geographical regions of Türkiye. The region where earthquakes are least observed is the Southeastern Anatolia Region, where the South Anatolian Fault passes through. When comparing earthquakes every month across regions, it is observed that, except for the Marmara Region, in the remaining six regions, earthquake frequencies and magnitudes generally change seasonally. It has been concluded that there is an increase in earthquakes during the winter months and fewer earthquakes during the summer months. According to their depth, earthquakes that occur between 0 and 60 km deep can be classified as "shallow earthquakes", those between 70 and 300 km as "medium depth earthquakes", and those with a depth of more than 300 km as "deep earthquakes" [35]. When the depths of the earthquakes occurring in the 7 geographical regions given in Figure 4 are examined, it is seen that the earthquakes occurring in Türkiye are generally shallow earthquakes that are felt in a narrow area but can cause great damage. In the Mediterranean and Aegean Regions, medium-sized earthquakes, which were felt over a wide area and whose damage was limited, although rare, occurred.

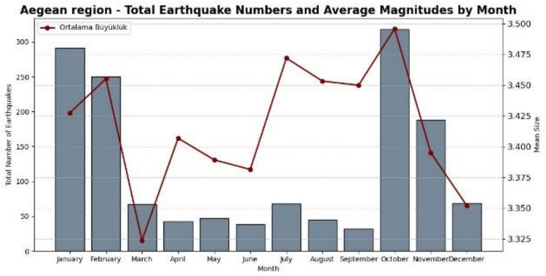
E-ISSN: 2564-7873



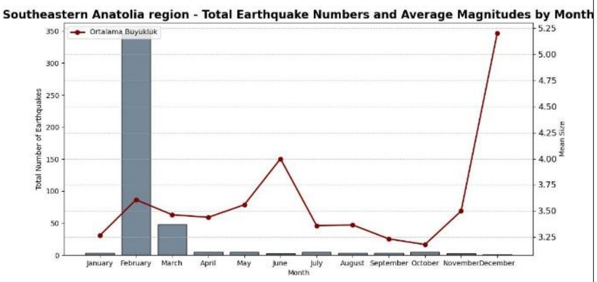
a. Mediterranean region



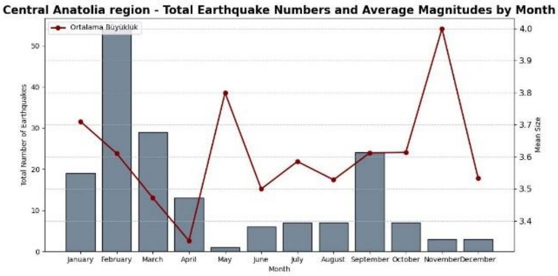
b. Eastern Anatolia region



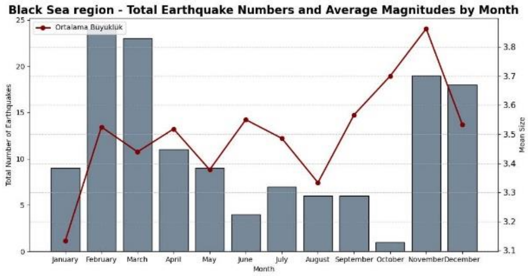
c. Aegean region



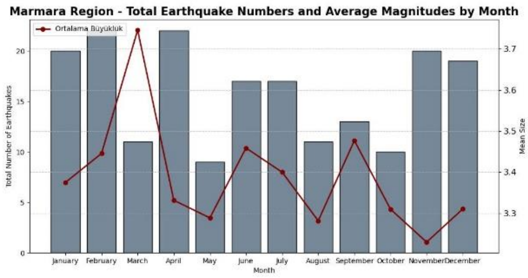
d. Southeastern Anatolia region



e. Central Anatolia region



f. Black Sea region



g. Marmara region

Figure 3. *Total number of earthquakes and average magnitudes by month*

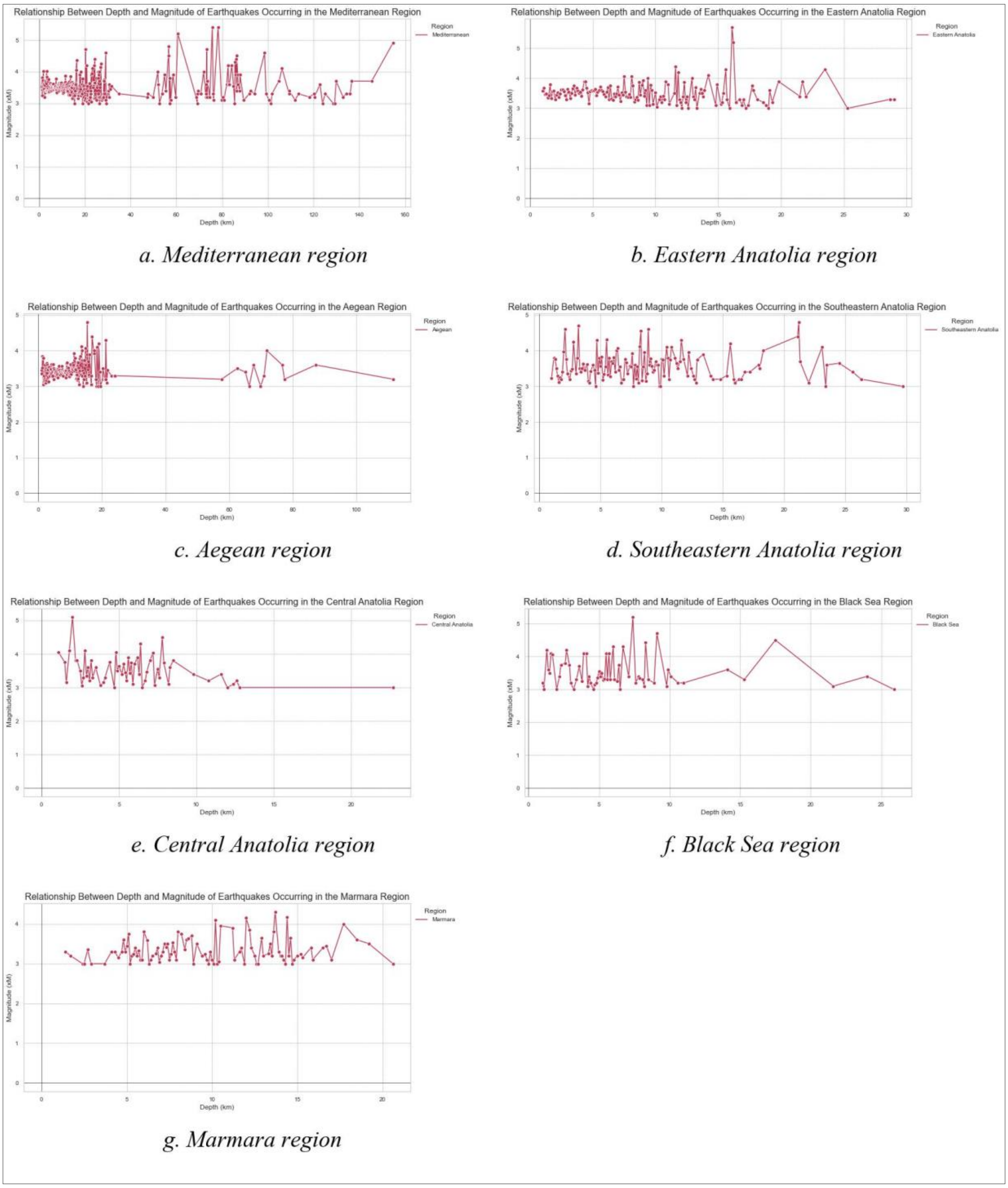


Figure 4. *Depth-magnitude relationship of earthquakes*

When the relationship between the depths and magnitudes of earthquakes was examined, it was observed that there was no significant relationship between these parameters. This shows that earthquake events are a complex and multifactorial process and that it is not possible to explain them with a single parameter.

Earthquake Prediction Model

In the second part of the study, a hybrid learning model in which machine learning algorithms LSTM, RNN, RF, GB, an attempt was made to estimate the magnitude of possible earthquakes. The testing phase of the created models was carried out using the SciKit-Learn library, which offers a wide set of machine learning tools and is widely preferred, especially in the Python programming language. LSTM, RNN, and ANN models are four-layered models with an input layer of 128 neurons with ReLU activation function, two hidden layers of 64 and 32 neurons with ReLU activation function, and an output layer with linear activation function. The first two layers of the models are configured with "return_sequences=True" to pass output to the subsequent layers. To prevent overfitting during training, 20% of the training data is reserved as the validation set. Adam’s optimization algorithm is preferred for optimization, combining adaptive learning rate and momentum for faster and more effective optimization. The mean squared error loss function is used to minimize the average squared differences between predictions and actual values. Finally, the batch size is set to 10, and the models are trained for 100 epochs. For the RF ensemble learning algorithm, hyperparameters are set as n_estimators: 200 and max_depth: 10 to provide a certain balance and stability during the training process to improve performance. In the GB model, n_estimators is set to 100. In the XGBoost model, which is an application of the GB algorithm, n_estimators is set to 50 and max_depth is set to 5. The hybrid model is an integration of two models to create a more comprehensive and powerful prediction model by combining ANN's deep learning capabilities and KNN's example-based similarity measurement. As depicted in Figure 5, an ANN consisting of an input layer with 128 neurons and ReLU activation function, two hidden layers with 64 and 32 neurons respectively, both using ReLU activation function, and an output layer with 1 neuron and linear activation function, has been trained on the dataset. The model uses the Adam optimizer function and mean squared error (MSE) loss function. The model will be trained for 50 epochs, and training data will be arranged into mini-batches of 32 samples each. To find similar examples in the dataset using the representations learned by the ANN model, these representations are given as input to the KNN model. The KNN model performs similarity measurement for n_neighbors=5 based on the learned representations.

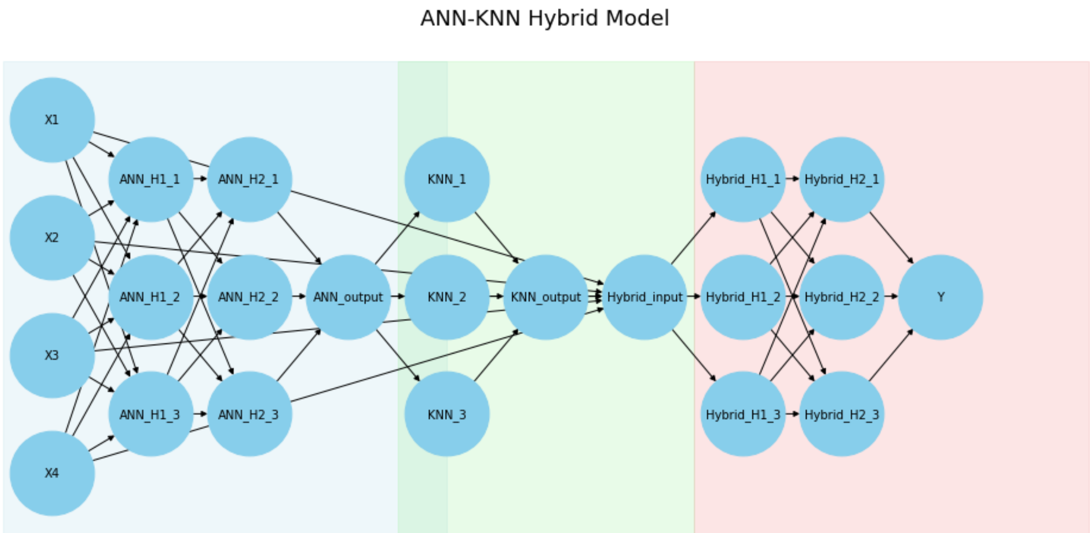


Figure 5. *Proposed model*

The prediction success of the models was evaluated using MSE, MAE, RMSE, and R^2 evaluation metrics as presented in Table 3.

Table 3. *Results obtained*

	MSE	MAE	RMSE	R^2
LSTM	0.0762	0.01030	0.1015	0.03132
RNN	0.0769	0.0102	0.1009	0.0428
RF	0.06280	0.0068	0.0823	0.3636
GB	0.07320	0.0093	0.09629	0.1283
XGBoost	0.0763	0.0100	0.1000	0.0605
ANN	0.07711	0.01026	0.1013	0.0357
ANN-KNN	**0.0418**	**0.0030**	**0.0552**	**0.7138**

The results obtained using MSE, MAE, RMSE, and R^2 metrics were analyzed to evaluate the predictive abilities of the models. Figure 6-7-8-9 shows the MSE, MAE, RMSE, and R^2 results obtained from the models used in the study. When evaluating the predictive abilities of the models based on error criteria, generally lower error values were obtained from RF, GB, XGB, ANN, and ANN-KNN models. Among them, the ANN-KNN model showed the highest performance with 0.0418 MSE, 0.0030 MAE, 0.0552 RMSE and 0.7138 R^2. On the other hand, LSTM, RNN, XGBoost, and ANN models gave high error rates and low R^2 values compared to other models. This means that these models show lower predictive performance in the context of a particular problem compared to others.

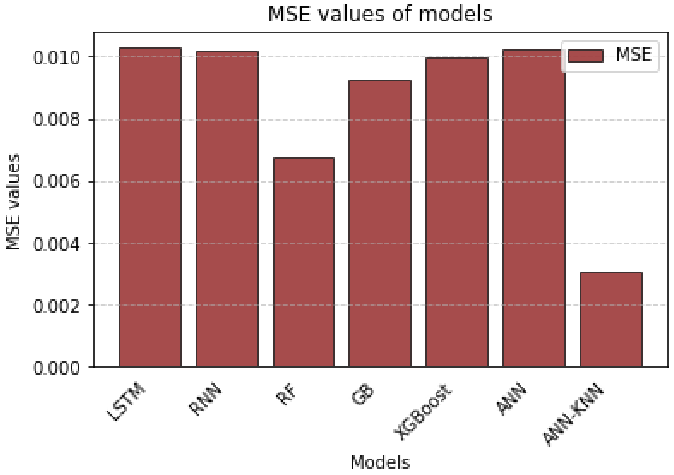


Figure 6. *Comparison of MSE values of the models*

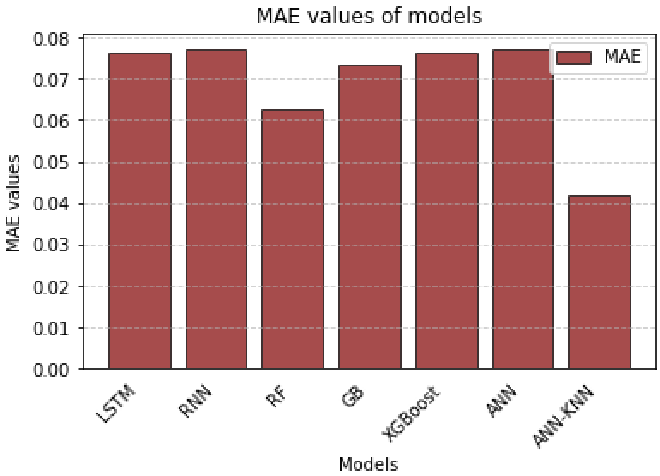


Figure 7. *Comparison of MAE values of the models*

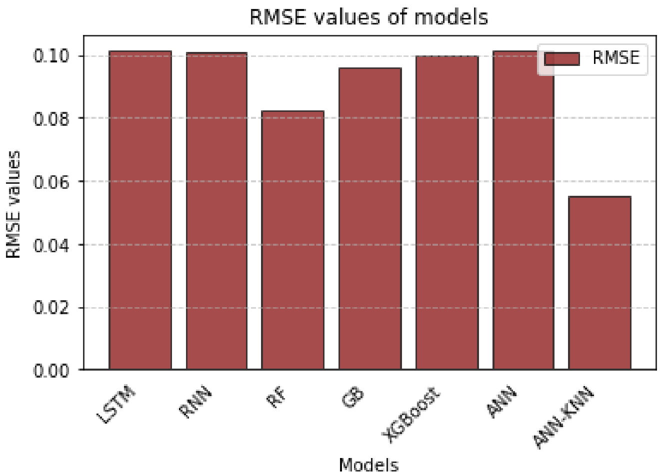


Figure 8. *Comparison of RMSE values of the models*

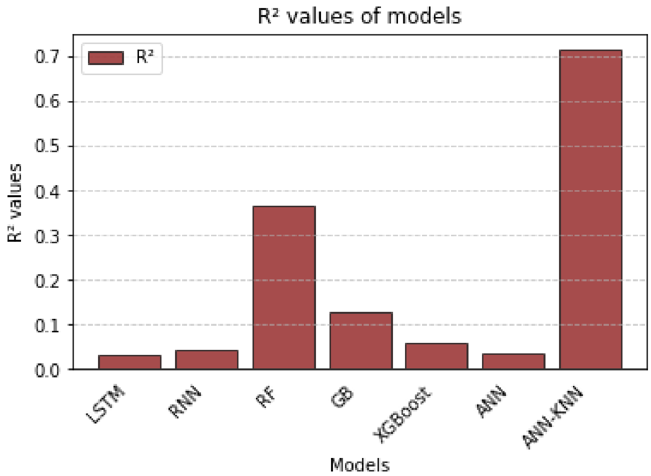


Figure 9. *Comparison of R^2 values of the models*

To visualize the results obtained from the models established in the study, scatter plots of the actual values and the predicted values produced by these models are presented in Figure 10. In these scatter plots, the x-axis represents the actual values, and the y-axis represents the predicted values. The distribution graphs created for each model are compared with the y=x line that runs through the center of the graph. Points that are close to the y=x line indicate good model performance, while points that are far from the line indicate poor performance.

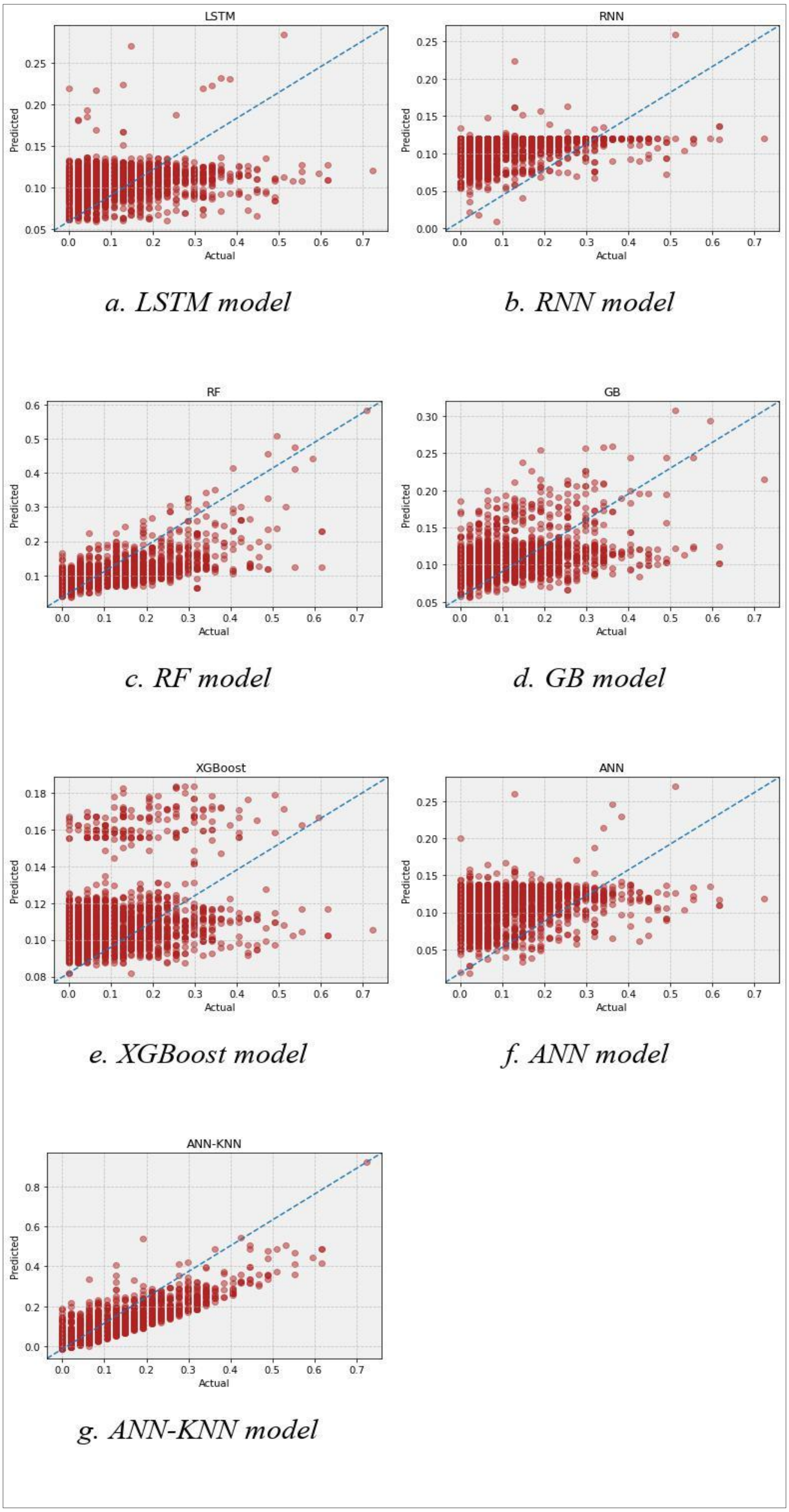


Figure 10. *Distributions of actual values and model predictions*

When examining the scatter plots presented in Figure 10, it can be observed that the LSTM, RNN, XGBoost, and ANN models exhibit points that are distant and widely distributed from the y=x line. Therefore, it is concluded that the predictions of these models do not align with the actual values and show low performance, as seen in the results obtained from Table 3 and Figures 6-7-8-9. On the other hand, in the RF and GB models, the points are closer to and more concentrated around the y=x line, indicating relatively better performance compared to the other four models. The best distribution among the models is achieved with the ANN-KNN model, where the points are close to the y=x line and have a smooth distribution.

Conclusion

Due to the major threat posed by earthquake disasters worldwide, continuous studies related to earthquakes are conducted; however, high success in predicting when and where future major earthquakes will occur has not yet been achieved. Nevertheless, in some countries where earthquakes frequently occur and cause significant destruction, earthquake prediction models are actively utilized [36] In this study, a two-phase earthquake analysis was carried out using the main earthquakes that occurred in Türkiye between 2000-2023 data collected from the Kandilli Observatory and Earthquake Research Institute; aiming to contribute to the literature with the analysis results. In the first phase, the dataset was divided into 7 sub-datasets based on the geographical regions of Türkiye using location information, and the table for each region was analyzed based on monthly total earthquake counts and average magnitudes; the depth-magnitude relationship of earthquakes was examined. In the second phase of the study, a hybrid learning model integrating LSTM, RNN, RF, GB, XGBoost, 2 hidden-layer ANN, and ANN with KNN algorithms was used to attempt magnitude estimation of potential future earthquakes; the models' prediction performances were compared using different evaluation metrics. The results present a positive outlook on the usability of earthquake prediction models.

Acknowledgements -

Funding/Financial Disclosure The authors have no received any financial support for the research, authorship, or publication of this study.

Ethics Committee Approval and Permissions The work does not require ethics committee approval and any private permission.

Conflict of Interests The authors stated that there are no conflict of interest in this article.

Authors Contribution Authors contributed equally to the study. All authors read and approved the final manuscript.

References

[1] AFAD. (2023, February 12). *Açıklamalı Afet Yönetimi Terimleri Sözlüğü.* https://www.afad.gov.tr/aciklamali-afet-yonetimi-terimleri-sozlugu

[2] EM-DAT. (2023, February 18). *Disaster Classification.* https://www.emdat.be/

[3] Santos, G. D. C. (2021). 2020 tropical cyclones in the Philippines: A review. Tropical *Cyclone Research and Review*, *10*(3), 191-199. https://doi.org/10.1016/j.tcrr.2021.09.003

[4] Winsemius, H. C., Aerts, J. C., Van Beek, L. P., Bierkens, M. F., Bouwman, A., Jongman, B., Kwadijk, J. C., Ligtvoet, W., Lucas, P. L., & Van Vuuren, D. P. (2016). Global drivers of future river flood risk. *Nature Climate Change*, *6*(4), 381-385. https://doi.org/10.1038/nclimate2893

[5] Özşahin, E. (2013, September 25-27). *Türkiye'de yaşanmiş (1970-2012) doğal afetler üzerine bir değerlendirme.* [Conference presentation]. Türkiye Deprem Mühendisliği ve Sismoloji Konferansı, Hatay, Türkiye.

[6] Bilham, R. (2009). The seismic future of cities. *Bulletin of earthquake engineering*, *7*, 839-887. https://doi.org/10.1007/s10518-009-9147-0

[7] Kavianpour, P., Kavianpour, M., Jahani, E., & Ramezani, A. (2023). A CNN-BiLSTM model with attention mechanism for earthquake prediction. *The Journal of Supercomputing*, *79*(17), 19194-19226. https://doi.org/10.1007/s11227-023-05497-5

[8] Jia, J. (2016). *Modern earthquake engineering: Offshore and land-based structures.* Springer.

[9] Çam, H., & Duman, O. (2016). Yapay Sinir Aği Yöntemiyle Deprem Tahmini: Türkiye Bati Anadolu Fay Hatti Uygulamasi. *Gümüshane University Electronic Journal of the Institute of Social Science/Gümüshane Üniversitesi Sosyal Bilimler Enstitüsü Elektronik Dergisi*, *7*(17).

[10] Mallouhy, R., Abou Jaoude, C., Guyeux, C., & Makhoul, A. (2019, December 18-20). *Major earthquake event prediction using various machine learning algorithms* [Conference presentation]. 2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), Paris, France. 10.1109/ICT-DM47966.2019.9032983

[11] Wang, X., Zhong, Z., Yao, Y., Li, Z., Zhou, S., Jiang, C., & Jia, K. (2023). Small Earthquakes Can Help Predict Large Earthquakes: A Machine Learning Perspective. *Applied Sciences*, *13*(11), 6424. https://doi.org/10.3390/app13116424

[12] Demirelli, E., Solak, H. İ., & Tiryakioglu, İ. (2023). Makine öğrenmesi algoritmaları ile deprem katalogları kullanılarak deprem tahmini. *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, *13*(4), 979-989. https://doi.org/10.17714/gumusfenbil.1268504

[13] Emre, Ö., Duman, T. Y., Özalp, S., Şaroğlu, F., Olgun, Ş., Elmacı, H., & Çan, T. (2018). Active fault database of Turkey. *Bulletin of earthquake engineering*, *16*(8), 3229-3275. https://doi.org/10.1007/s10518-016-0041-2

[14] Karcı, M., & Şahin, İ. (2022). Derin öğrenme yöntemleri kullanılarak deprem tahmini gerçekleştirilmesi. *Artificial Intelligence Studies*, *5*(1), 23-34.

[15] Kandilli. (2023, January 10). *Kandilli Rasathanesi BDTİM Deprem Sorgulama Sistemi.* http://www.koeri.boun.edu.tr/sismo/zeqdb/

[16] AFAD. (2023, February 18). *Genel Bilgiler.* http://www.koeri.boun.edu.tr/sismo/bilgi/sss_tr.htm

[17] USGS. (2023, January 21). *Earthquakes*. https://www.usgs.gov/

[18] Habek, G. C. (2022). *Makine Öğrenmesi Teknikleri ile Kripto Para Duygu Analizi.* (Tez no. 763894) [Yüksek Lisans Tezi, Manisa Celal Bayar Üniversitesi].

[19] Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill New York.

[20] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

[21] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(5), 771-780.

[22] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785-794. https://doi.org/10.1145/2939672.29397

[23] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

[24] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

[25] Kurtgoz, Y., Karagoz, M., & Deniz, E. (2017). Biogas engine performance estimation using ANN. *Engineering Science and Technology, an International Journal*, *20*(6), 1563-1570. https://doi.org/10.1016/j.jestch.2017.12.010

[26] Akhter, M. N., Mekhilef, S., Mokhlis, H., Almohaimeed, Z. M., Muhammad, M. A., Khairuddin, A. S. M., Akram, R., & Hussain, M. M. (2022). An hour-ahead PV power forecasting method based on an RNN-LSTM model for three different PV plants. *Energies*, *15*(6), 2243. https://doi.org/10.3390/en15062243

[27] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166. https://doi.org/10.1109/72.279181

[28] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

[29] Akandeh, A., & Salem, F. M. (2019, August 04-07). *Slim lstm networks: Lstm_6 and lstm_c6* [Conference presentation]. 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA. https://doi.org/10.1109/MWSCAS.2019.8884912

[30] Clarke, M. (1974). Pattern classification and scene analysis. *In: Wiley Online Library* https://doi.org/10.2307/2344977

[31] Akbulut, Y., Sengur, A., Guo, Y., & Smarandache, F. (2017). NS-k-NN: Neutrosophic set-based k-nearest neighbors classifier. *Symmetry*, *9*(9), 179. https://doi.org/10.3390/sym9090179

[32] Demir, F. (2021). Siber saldırı tespiti için makine öğrenmesi yöntemlerinin performanslarının incelenmesi. *Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *23*(2), 782-791. https://doi.org/10.25092/baunfbed.876338

[33] Nguyen, V. V., Pham, B. T., Vu, B. T., Prakash, I., Jha, S., Shahabi, H., Shirzadi, A., Ba, D. N., Kumar, R., & Chatterjee, J. M. (2019). Hybrid machine learning approaches for landslide susceptibility modeling. *Forests*, *10*(2), 157. https://doi.org/10.3390/f10020157

E-ISSN: 2564-7873
[34] Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PloS One*, *18*(2), https://doi.org/10.1371/journal.pone.0279774

[35] Demirtaş, R., & Erkmen, C. (2000). Odak mekanizması çözümü. *Deprem ve Jeoloji, TMMOB Jeoloji Mühendisleri Odası Yayınları*, *52*, 91-94.

[36] Chen, J., & Zhu, S. (2020). Spatial and temporal b-value precursors preceding the 2008 Wenchuan, China, earthquake (Mw= 7.9): implications for earthquake prediction. *Geomatics, Natural Hazards and Risk*, *11*(1), 1196-1211. https://doi.org/10.1080/19475705.2020.1784297