Journal of Basrah Researches (Sciences) **49**(2), 112 (2023)
DOI: https://doi.org/10.56714/bjrs.49.2.10

Processing and analyzing data to predict earthquakes in Iraq

Nada B. Jarah[1*], Abbas H. AlAsadi[2] , Kadhim M. Hashim[3]

[1] Department of Computer Science, Faculty of Computer Sciences and Maths, University of KUFA, Iraq.
[2] Department of Computer Information Systems, College of Computer Sciences and Information Technology, University of Basrah, Basrah, Iraq.
[3] Computer Technology Engineering Department, College of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq.

ARTICLE INFO

Received 06 August 2023
Accepted 02 November 2023
Published 30 December 2023

Keywords:

Earthquake, Data, catalog, Model, Predaction.

Citation: N. B. Jarah et al., J. Basrah Res. (Sci.) **49**(2), 112 (2023).
DOI:https://doi.org/10.56714/bjrs.49.2.10

ABSTRACT

The An earthquake is a devastating natural disaster that causes great economic and human losses because it occurs without warning. The increase in earthquakes in Iraq has raised concerns about the future of the region. It is necessary to study earthquake prediction and determine the location, size and time of the earthquake. A machine learning model was proposed to predict earthquakes in Iraq using two sources: the first is a catalog of data from 1900 to 2019, which includes 36,663 earthquakes, and the second is from the USGS for one year from 2022 to 2023, which includes 25,000 earthquakes. Preliminary processing of the data was done, removing outliers and integrating Date and time data in timestamps, and the five important features for prediction were identified and the data was divided into 80% for training and 20% for testing. After applying several attempts in using different models, the best results were achieved using NN and the accuracy was about 0.7. The most important reason for this result is training over many years that may change geologically. The study compared its results with other studies to predict earthquakes across different regions of the world.

1. Introduction

The Earth is a complex system. Some phenomena occur in an instant, such as earthquakes, and they are not new events on Earth, but they occurred naturally in the past, seeing the past is the same as seeing the future, and therefore it is necessary to rely on monitoring a network consisting of seismographs installed deep in tunnels in locations. Different. Multiple sites across Iraq to accurately record tremors at different periods, where observed waveform data, information to resolve earthquake mechanisms, etc. are available. Table 1 shows seismic monitoring stations in Iraq. Machine learning generally requires a large amount of training data. A catalog of earthquakes in Iraq and neighboring regions was compiled for 119 years and used USGS data via its website for one year from 2022 to 2023. It identified the five important features that are currently available in earthquake data for Iraq. The adjacent areas are earthquake Magnitude, depth, longitude, latitude, and Timestamp. Thus, the structure of developing the proposed model was represented in the following three stages: First: data collection, data cleaning, feature selection, model building,

[*]**Corresponding author email :** nadabadrjarah@yahoo.com

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.iobrs.edu.iq

model testing, and model performance measurement. Second: Mathematical analysis of the proposed model. Third: Forecasting results for early prediction of earthquakes and how to preserve people's lives and property. Figure 1 shows the structure of the proposed model.

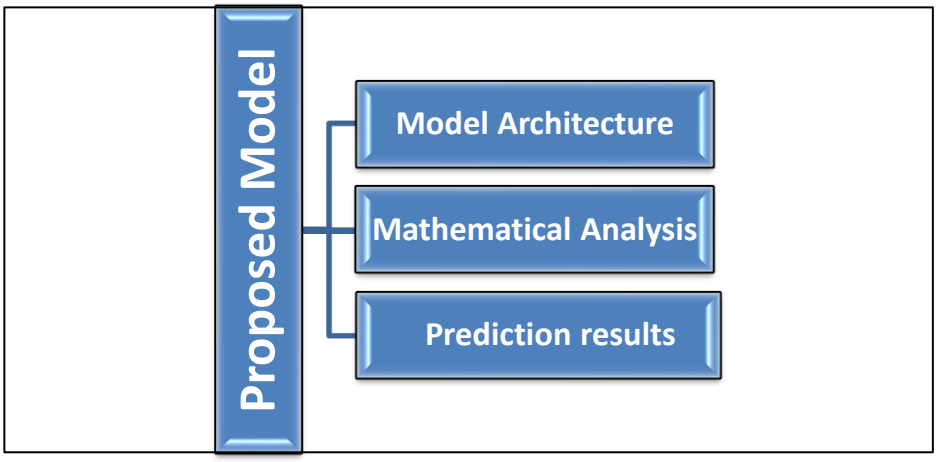

Figure 1 . structure of the proposed model

The model preprocesses the data to remove outliers and combines date and time data into timestamps. Divide the data into 80% for training and 20% for testing. Several algorithms were used, the most important of which were linear regression, random forest regression, and neural network (NN), to predict upcoming earthquakes in Iraq.The best results were achieved using NN, where the first type of data sourced from the Iraqi Seismic Network (ISN) achieved an accuracy of 0.7754 for the input features Timestamp, Depth, Magnitude, and the predicted features Longitude, Latitude, While USGS seismic data showed a resolution of 0.7328 for the input features Timestamp, Longitude, Magnitude, the features were the expected depth and latitude.

2. Related Works

Rouet-Leduc et al. [1] applied laboratory data to identify the hidden signal that precedes earthquakes, which was previously believed to be low-amplitude noise that allows the prediction of faults throughout the laboratory earthquake cycle. Predictions are based on the momentary physical properties of the acoustic signal, not its history. The study concludes that the acoustic signal emitted by the laboratory fault is the displacement of fault blocks and that applying this approach to continuous earthquake data identifies earthquake signals. This was the first application of ML to continuous acoustic/earthquake data to infer failure times.

In 2019, Corte et al. [2] reviewed the analysis of huge data sets using powerful computational techniques and cloud structures, in the areas most exposed to seismic activity, which is the state of California in the period from 1970 to 2017, where data is available in this area and 1GB of information has been processed, and four methods of common regression algorithms are used. To predict earthquake magnitude over the next 7 days and apply stacking-based group learning, relative error reporting.In Mallouhy et al. [3], eight different machine-learning algorithms were applied to an earthquake data set collected at the California Center for 36 years to classify major earthquake events between negative and positive: Random Forest, Naive Bayes, Logistic Regression, Multilayer Perceptron (MLP), AdaBoost, K-nearest neighbors (KNN), Support Vector Machine, and Classification and Regression Trees. Different hyperparameters were chosen for each model, and the prediction results were compared using different measures for each algorithm. Fairly, it was found that the prediction is reliable for the important events of the three algorithms: KNN, Random Forest, and MLP, by producing the least false output.Bangar et al. [4] designed and developed an earthquake disaster prediction system based on detecting early signs of earthquakes using machine-learning algorithms. Data were collected for the Indian subcontinent with the rest of the world from government sources. Then, the data was preprocessed by building a model that combines random forest algorithms and vector machine support. The mathematical model is developed based on the "training data set" Thus, integrating earthquake activity with machine learning

technology produces an effective and important result for large-scale earthquake prediction.

1. Materials and methods :

1.1. Model Architecture:

Based on our review and analysis of the literature for the study. Different structures were discovered that were used in the proposed models and frameworks but the most common steps for building the models are as follows:

3.1.1 Data Collection Stage

This stage describes the process of collecting research data necessary to develop the proposed model. There are four types of data, and **Figure** 2. shows the components of data collection methods that can be collected to support research development, namely:

- **Questionnaire**: data is collected by using scaled answers questionnaires which typically involve presenting a series of questions or statements to research participants and asking them to rate their level of agreement or disagreement on a pre-defined scale. For example, a commonly used scale is the Likert scale, which ranges from "strongly agree" to "strongly disagree"[5].

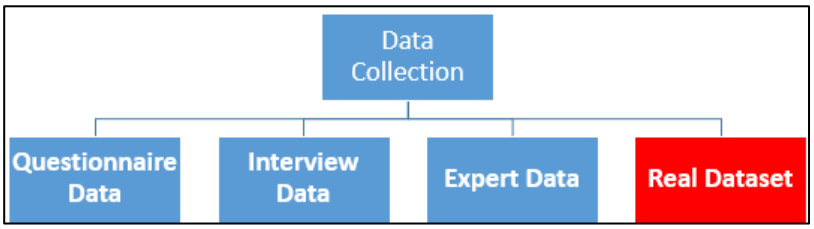

Figure 2. Data Collection Methods

- **Interview**: data is collected by interview involves obtaining information from participants through face-to-face or virtual conversations, typically guided by a set of pre-defined questions or topics. The interviewer may ask open-ended questions that allow participants to share their experiences, opinions, or perspectives in their own words or close-ended questions that require participants to select from a set of predetermined response options [6].
- **Expert**: Expert data collection involves collecting information from individuals who have specialized knowledge or skills related to a particular research topic. Experts may include professionals, practitioners, academics, or other individuals who have extensive experience or training in a specific field [7].
- **Real Dataset**: The real data set during this research is based on the Iraqi earthquake prediction centers and directly from the earthquake laboratory in the Department of Earth / College of Science, University of Basra. This center collects earthquake data based on a digital network of the latest seismic and geophysical sensors connected to a communications network. The earthquake data were recorded based on the extent of Iraq, which was used as one of the sources of structural information while these stations were spread in the region of Iraq and connected to the network, as these stations represent three sensor networks and Table 1 shows the seismic monitoring stations in Iraq [8].

Table 1: Data collection Network

	Location	No of Sensors
ıern Iraq Seismic Network (NISN)	Khanaqin	3
seismic network (ISN)	Al-Jadriya	16
mic Observatory (ISO)	University of Basra	9

Here, a catalog of earthquakes in Iraq and the surrounding and influential region has been compiled for the

study of a large project of Probabilistic seismic-hazard assessments (PSHA) and from various sources, it is characterized by the lack of detailed information, the presence of many known active faults, and the loss of many local earthquakes. in the study area. The first part of the earthquake catalog is from 1900 to the end of 2009. It includes more than 18,000 earthquakes [9], and the following are the sources of seismic data for the study area within the catalog used in this study:

1) Iraqi Seismological Network (ISN)
2) International Seismological Center - Global Seismic Model (ISC-GEM)
3) International Seismological Center (ISC)
4) Euro-Mediterranean Seismological Center (EMSC)
5) Geological Survey Catalog
6) Universal central moment tensor (CMT universal)
7) National Geophysical Data Center/Wetlands Database (NGDC/WDS)
8) USGS Centennial Catalog
9) Ambrasys' extensive work cataloging earthquakes in the Mechanistic Age in the Middle East

Previously recorded earthquake data between 2009 and 2014 have also been added to ISN and the stations are: Baghdad (BHD), Mosul (MSL), Rutba (RTB), Nasiriyah (NSR), Badra (IBDR), Kirkuk (IKRK) [10].
And from 2014 to 2019 from the earthquake laboratory at the University of Basra / College of Science / and from the stations: Basra, Nasiriyah, Samawah, Amarah, Karbala, Anbar, Kirkuk, Sulaymaniyah, Dohuk [11].
Finally, the earthquake catalog used in this study covers the area bounded by latitudes 26°-40°N and longitude 36°-51°E. For a period of 119 years from 1900 to 2019, it consisted of (34,663 aftershocks).
The earthquake intensity ranged up to 7.7 [12], Table 2 shows the earthquake catalog.

Table 2. Catalog of earthquakes in Iraq and the surrounding area (1900-2019)

EVENTID	ISC EVENTID	YEAR	MONTH	DAY	TIME	LAT	LON	DEPTH (km)	MAG
10001	N/A	1900	2	24	30:00.0	38.45	44.87	0	5.4
10002	N/A	1900	4	17	17:00.0	38	46	0	6.2
10003	N/A	1900	10	10	00:00.0	39.1	42.5	0	5.2
10004	N/A	1901	2	6	48:00.0	33	49	0	7.4
10005	N/A	1901	5	20	38:36.0	38.38	42.23	10	5.5
:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:
44659	6.17E+08	2019	12	29	28:27.4	38.8709	43.5207	14.3	2.1
44660	6.17E+08	2019	12	30	56:21.8	30.8205	50.0474	10	3.7
44661	6.17E+08	2019	12	31	10:58.9	39.7427	43.5995	9	2.4
44662	6.17E+08	2019	12	31	50:57.3	32.4992	46.9521	10	2.8

3.1.2 Data Cleaning Stage

Data cleaning is an essential step in the process of preparing data for machine learning models. The quality of the data used to train a model is crucial, and poor quality data can lead to inaccurate or biased results[13]. These are some of the common techniques used for data cleaning in machine learning. However, the choice of data cleaning technique depends on the dataset and the specific problem being addressed[14]. Here are some common techniques used for data cleaning in machine learning:

- Handling Missing Values: Missing data is a common problem in datasets. One way to handle missing values is to simply remove the rows or columns with missing data. However, this can lead to a significant loss

of data. Another approach is to impute the missing values by using techniques like mean imputation or regression imputation[15].

- Outlier Detection: Outliers are data points that are significantly different from other data points. Outliers can skew the results of a machine learning model. One way to deal with outliers is to remove them from the dataset. Another approach is to use robust statistical methods that are less affected by outliers[16].
- Data Normalization: Normalizing data involves scaling the data so that it has a mean of zero and a standard deviation of one. This helps to ensure that the features in the dataset have a similar scale, which can improve the performance of the machine learning model[17].
- Removing Duplicates: Duplicate data points can bias the model's training and reduce its effectiveness. Therefore, it is essential to identify and remove duplicate data points from the dataset.
- Text Preprocessing: Text data requires specific preprocessing steps like removing punctuation, lowercasing the text, removing stop words, and stemming or lemmatizing the text. This can improve the quality of the text data and help the machine learning model to perform better.
- Balancing Class Distribution: If the dataset is imbalanced, i.e., it has significantly more samples from one class than the others, the model's performance may be biased towards the majority class. To overcome this issue, data augmentation techniques such as oversampling or under sampling can be applied to balance the class distribution.
- These are some of the common techniques used for data cleaning in machine learning. However, the choice of data cleaning technique depends on the dataset and the specific problem being addressed.

The main cleaning techniques have been applied on the research dataset are:

- Remove the missing data, there are some missed data that affect the quality of proposed model. These rows have been removed from the dataset and the final size of data set consist of 26397[18].

1. Timestamps are used to represent time in various formats. Some common timestamp formats are Unix Timestamp which represents the number of seconds that have elapsed since January 1, 1970, 00:00:00 UTC. This format is widely used in computing systems and is also known as Epoch time. Because the data collected by different stations with different data and time format we facing such difficulties to train our model in different format of data. The date and time converted to float number as time stamp by using the following formulate:

$$\text{Timestamp} = \text{current data} - (1970,1,1) * 86400 \quad \ldots.(1)$$

A new field was added with the name of timestamp, which represents column data: Year, Month, Day, Time, and Date, all read in seconds, as in Table 3.

Table 3. Cleaning Dataset

	EVENTID	YEAR	MONTH	DAY	Time	Date	Latitude	Longitude	Depth	Magnitude	ISC EVENTID	Timestamp
40	10041	1909	2	9	0:24:06	2/9/1909	40.0000	38.0000	60.0	6.8	914221.0	-1.921621e+09
82	10083	1923	4	29	0:34:35	4/29/1923	40.0000	37.0000	35.0	5.6	911346.0	-1.473032e+09
94	10095	1924	9	13	0:34:05	9/13/1924	39.8640	41.8760	35.0	6.8	911110.0	-1.429572e+09
34660	44661	2019	12	31	5:10:58	12/31/2019	39.7427	43.5995	9.0	2.4	617151713.0	1.577769e+09
34661	44662	2019	12	31	12:50:57	12/31/2019	32.4992	46.9521	10.0	2.8	617208655.0	1.577797e+09
34662	44663	2019	12	31	14:49:12	12/31/2019	34.7548	45.5659	12.2	2.8	617208656.0	1.577804e+09

26397 rows × 15 columns

3.1.3 Feature selection Stage

Feature selection is the process of identifying the most relevant features in a dataset to improve the performance of the proposed model. There are some common techniques used for feature selection to develop a models:

- Correlation Analysis: This technique involves analyzing the correlation between the input features and the output variable. Highly correlated features can be removed as they are redundant and do not provide additional information to the model [19].
- Recursive Feature Elimination: This is an iterative technique that involves removing the least important features in a dataset and evaluating the performance of the model. The process is repeated until the desired number of features is reached or the performance of the model is optimized.

These are some of the common techniques used for feature selection in machine learning models. However, the choice of technique depends on the dataset, the size of the dataset, and the specific problem being addressed. It is essential to experiment with different techniques to find the optimal set of features that can improve the performance of the machine learning model. Based on this research dataset there are only five features related to identify the earthquake efficiently as the explained in Table 4.

Table 4. Features Selection

No	Feature	Selection	Reasons
1	EVENTID	N	Has no effect on prediction
2	YEAR	N	Used in Timestamp
3	MONTH	N	Used in Timestamp
4	DAY	N	Used in Timestamp
5	TIME	N	Used in Timestamp
6	**Timestamp**	**Y**	**Effecting the prediction**
7	**LAT**	**Y**	**Effecting the prediction**
8	**LON**	**Y**	**Effecting the prediction**
9	**DEPTH**	**Y**	**Effecting the prediction**
10	**MAG**	**Y**	**Effecting the prediction**
11	AUTHOR	N	Has no effect on prediction
12	TYPE	N	Has no effect on prediction
13	ISC EVENTID	N	Has no effect on prediction
14	SOURCE	N	Has no effect on prediction

3.1. 4 Model Building

The model building consist of 5 steps : importing, cleaning, Split into training and testing, then fitting Figure 3. shows the main steps of model building.

The research dataset was uploaded on Google Drive to read directly online on Google Colab. The most needed libraries have been imported to be utilized in dataset reading and viewing .

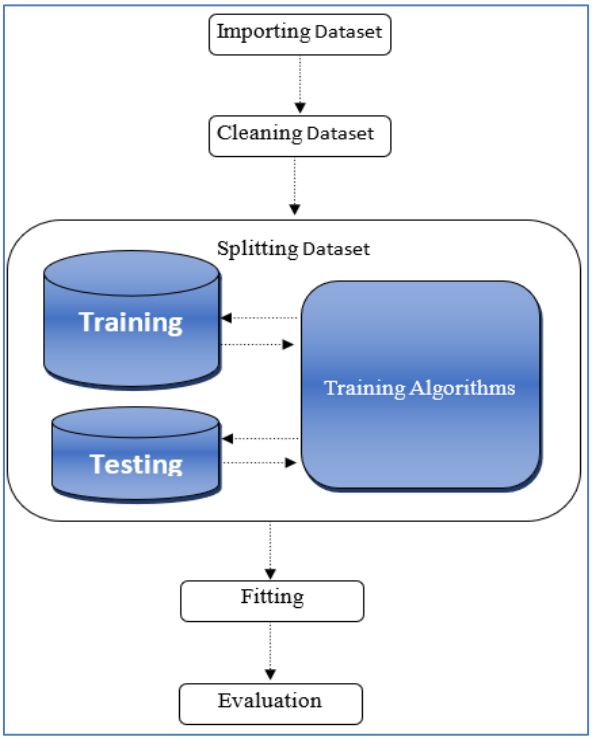

Figure 3. Earthquake Prediction Model Building

Cleaning dataset to remove the records that include missing values by applying some functions
The data is divided into training data and test data, so 80% is for training and 20% for testing, as shown in Algorithm 3. According to the Iraqi data above, which consist of 26397 earthquakes, it is: for training 21117 and for testing 5280.
The model is built, trained and fit to predict the target traits .then run ten times, each one changes the input to predict a different output and the result is shown in Table 5.

Table 5. The prediction Accuracy

No	Input Features	Predicted Features	ISN	USGS
1	Timestamp, Latitude, Magnitude	Depth, Longitude	0.5613	0.4717
2	Timestamp, Depth, Longitude	Magnitude, Latitude	0.5728	0.4306
3	Timestamp, Latitude, Depth	Magnitude, Longitude	0.6254	0.6128
4	Timestamp, Latitude, Longitude	Magnitude, Depth	0.6294	0.5865
5	Depth, Longitude, Magnitude	Timestamp, Latitude	0.6459	0.5135
6	Latitude, Depth, Longitude	Magnitude, Timestamp	0.7152	0.565
7	Latitude, Depth, Magnitude	Timestamp, Longitude	0.7224	0.6249
8	Timestamp, Longitude, Magnitude	Depth, Latitude	0.7328	**0.7328**
9	Longitude, Latitude, Magnitude	Depth, Timestamp	0.7733	0.716
10	Timestamp, Depth, Magnitude	Longitude, Latitude	**0.7754**	0.664

3.2. Mathematical Analysis

Seismic ground motion is controlled by three factors: the earthquake source, seismic wave path and local site effects. Under certain conditions - the mathematical formulates will be elaborated within this section- it is possible to describe the particle Seismic $E(x,t)$ caused by an earthquake at some point x at time t through a convolution of these effects, or, equivalently, by a multiplication in the frequency domain with.

$$E(x,y) = A(x,y).T(x,y).M(x,y).D(x,y) \text{ where : } LO(x,y) \text{and } LA(x,y) \in A(x,y) \ldots\ldots (2)$$

$A(x, y)$ mean the area will affected by the earthquake that consist of $LA(x, y)$ Latitude and Longitude $LO(x, y)$ The multilayer perceptron network (MLP) has been utilized and classify the dataset to five features which are $(T(x, y). M(x, y). D(x, y) . LO(x, y) LA(x, y)$. Based on the features engineering there only three features used as input $D(x, y) . LO(x, y) LA(x, y)$ to predict two output $T(x, y). M(x, y)$ which critical challenge because the number of training feature too small in compare with the predicted values. The proposed model use multilayer perceptron (MLP) network because the list of input cannot separate by one direction. Figure 4. shows the network used.

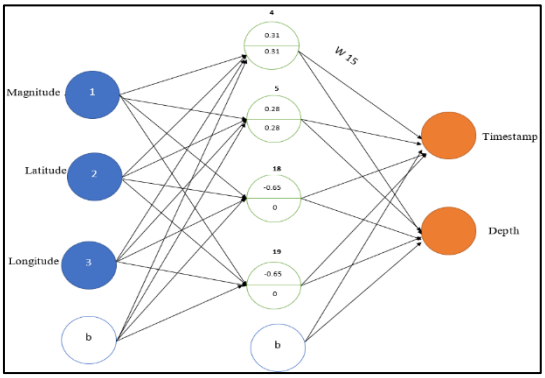

Figure 4. multilayer perceptron (MLP) network

Input Signals$(X1, X2, \ldots\ Xn)$
Initial weights$(W1, W2, \ldots \ldots, Wn)$
Output Signals$(Y\ )$
Desired Valu(d)
Let assume the learning rate is 0.3

Forward Propagation

Each neuron includes two values which are net and observation
Net calculate by the following equation:

$$x = \sum_{i=1}^{n} \text{wixi} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)\text{net}$$

Observation value can be calculated by activation function which is Relu as the following equation:

$$\text{Relu} = \sum_{0}^{z} \begin{matrix} z>0 \\ z\leq 0 \end{matrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)\ \text{Observation vlaue}$$

Error Calculation:

$$\text{Err}_{\text{Val}} = \frac{1}{2} * (d - y)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)\ \text{Error Value}$$

Here is the real applying of the formula 3 and 4

Formula 3 & 4 can apply for all neuron from 4 to 19 to find the net and observation values which will used as input to neuron 20 & 21.
To find the all error has been occurred dunning the Epoch 1 need to apply the following formula:

$$\text{Err}_{\text{sum}} = \sum_{\text{eo}}^{n} \frac{1}{2} * (d - y)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)\ \text{Error Value}$$

Back propagation changes of the prediction Error by weight changing

$$\delta(\text{Output neuron}) = y_n(1 - y_n)(t_n - y_n) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

$$\delta\ = Y(1 - y)(\ \text{Timestamp} - y)$$

$$\delta\ = Y(1 - y)(\ \text{Magnitude}\ - y)$$

$$\delta(\text{Hidden Neuron}) = o_h(1 - o_n)(\delta_l\ w_{nl}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where the predicted values are not meet the desired values as outputs then the weight need to be adapted according to the equation 8 & 9:

$$w_{(n+1)} = w_n + n[d(n) - Y(n)]X(n) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)\text{new weight}$$

Where $W_{i\ is}$ the weight of neuron, X_i input values and n learning rate.

4. Prediction results

A descriptive analysis of the results of the proposed model appears in Table 6. To ensure that the proposed model meets the best predictive accuracy, we run the model ten times while changing the input features each time, and the results obtained from running the proposed model are about 0.7152, 0.7224, and 0.7328, 0.7733, and 7754 to predict magnitude and timestamp, longitude and timestamp, depth and latitude, and timestamp and depth, longitude and latitude, respectively. The result will be approximately 0.7 which helps to understand the size, depth and timing of the earthquake well. The result of this research is very important and useful in making decisions regarding population conservation in earthquake-affected areas. This is because the result provides us with valuable insights and numerical data, and this data can add a layer of certainty and understanding to our decision-making process.
We found that the best result is 0.7754 to find the location of the next earthquake, and this result can help people in the affected area to evacuate and take shelter in a safe place as well as emergency response to attend at the same location.

5. Comparison of the Research Results

Comparing the accuracy of the proposed model with the results of other researchers shown in Table 6, the proposed model achieved 0.78 as a validation score while others were classified in this upper and lower value. (Malouhi and Jawda, 2020) recorded 0.76 as accuracy results while Assem et al. (2016) reported 0.64 and Leduc et al. (2017) showed that the highest achieved result was 0.62, and finally 0.76 was reached by Karimzadeh et al., 2019. These researchers conducted their studies in different datasets for websites. As different as Iran, India and the world. The proposed model achieved better results than other researchers. Three studies reported higher accuracy than the proposed model, with Southwick et al. (2022) recording 0.9, Bangar et al. (2020) achieving 0.83 and finally Curtis et al. (2019) showing 0.8 as the accuracy value. Southwick et al. (2022) achieved better results than the proposed model for several reasons, such as the number of features to train the model, which is 14 vectors included with the proposed model and only 3 vectors for the training model, and this is one of the most important effects on the model's accuracy. Bangar et al. (2020) conducted the study and checked the model on an Indian database and the number of training features was 6 vectors which resulted in an accuracy of 0.83. The main reason was related to the size of the data set used in this research, which was approximately 600 million data records

Table 6. Comparison of the Research Results

No	Authors	Region	Algorithm	Accuracy	Highest Accuracy
1	(Sathwik et al, 2022) [20]	Public	Logistic Regression, Support Vector Machine, Random Forest Classifier K-Nearest Neighbors	0.9	0.9
2	(Bangar et al, 2020) [4]	Indian	Regression Model, Boosting Model Stacking Model	0.74 0.76 0.83	0.83
3	(Mallouhy & Jaoude, 2020) [3]	Public	Random Forest Support Vector Machine Logistic Regression Naive Bayes KNN Multilayer Perceptron AdaBoost CART	0.76 0.66 0.68 0.74 0.72 0.75 0.74 0.70	0.76
4	(Cortés et al, 2019) [2]	California	Regression algorithms	0.8	0.8
5	(Asim et al,2016) [21]	Hindukush India	PRNN, RNN, Random forest, LPBoost ensemble	0.58, 0.64, 0.62 0.65	0.64
6	(Leduc et al.,2017) [1]	Hindukush	Neural Network Random Forest	0.58 0.62	0.62
7	Karimzadeh et al, 2019 [22]		Logistic Regression Naïve Bayes	0.74 0.78 0.58	0.76

			K-Nearest Neighbours, SVM Random Forest	0.75 0.76	
8	Our Research	Iraq	Linear Regression Random Forest Neural Network	0.33 0.72 0.83	0.83

6.Conclusion

The most basic observation about the results of the study is that the predicted results are not very accurate due to the insufficient number of training features to predict two outcomes. And the size of the data set used in the training was less than 26,000 as a result of the lack of data, so the number of training features will be increased in the future to obtain the best prediction value.

Most importantly, the data gathered from this research provides a solid basis for long-term planning. It allows us to predict possible events, anticipate developments, and adapt accordingly. This leads to proactive decisions aimed at preserving and improving housing, ultimately improving the quality of life for residents. In conclusion, the important and useful findings of this research are vital in making informed decisions regarding housing upkeep and maintenance. By providing insights into safety, sustainability, resident preferences and long-term planning, this research provides us with the knowledge needed to ensure that dwellings remain desirable and conducive places to live.

7.References

[1] B. R-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, P. A. Johnson, Geophysical Research Letters, **44**(18), 9276 (2017). Doi: https://doi.org/10.1002/2017GL074677

[2] G.A. Corte's, A.M. Esteban, X. Shang, F.M. Alvarez, Computers & Geosciences, **115**, 198 (2019). Doi: https://doi.org/10.1016/j.cageo

[3] R. Mallouhy, C. A. Jaoude, C. Guyeux, A. Makhoul International Conference on Information and Communication Technologies for Disaster (2019). Doi: https://doi.org/10.1109/ICT-DM47966.2019.9032983

[4] P. Bangar, D. Gupta, S. Gaikwad, B. Marekar J. Patil, International Journal of Recent Technology and Engineering (IJRTE) **8**(6), 4684 (2020). Doi: https://doi.org/10.35940/ijrte.E9110.018620

[5] A. Joshi , S. Kale , S. Chandel , D. K. Pal, Current Journal of Applied Science and Technology **7**(4), 396 (2015). Doi: https://doi.org/10.9734/BJAST/2015/14975

[6] S.Jamshed, J Basic Clin Pharm. **5**(4), 87 (2014). Doi: https://doi.org/10.4103/0976-0105.141942

[7] G.Bowen, Qualitative Research Journal **9**(2), 27. Doi: https://doi.org/10.3316/QRJ0902027

[8] S. Dash, S.K.Shakyawar, M. Sharama, S. Kaushik, Journal of Big Data **6(1)**, 54 (2019). Doi: https://doi.org/10.1186/s40537-019-0217-0

[9] T. Onur, R. Gök, W. Abdulnaby, H. Mahdi, N. M. S. Numan, H. Al-Shukri, Ammar M. Shakir, H. K. Chlaib, Taher H. Ameen, Najah A. Abd, Seismological Research Letters **88**(3), 798 (2017). Doi: https://doi.org/10.1785/0220160078

[10] H. J. Mohammed S. H. Faraj, Iraqi Geological Journal **39–49**(2), 104 (2016). Doi: https://doi.org/10.46717/igj.39-49.2.8Ms-2016-12-31

[11] W. Abdulnaby ,T.Onur ,R.Gok, A.M. Shakir, H. Mahdi, H. Al-Shukri, N.S. Numan, H.K.Chlaib, T.H. Ameen , A.Ramthan, Journal of Seismology **24**(3), 2020. Doi: https://doi.org/10.1007/s10950-020-09919-2

[12] M. A. Salam , L. Ibrahim and D. S Abdelminaam , Article Published in International Journal of Advanced Computer Science and Applications (IJACSA), **12**(5), 2021. Doi: https://doi.org/10.14569/IJACSA.2021.0120578

[13] M. A. Priestley, F. O'Donnell, E. Simperl, ACM Journal of Data and Information Quality, 2023. Doi:https://doi.org/10.1145/3592616

[14] F. Ridzuan, W. M. Zainon, Procedia Computer Science **161**, 731 (2019). Doi:https://doi.org/10.1016/j.procs.2019.11.177

[15] A. Palanivinayagam, R. Damaševičius, Information **14**(2), 2023.Doi: https://doi.org/10.3390/info14020092

[16] D. Cousineau, S. Chartier, International Journal of Psychological Research **3**(1).Doi: https://doi.org/10.21500/20112084.844

[17] P. J. M. Ali, R.H. Faraj, Machine Learning Technical Reports **1**(1), 1 (2014).Doi: https://doi.org/10.13140/RG.2.2.28948.04489

[18] H.Kang, Korean Journal of Anesthesiology **64**(5), 402 (2013).Doi: https://doi.org/10.4097/kjae.2013.64.5.402

[19] J.Y.Chan,S.M. Leow,K.T. Bea,W.K. Cheng, S.W.Phoong, Z.W.Hong, Y.L.Chen, Mathematics **10**(8), (2022).Doi: https://doi.org/10.3390/math10081283

[20] G. Sathwik, V.S. Patil, S. Br, S. Vishesh, J. Yatish, S. Soham, International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE) **11**(11), 110 (2022).Doi: https://doi.org/10.17148

[21] K. Asim, A. Idris, T. Iqbal, F. Álvarez, PloS one **13**(7), (2018), Doi:https://doi.org/10.1371/journal.pone.0199004

[22] S. Karimzadeh, M. Matsuoka, J.Kuang, L. Ge, ISPRS International Journal of Geo-Information **8** (10), (2019).Doi:https://doi.org/10.3390/ijgi8100462

Journal of Basrah Researches (Sciences) **49**(2), 112 (2023)
DOI: https://doi.org/10.56714/bjrs.49.2.10

معالجة وتحليل البيانات للتنبؤ بالزلازل في العراق

ندى بدر جراح[1]*، عباس حنون الاسدي[2]، كاظم مهدي هاشم[3]

[1]كلية علوم الحاسوب والرياضيات، قسم علوم الحاسوب، جامعة الكوفة ، العراق.
[2]قسم نظم المعلومات ، كلية علوم الحاسب وتقنية المعلومات، جامعة البصرة ، البصرة، العراق.
[3]كلية تكنولوجيا المعلومات ، جامعة الإمام جعفر الصادق ، ذي قار، العراق.

معلومات البحث

الاستلام 06 اب 2023
القبول 02 تشرين الثاني 2023
النشر 30 كانون الأول 2023

الكلمات المفتاحية

الزلزال ، البيانات ، الكتالوج ، النموذج ، التنبؤ

الملخص

يعد الزلزال كارثة طبيعية مدمرة تتسبب في خسائر اقتصادية وبشرية كبيرة لأنه يحدث دون سابق إنذار. وأثارت زيادة الزلازل في العراق مخاوف بشأن مستقبل المنطقة. ومن الضروري دراسة التنبؤ بالزلزال وتحديد موقع وحجم ووقت وقوع الزلزال. تم اقتراح نموذج التعلم الآلي للتنبؤ بالزلازل في العراق باستخدام مصدرين: الأول هو كتالوج البيانات من عام 1900 إلى عام 2019، والذي يتضمن 36663 زلزالاً، والثاني من هيئة المسح الجيولوجي الأمريكية لمدة عام واحد من عام 2022 إلى عام 2023، والذي يتضمن 25000 زلزال. الزلازل. تم إجراء المعالجة الأولية للبيانات وإزالة القيم المتطرفة ودمج بيانات التاريخ والوقت في الطوابع الزمنية، وتم تحديد الميزات الخمس المهمة للتنبؤ وتم تقسيم البيانات إلى 80% للتدريب و20% للاختبار. وبعد تطبيق عدة محاولات في استخدام نماذج مختلفة تم الحصول على أفضل النتائج باستخدام NN وكانت الدقة حوالي 0.7. وأهم سبب لهذه النتيجة هو التدريب على مدى سنوات طويلة قد تتغير جيولوجياً. وقارنت الدراسة نتائجها مع دراسات أخرى للتنبؤ بالزلازل في مناطق مختلفة من العالم.

Citation: N. B. Jarah et al., J. Basrah Res. (Sci.) **49**(2), 94 (2023).
DOI:https://doi.org/10.56714/bjrs.49.2.10

***Corresponding author email :** nadabadrjarah@yahoo.com

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.iobrs.edu.iq