Earthquake Magnitude Prediction Using Machine Learning Techniques

Conference Paper · April 2024

DOI: 10.1109/IATMSI60426.2024.10502770

6 authors, including:

Fardin Ahmed
North South University
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Shapla Akter
North South University
7 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Jayed Bin Harez
North South University
2 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Riasat Khan
North South University
96 PUBLICATIONS **1,311** CITATIONS
SEE PROFILE

2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3503-6052-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IATMSI60426.2024.10502770

Earthquake Magnitude Prediction Using Machine Learning Techniques

Fardin Ahmed
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
fardin.ahmed01@northsouth.edu

Shapla Akter
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
shapla.akter@northsouth.edu

SM Minhazur Rahman
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
minhazur.rahman3@northsouth.edu

Jayed Bin Harez
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
jayed.harez@northsouth.edu

Amatullah Mubasira
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
amatullah.mubasira@northsouth.edu

Riasat Khan
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
riasat.khan@northsouth.edu

***Abstract*—This research investigates machine learning techniques to predict earthquake magnitude scales. The purpose of training machine learning algorithms on seismic data is to create models that can reliably predict the continuous output earthquake magnitude. A dataset from the United States Geological Survey has been used in this work. The research aims to increase our insight into seismic events and help us better prepare for disasters. Various machine learning algorithms have been applied in this work, such as Decision Tree, KNN, Random Forest, Gradient Boost, XG Boost, SVM and Ridge Regression. For data pre-processing, the duplicate and null values have been managed first. Specifically, a feature named "nst" has been discarded because it had plenty of null values. Next, we replaced the null values with the mean and median imputation techniques. We applied one-hot-encoding and feature scaling techniques as well. The findings illustrate the potential of machine learning in earthquake prediction, providing significant insights for gauging earthquake intensity and limiting its damage. SVM with optimized hyperparameters achieved a root-mean-squared error of 0.10 and 0.93 coefficient of determination (R^2). The method used to forecast the attribute underwent a thorough investigation, followed by a data analysis that revealed information that might be used to lessen the effects of an earthquake in the future.**

Keywords—magnitude estimation, seismic data analysis, seismic event forecasting, performance evaluation, prediction model.

I. INTRODUCTION

An earthquake is one of the most dangerous natural catastrophes, mainly because there is rarely advance notice and little opportunity to prepare [1]. Despite the increasing scientific interest, the possibility of earthquake prediction with sufficient precision remains dubious [2]. Earthquakes are a standard feature of the planet's geology since they occur during the movement of tectonic plates underneath the earth's surface [3]. The following are typical earthquake causes and factors: Plate Boundaries, Faults, Depth, Magnitude, and Geological Conditions. Earthquakes are a common phenomenon, affecting large numbers of people worldwide. Mexico recently experienced a 7.6 magnitude earthquake, but the death toll is negligible. Bangladesh has had 284 earthquakes in the last 40 years, and Chittagong faced a 6.2-magnitude earthquake in 2021.

Sadly, earthquakes are natural, but people can reduce earthquake effects and increase safety for impacted people, such as enforcing strict building codes and constructing structures to resist seismic activity according to standards [4]. Although people cannot prevent earthquakes from occurring, they may take actions to lessen their effects and increase safety for impacted people.

Machine learning techniques have been employed to identify potential earthquake precursors and forecast the likelihood of future seismic activity by analyzing and interpreting complicated patterns in big datasets. Additionally, these automatic systems can offer early warning systems to warn communities of a possible threat and give them time to take preventative action.

Machine learning-based earthquake prediction has become a promising strategy to improve the effective comprehension of seismic activity and reduce potential hazards. This paragraph provides a brief discussion of recent articles related to automatic earthquake and seismic activity detection. For instance, Gaba et al. [5] predicted earthquakes using machine learning models. The authors used fuzzy analysis and artificial neural networks to expect earthquake damage using a seismic parameter. The dataset had 1,038,900 records, with the test dataset having 421,175 and the training dataset having 617,725. The applied decision tree and SVM techniques accomplished F1 scores of 0.85 and 0.89, respectively. Mallouhy and his colleagues [6] used several machine learning algorithms to predict whether an incident is categorized as a negative or positive earthquake. The authors used eight different algorithms. Several hyperparameters have also been evaluated for each chosen model. Prediction results were fairly compared using several parameters, resulting in a reliable forecast of three key events. The method that performs best in terms of accuracy is Random Forest, which, with a score of 0.769, is quite close to KNN's (0.755), MLP's (0.748), and SVM's (0.748). Chelidze and his co-authors [7] sought to reduce the imbalanced data for M>3.5 and used ML techniques to predict the likely course of the following seismic activity event. The researchers employed hydrodynamic, magnetic, and tidal measurements from the day before the seismic event to anticipate using deep learning SVM and Decision Tree techniques. Synthetic oversampling approach, SMOTE is used to oversample/balance the seismic to aseismic days ratio dataset. Matthews correlation coefficient (MCC) has been used to assess the randomization of the dataset. To evaluate the forecasting potential, "sklearn," a machine learning

algorithm from the decision tree classifier library, was used. MCC of the SVM model after the randomization process was 0.17. With two incorrect forecasts, 12 instances were detected correctly out of 14 events. Johnson and his team [8] predicted laboratory earthquakes using machine learning algorithms. The authors employed an open-source dataset. Their Training and Testing Data were 'Galaxy Images' and 'Star Images'. They used 40,000 postage stamps png-files for both 'Galaxy' and 'Star' images in their training dataset. Moreover, the testing dataset used 60,000 postage stamps 'png' files for 'Galaxy' and 'Star' images. They used various ML models, and the Random Forest model attained a maximum accuracy of 0.90. Chittora and his teammates [9] tried to predict the earthquake's magnitude range before the event occurred. Multiple USGS earthquake datasets were used in their work. They have used six classifier models, including 26,978 instances with fifteen attributes. The XG Boost tree achieved the most remarkable accuracy of 92.53% after implementing multiple classifier models.

In this paper, machine learning techniques have been used to predict magnitude of earthquake events. Seven supervised machine learning methods have been applied to analyze seismic data and extract valuable patterns and features. We sought to create predictive algorithms to recognize precursory seismic signals and gauge the likelihood of upcoming earthquake events by training them on past earthquake data. Our study aims to contribute to the area by improving earthquake prediction skills and supporting proactive disaster management and mitigation steps through machine learning algorithms. Hyperparameters of the applied machine learning models have been optimized. LIME-based explainable AI technique has been implemented to interpret the prediction results.

Using the necessary tables, figures, or flowcharts, we display relevant data as we explore the proposed system in detail in Section II. The methodology, including the pre-processing procedures, feature extraction methods, and machine learning algorithms used, are thoroughly described in this section. Section III provides statistical measures, evaluation metrics, and visualizations to improve the presentation of the results.

We summarize the paper by summarizing our work's main conclusions and contributions in Section IV. We also discuss possible directions for future development and improvement in this area. These suggestions for further research guide researchers looking to build on our discoveries and overcome the constraints and challenges identified throughout our investigation.

II. Proposed System

A. Dataset

The dataset for earthquake prediction using machine learning is a significant resource that provides a wide range of seismic event information. It usually includes information like earthquake magnitudes, depths, localities, and timestamps. The US earthquake dataset has been collected from the United States Geological Survey (USGS) online repository [10], over the preceding ten years (2013-2023). The initial dataset comprises a total of 2,013 instances and 22 features with the continuous output of earthquake magnitude.

TABLE I. SUMMARY OF THE EMPLOYED EARTHQUAKE MAGNITUDE DATASET

	latitude	longitude	depth	mag	nst	gap	dmin	RMS
Mean	35.4	139	51.3	4.76	106	101	1.54	0.81
Std	2.47	3.13	64.9	0.35	87.4	36.9	0.95	0.23
min	30.9	130	1.61	4.5	12	10	0.04	0.19
25%	33.3	139	12.6	4.5	53	76	0.82	0.65
75%	37.3	141	56.9	121	127	2.11	0.94	7.3
max	42.1	142	545	7.3	535	254	19.7	1.51

Table I provides the number of non-empty values, the mean (average) value, the standard deviation, the 25th percentile, the 50th percentile, the 75th percentile, and the maximum value of the numerical features of the employed earthquake magnitude dataset. According to this table, the number of seismic stations features comprises approximately 88.5% null values.

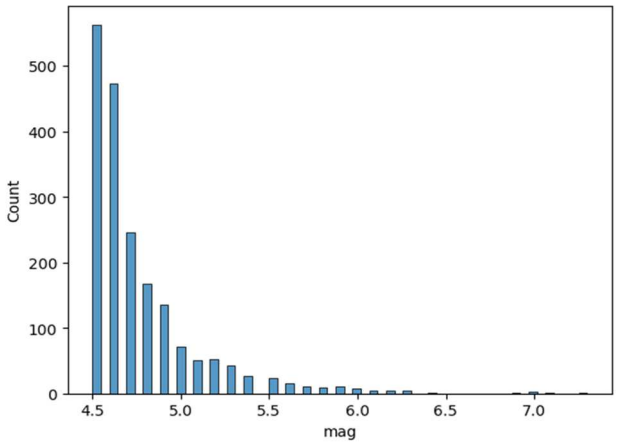

Fig. 1. Histograms of the earthquake magnitude.

The histograms of the earthquake magnitude against count have been illustrated in Fig. 1. Most of the earthquake instances lie between magnitudes of 4.5 and 5.5 with mean magnitude of 4.76.

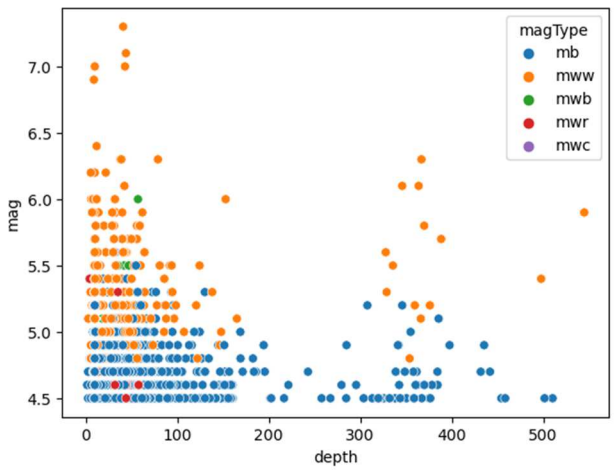

Fig. 2. Scatterplot for magnitude and depth of various earthquake types.

From Fig. 2, we can observe that the short period earthquake (mb) has the highest value and body wave (mwb) has the lowest earthquake magnitude.

B. Dataset Preprocessing

The employed USGS dataset used in this work is extensive and contains duplicate and missing values. Initially, the missing value or null value for each feature is determined [13]. We eliminate the feature with the highest null value. Then we use mean imputation [14] for the missing values for the remaining attributes. IQR and boxplot have been used to detect and manage outliers by removing them. Then we apply features selection to get rid of the unnecessary features. As MagType is a categorical feature, we apply a one-hot encoding technique for MagType.

The equation for mean imputation is relatively straightforward. Let us consider a dataset with a feature or variable X that contains missing values. The mean of the available data for that feature is calculated as:

$$X_{mean} = \frac{1}{2}\sum_{i=1}^{n} x_i \quad (1)$$

where n illustrates the number of non-missing values and x_i represents each non-missing value. The missing values are replaced with the calculated mean. For each missing value, it is substituted with the mean as:

$$X_{imputed} = X_{mean} \quad (2)$$

By performing mean imputation, the missing values are replaced with the average value of the available data for that feature. It is a simple method but may only sometimes capture the accurate underlying patterns or variability in the data.

C. Machine Learning Models

1) Decision tree: The decision tree [15] is a part of supervised machine learning models and effectively presents an organizational structure. So that the trained model may have accurate accuracy, the data in this particular model will be simultaneously divided based on their pertinent parameters. The Decision Tree model has been evaluated from "sklearn.tree" by importing "DecisionTreeRegressor", yielding a test mean squared error (MSE) of 0.087, suggesting a moderate error level in the model's predictions. The test's root mean squared error (RMSE), representing the average size of the prediction mistakes, was 0.298. The test gives the average absolute difference between the anticipated and actual values for mean absolute error (MAE), which was 0.184. Furthermore, the model explains roughly 26.9% of the variation in the test data, according to the test R2 value of 0.269.

2) KNN: The K nearest neighbor model is one of the most intriguing models in machine learning since it can accumulate similar types of data available and classify those according to those similarities [16]. It is also among the most fundamental models. Importing "KNeighborsRegressor" allowed the K-Nearest Neighbors (KNN) model to be evaluated from "sklearn.neighbors" evaluating the model's test mean squared error (MSE) value of 0.073 shows a moderate amount of prediction error. The test's root represented the average size of the prediction error mean squared error (RMSE), which was 0.271. Between projected and actual values, there was an average absolute difference of 0.199, according to the test's mean absolute error (MAE). Furthermore, the model explains about 37.9% of the variation in the test data, according to the test R2 value of 0.379.

3) Random Forest: The bagging ensemble technique, which creates a specific number of decision trees, includes Random Forest. A random forest classifier employs several distinct decision trees for training purposes. Additionally, a few chosen trees will also contribute to the result.

4) Gradient Boost: Gradient boost is related to greed function approximation. It combines previous models to minimize error.

5) XG Boost: XG Boost stands for extreme gradient boosting. It conducts a parallel process of decision tree boosting.

6) SVM: SVM is a robust machine learning model for classification and regression problems. It works by producing an ideal hyperplane. SVM effectively deals with high-dimensional data, can handle linear and non-linear issues using kernel functions, and is insensitive to overfitting because of the margin maximization objective.

7) Ridge Regression: Ridge regression is a linear regression technique incorporating a regularization term to mitigate overfitting and address multicollinearity in the data. Ridge regression reduces the effect of linked predictor variables by shrinking the coefficient estimates towards zero by including a penalty component in the least squares objective function. This regularization helps improve the stability and generalization ability of the model, making ridge regression a valuable tool in scenarios with high dimensionality or collinearity among predictors.

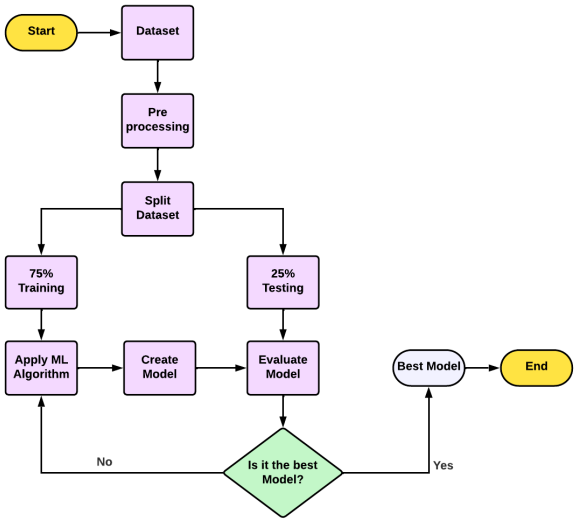

Fig. 3. Working sequences of the proposed earthquake prediction using machine learning techniques.

Working sequences of the proposed earthquake prediction using machine learning have been depicted in Fig. 3. This work is initiated by collecting data. Then we have to apply the pre-processing techniques for the duplicate and missing or null values, categorical features, etc. After finishing pre-processing techniques, the dataset has been split into 75% for training and 25% for testing facilitating the stratifying preference. Then the ML algorithms have been applied to create and evaluate the model for training data with enabling the hyperparameter optimization techniques. Next, the applied models are evaluated using test data. This process will terminate after getting the best model with the maximum score.

III. Results and Discussion

This section discusses the simulation results of the proposed machine learning-based automatic earthquake magnitude prediction system. Using the default hyperparameter, we train our dataset in Jupyter Notebook with seven regressor models. It provides us with performance metrics, among which XG Boost and Gradient Boost models perform the best with less error percentage and least accurate performance by a decision tree. After that, hyperparameter optimization- GridSearchCV and RandomizedSearchCV have been used to get the best parameter to train the model and achieve better results.

TABLE II. HYPERPARAMETER VALUES' RANGES FOR VARIOUS ML MODELS

Model	Hyperparameter Value Range	Optimized value
Decision Tree	splitter: ['best', 'random'], max_depth: [10-50], min_samples_split: [2,5,8,10], min_samples_leaf: [1,2,4,6].	splitter: 'random', max_depth: 27, min_sample_split: 10, min_sample_leaf: 6
KNN	n_neighbours: [200-1000], weights: ['uniform, 'distance'], leaf_size: [2,5,8,10], metric: ['euclidean', 'manhattan', 'chebyshev']	n_neighbours: 368, weights: 'distance' leaf_size 5, metric: 'euclidean'
Random Forest	n_estimators: [100-800], max_features: ['auto', 'sqrt', 'log2'], max_depth: [10-100], min_samples_split: [2,5,8,10], min_samples_leaf: [1,2,4,6]	n_estimators: 578, max_features: 'sqrt', max_depth: 40, min_samples_split: 2, min_samples_leaf: 1
Gradient Boost	n_estimators: [100-800], max_features: ['auto', 'sqrt', 'log2'], max_depth: [10-100], learning_rate: [0.001, 0.1, 0.25, 0.5, 0.3]	n_estimators: 911, max_features: 'sqrt', max_depth: 40, learning_rate: 0.1
XG Boost	n_estimators: [200-800], max_features: ['auto', 'sqrt', 'log2'], max_depth: [10-100], learning_rate: [0.001, 0.1, 0.25, 0.5, 0.3]	n_estimators: 284, max_features: 'sqrt', max_depth: 40, learning_rate: 0.1
SVM	C: [2-10], gamma: [0.1-1]	C: 8.88, gamma: 0.13
Ridge Regression	alpha: loguniform (1, 100), fir_intercept: [True, False], solver: ['svd', 'cholesky', 'lsqr', 'sag'], n_estimators: [1, 50]	alpha: 1.22, fir_intercept: True, solver: 'svd'

Table II illustrates the hyperparameter values' ranges and the corresponding optimized hyperparameters (obtained from hyperparameter optimization) for all the ML models.

TABLE III. PERFORMANCE METRICS OF VARIOUS ML MODELS WITH DEFAULT HYPERPARAMETERS

Model	**MAE**	**MSE**	**RMSE**	**R-squared (R^2)**
Decision Tree	0.18	0.08	0.29	0.26
Random Forest	0.13	0.04	0.20	0.65
Gradient Boost	**0.12**	**0.03**	**0.18**	**0.75**
XG Boost	0.13	**0.03**	**0.18**	**0.75**
KNN	0.19	0.07	0.27	0.38
SVM	0.17	0.07	0.26	0.51
Ridge Regression	0.14	0.05	0.22	0.66

Performance metrics of various ML models with default hyperparameters have been illustrated in Table III. According to this table, the Gradient Boost model accomplished the best performance.

TABLE IV. PERFORMANCE METRICS OF VARIOUS ML MODELS WITH OPTIMIZED HYPERPARAMETERS

Model	**MAE**	**MSE**	**RMSE**	**R-squared (R^2)**
Decision Tree	0.16	0.05	0.23	0.54
Random Forest	0.12	0.03	0.19	0.69
Gradient Boost	0.12	0.03	0.18	0.76
XG Boost	0.13	0.03	0.19	0.75
KNN	0.18	0.08	0.28	0.34
SVM	**0.09**	**0.01**	**0.10**	**0.93**
Ridge Regression	0.14	0.05	0.22	0.66

Performance metrics of various ML models with optimized hyperparameters have been illustrated in Table IV. The SVR model with optimized $C = 8.88$ and $gamma = 0.13$ achieved the lowest errors.

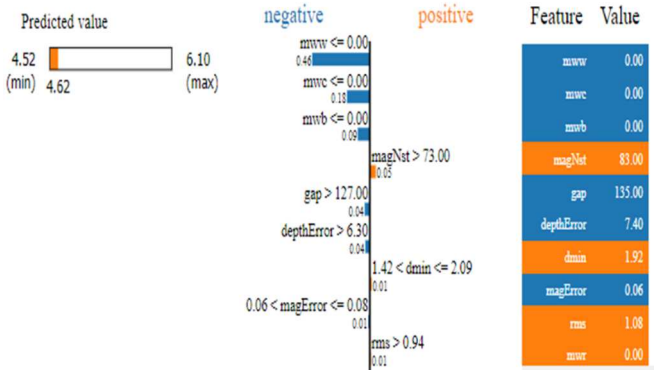

Fig. 4. Machine learning model prediction interpretation by LIME explainable AI library.

Fig. 4 provides a LIME explainable AI library that interprets the Gradient Boost machine learning model predictions. According to this figure, the model predicts an earthquake magnitude of 4.62 for the specific sample.

TABLE V. COMPARISON OF THE PROPOSED SYSTEM WITH EXISTING WORKS

Ref.	Model	RMSE	Other metrics
[5]	Random Forest	0.358	R-Squared = 0.75127
[6]	Logistic Regression	0.316	MAE = 0.3851
[12]	BiLSTM	0.383	MAE = 0.28, MSE = 0.15
This work	SVM	0.10	MAE = 0.09, R-Squared = 0.93

Table V illustrates the comparison of the proposed earthquake prediction system with other existing works.

IV. Conclusions

This work uses various machine learning approaches to predict earthquake magnitudes by observing patterns in seismic data. An open-source USGS seismic dataset with suitable preprocessing techniques has been used. These methods could help us learn more about earthquakes and make earthquake predictions more accurate. However, the success of machine learning models depends on the dataset, features, and methods used. Even though they can find critical seismic features and set up early warning systems, it is essential to remember that earthquake modelling is complicated and has limits. Due to the lack of information and many different things that cause earthquakes, machine learning models may need help picking up on rare or never-before-seen seismic events. Machine learning and geophysics experts must keep

researching and working together to move the field forward and make earthquake forecast models more useful and accurate. In future, it would be valuable to explore incorporating additional data sources, such as geospatial and geological data, to improve the accuracy and robustness of the models. Additionally, developing ensemble models that combine multiple machine learning algorithms or techniques could be beneficial for enhancing prediction performance and capturing the complexities of seismic activity.

REFERENCES

[1] A. Volvach *et al*., "Statistical precursors of a strong earthquake on April 6, 2009 on the Apennine Peninsula," Heliyon, vol. 8, 2022.

[2] S. Cho, J. K. Ahn and E. H. Hwang, "Optimization of Network-Based Earthquake Early Warning Systems on the Korean Peninsula," IEEE Access, vol. 10, pp. 83931-83939, 2022.

[3] Z. Jiang, "Attention Behavior Evaluation during Daily Living Based on Egocentric Vision," Journal of Advances in Information Technology, vol. 8, pp. 126-134, 2017.

[4] E. Cochran, J. Lawrence, C. Christensen and A. Chung, "A novel strong-motion seismic network for community participation in earthquake monitoring," IEEE Instrumentation & Measurement Magazine, vol. 12, pp. 8-15, 2009.

[5] A. Gaba, A. Jana, R. Subramaniam, Y. Agrawal, and M. Meleet, "Analysis and Prediction of Earthquake Impact-a Machine Learning approach," Computational Systems & Information Technology for Sustainable Solution, pp. 1-5, 2019.

[6] R. Mallouhy, C.A. Jaoude, C. Guyeux and A. Makhoul, "Major earthquake event prediction using various machine learning algorithms," International Conference on Information and Communication Technologies for Disaster Management, pp. 1-7, 2019.

[7] T. Chelidze, T. Kiria, G. Melikadze, T. Jimsheladze and G. Kobzev, "Earthquake Forecast as a Machine Learning Problem for Imbalanced," Frontiers in Earth Science, pp.1-11, 2022.

[8] P. A. Johnson *et al*., "Laboratory earthquake forecasting: A machine learning competition," Proceedings of the National Academy of Sciences, vol. 118, pp. 1–10, 2021.

[9] P. Chittora *et al*., "Experimental analysis of earthquake prediction using machine learning classifiers, curve fitting, and neural modelling," Internet Archive Scholar, pp. 1-25, 2022.

[10] USGS. Accessed on: July 10, 2023. [Online]. Available: https://earthexplorer.usgs.gov.

[11] M. N. I. Suvon, S. C. Siam, M. Ferdous, M. Alam and R. Khan, "Masters and doctor of philosophy admission prediction of Bangladeshi students into different classes of universities," International Journal of Artificial Intelligence, vol. 11, pp. 1545-1553, 2022.

[12] E. Abebe, H. Kebede, M. Kevin and Z. Demissie, "Earthquakes magnitude prediction using deep learning for the Horn of Africa," Soil Dynamics and Earthquake Engineering, vol. 170, 2023.

[13] N. E. J. Asha, E. U. Islam and R. Khan, "Low-Cost Heart Rate Sensor and Mental Stress Detection Using Machine Learning," International Conference on Trends in Electronics and Informatics, pp. 1369-1374, 2021.

[14] N. N. Prachi, M. Habibullah, E. H. Rafi, E. Alam and R. Khan, "Detection of fake news using machine learning and natural language processing algorithms," Journal of Advances in Information Technology, vol. 13, pp. 652-661, 2022.

[15] R. Siddiqua, N. Islam, J. F. Bolaka, R. Khan and S. Momen, "AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students," Array, vol. 18, 2023.

[16] G. Aurélien, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.," O'Reilly Media, Inc., 2017.