Earthquake Magnitude Estimation from Seismological and Geographical Data Using Machine Learning

Rafiad Zaman Khan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rafiad.zaman.khan@g.bracu.ac.bd

Kazi Wahidul Islam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
kazi.wahidul.islam@g.bracu.ac.bd

Rifat Mahmud Tamim
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rifat.mahmud.tamim@g.bracu.ac.bd

***Abstract*—Earthquakes are among the most destructive natural disasters, yet their complex dynamics make accurate prediction extremely challenging. Reliable estimation of earthquake magnitude prior to or immediately after seismic events is crucial for mitigating risks, enhancing disaster preparedness, and reducing socio-economic losses. This study, titled "Earthquake Magnitude Prediction Using Machine Learning," investigates the application of multiple machine learning approaches for modeling and predicting earthquake magnitudes using seismological and geospatial parameters. The dataset was systematically preprocessed through missing value imputation, categorical encoding, and outlier removal employing both interquartile range and Z-score methods, followed by feature normalization to ensure robustness. Several regression models were implemented, including Random Forest Regressor, Support Vector Regression (SVR and LinearSVR), K-Nearest Neighbors (KNN), XGBoost Regressor, and a Multilayer Perceptron (MLP) neural network. Model performance was rigorously evaluated using the coefficient of determination (R^2), Root Mean Square Error (RMSE) and Mean Squared Error (MAE). Experimental results demonstrate that ensemble-based methods and neural networks consistently outperform traditional regressors, effectively capturing the nonlinear and multidimensional relationships among seismic features such as depth, latitude, longitude, signal strength, and ground intensity indices. The findings underscore the promise of machine learning for advancing earthquake magnitude prediction, offering valuable insights for the integration of intelligent predictive systems into seismic risk assessment frameworks and early-warning infrastructures.**

***Index Terms*—Earthquake Magnitude Prediction, Machine Learning, Seismic Data Analysis, Random Forest, Support Vector Regression, Multilayer Perceptron**

I. INTRODUCTION

Earthquakes represent some of the most catastrophic geological events, capable of triggering massive human and infrastructural losses along with prolonged socio-economic disruption. Despite advancements in seismology, accurately predicting the magnitude of earthquakes remains a formidable scientific challenge due to the inherently complex and nonlinear nature of tectonic processes. Traditional seismological models often rely on historical records, fault-line analysis, and geophysical simulations, which, while valuable, are limited in their ability to capture the intricate interactions among multiple seismic parameters. Consequently, there is an urgent need for methodologies that can model the multifaceted dependencies inherent in earthquake data and provide more reliable predictions.

Recent developments in machine learning (ML) offer promising avenues for enhancing earthquake prediction. ML techniques are capable of learning complex, nonlinear relationships from large, multidimensional datasets, making them particularly well-suited for seismic analysis [1]. By leveraging historical seismic records, geospatial information, and signal measurements, machine learning models can uncover latent patterns that traditional regression or statistical models may fail to identify. The integration of ML approaches into seismology has the potential not only to improve predictive accuracy but also to contribute to early-warning systems and disaster preparedness strategies, thereby mitigating the devastating impacts of seismic events [2].

This study investigates the application of several machine learning algorithms to estimate earthquake magnitudes based on a diverse set of seismological and geospatial features. The research emphasizes rigorous data preprocessing, including handling of missing values, categorical encoding, outlier treatment using interquartile range (IQR) and Z-score methods, and feature normalization, to enhance model reliability and generalization. A variety of regression models were implemented, including Random Forest Regressor, Support Vector Regression (SVR and LinearSVR), K-Nearest Neighbors (KNN), XGBoost Regressor, and a Multilayer Perceptron (MLP) neural network, with model performance evaluated using standard metrics such as the coefficient of determination (R^2), Root Mean Square Error (RMSE) and Mean Squared Error (MAE).

The research aims to identify the most effective machine learning strategies for capturing the nonlinear and multidimensional dependencies among critical seismic features, including earthquake depth, latitude, longitude, signal strength, and ground intensity indices. By systematically comparing the per-

formance of ensemble-based, neural network, and traditional regression models, this study contributes to the growing body of knowledge on data-driven earthquake prediction and highlights the practical utility of machine learning in enhancing seismic risk assessment frameworks. Ultimately, the findings of this research are expected to inform the development of intelligent predictive systems that can support early-warning mechanisms and improve disaster management policies globally.

II. LITERATURE REVIEW

The application of machine learning (ML) to earthquake prediction has gained considerable traction in recent years, driven by the limitations of traditional statistical and physical models in forecasting seismic events. This section reviews key research studies illustrating the evolution of machine learning techniques in seismic forecasting, with a focus on their methodologies, findings, and limitations.

Ahmed et al. [3] explored the use of machine learning to predict earthquake magnitudes using past data from the USGS Earthquake Database. The data was cleaned and prepared for training, examining how features like location, depth, and time might be related to earthquake magnitude. Various machine learning models were tested, including linear regression, decision tree regression, random forest regression, K-nearest neighbors (KNN), and support vector regression (SVR). The results showed that random forest, an ensemble model, provided the best accuracy and lowest error, indicating that predicting magnitudes involves complex patterns rather than simple linear relationships. The authors conclude that machine learning can identify useful patterns in earthquake data, but accuracy depends on data quality, the study area, and available data. They suggest adding more geological and physical information to improve future predictions.

Asim et al. [4] explored the use of Long Short-Term Memory (LSTM) neural networks to predict earthquakes using past data. The authors used earthquake records from the Japan Meteorological Agency (JMA), which included date, time, location, depth, and magnitude. They organized the data into sequences and normalized the numbers to work with the LSTM model. The target area was divided into grids and each time series was treated as a time series. The LSTM model was trained to predict if an earthquake would occur in each grid cell based on previous patterns. The results showed that the LSTM model performed better than traditional methods, but predicting large earthquakes was still challenging. The authors suggest that adding more geological and geophysical data could improve future predictions.

Narayanakumar and Raja [5] presented a method for predicting earthquake magnitudes in the Himalayan region using a Back Propagation Artificial Neural Network. Researchers gathered earthquake data from 1887 to 2015 from sources like the US Geological Survey and the Indian Meteorological Department. They transformed this data into eight mathematically computed seismicity indicators, which served as input features for a three-layer feed-forward BP neural network trained using the Levenberg–Marquardt algorithm. The network performed well for earthquakes between magnitudes 3.0 and 5.4, with reduced accuracy for larger magnitudes due to limited training examples. The study demonstrates that BP-ANN can effectively capture complex, non-linear seismic activity patterns, offering better prediction accuracy than conventional statistical approaches for small to moderate earthquakes.

Galkina and Grafeeva [6] reviewed recent research on using machine learning to predict earthquakes. The authors highlighted the historical difficulty of predicting earthquakes and the skepticism surrounding traditional precursors. The paper argues that modern machine learning techniques, such as neural networks and other classifiers, are a promising new approach. The survey discusses various studies that have applied different models, including Artificial Neural Networks (ANN) and Probabilistic Neural Networks (PNN) to predict earthquake magnitudes, and models like Random Forest (RF) and Support Vector Machines (SVM) for both classification and regression tasks. It also mentioned more complex ensemble methods like LPBoost and a system combining a Support Vector Regressor (SVR) with a Hybrid Neural Network (HNN). The authors concluded that these diverse machine learning models show great potential for creating more accurate and timely earthquake predictions by uncovering complex, non-linear patterns within seismic data.

Abdul Salam et al. [7] introduced two hybrid machine learning models to predict earthquake magnitudes in Southern California over a 15-day period. The models, FPA-ELM (Flower Pollination Algorithm and Extreme Learning Machine) and FPA-LS-SVM (Flower Pollination Algorithm and Least Square Support Vector Machine), used seven seismic indicators calculated from historical earthquake data. The study compared the performance of these hybrid models with their non-hybrid counterparts using four evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and Percent Mean Relative Error (PMRE). The results demonstrated that the FPA-LS-SVM model delivered superior prediction accuracy, suggesting that combining optimization algorithms with machine learning can enhance earthquake prediction.

Yavas et al. [8] achieved a remarkable 97.97% accuracy in predicting the maximum earthquake category in Los Angeles within a 30-day period. This was accomplished by utilizing a Random Forest machine learning model and a meticulously crafted feature matrix based on data from the Southern California Earthquake Data Center (SCEDC). The study, which builds on previous work in other seismic regions, demonstrates a significant improvement over a prior Los Angeles prediction accuracy of 69.14%. The researchers meticulously processed data from January 2012 to September 2024, converting various magnitude types to a consistent local magnitude (ML) for model training. Their findings underscore the immense potential of machine learning and neural networks to enhance seismic risk management and preparedness in highly active regions. The study's success is attributed to a robust method-

ology, including the selection of an optimal predictive model and a comprehensive dataset, setting a new benchmark for earthquake forecasting.

Mondol's [9] article explored the use of machine learning techniques to predict earthquake magnitude and depth range using real-life seismic data. The study uses USGS data from significant earthquakes between 1965 and 2016, focusing on key attributes like latitude and longitude. Four models (Random Forest, Linear Regression, Polynomial Regression, and Long Short-Term Memory) are trained and their performance measured using indicators like R^2 score, explained variance, and mean squared error. The results show that earthquake magnitude is not easily predicted, with Polynomial Regression of level 16 being the best overall model for predicting magnitude. Random Forest was the best in predicting depth with an R2 score of 0.8574. The study also identified patterns such as frequent magnitude ranges, outliers, and global spatial distribution of earthquakes. Although accurate prediction of earthquakes is yet to be determined, machine learning offers viable approximations, particularly in detecting depth patterns.

Ridzwan and Yusoff [10] provides a comprehensive overview of machine learning (ML) methods for predicting earthquakes, highlighting their limitations and future directions. The study identifies three main tasks: earthquake detection, parameter estimation, and forecasting. It discusses various algorithms like Support Vector Machines, Artificial Neural Networks, Decision Trees, Random Forests, and Deep Learning architecture, and their applications, advantages, and disadvantages. The paper emphasizes the importance of quality seismic data and feature engineering, considering geological and geophysical parameters. However, the paper also notes limitations in generalizability due to seismic activity differences and insufficient data. The authors suggest hybrid methods combining physics-based models with data-driven ML techniques and the integration of deep learning and big data. They also emphasize the morality and social dimension of prediction systems and advise against reckless communication of findings to prevent panic and misinformation.

Mallouhy et al. [11] presented a machine learning-based method for predicting earthquake magnitude using historical data from various global sources. The authors use multiple machine learning models, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor, to forecast earthquake magnitude using parameters like latitude, longitude, depth, and time of occurrence. The performance of these models is evaluated using indicators like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R 2 score. The results show that ensemble methods, specifically Random Forest and Gradient Boosting, have a higher predictive score and generalization capability compared to simpler models. The research also explores the spatial coverage of earthquakes to identify the most frequent areas and the relationship between input variables. The authors conclude that machine learning can improve earthquake magnitude forecasts, but issues still arise due to the chaotic nature of earthquake phenomena.

III. DATASET DESCRIPTION

The dataset employed in this study consists of 1000 earthquake events collected from different regions worldwide. Each record provides a rich set of 19 attributes, encompassing both seismic measurements and contextual information. The primary variable of interest is the earthquake magnitude, which serves as the target for predictive modeling. Alongside magnitude, the dataset includes essential seismological parameters such as depth, latitude, and longitude, which describe the physical characteristics and spatial distribution of the seismic events.

To capture the perceived intensity of earthquakes, two crowd- and observation-based measures are included: the Community Internet Intensity (cdi) and the Modified Mercalli Intensity (mmi). These variables provide insights into the severity of earthquakes as experienced by populations. The dataset further includes the number of reporting stations (nst), significance value (sig), and seismic gap (gap), which together characterize the reliability and quality of the recorded event. An alert level attribute, along with a binary tsunami indicator, provides information on hazard assessment.

In addition, categorical variables such as the seismic network code (net) and magnitude type (magType) are present, offering further classification of the recorded events. Contextual information is provided through descriptive attributes including the event title, reported location, continent, and country. Temporal characteristics are captured by the date and time of occurrence, allowing for chronological analysis of seismic activity.

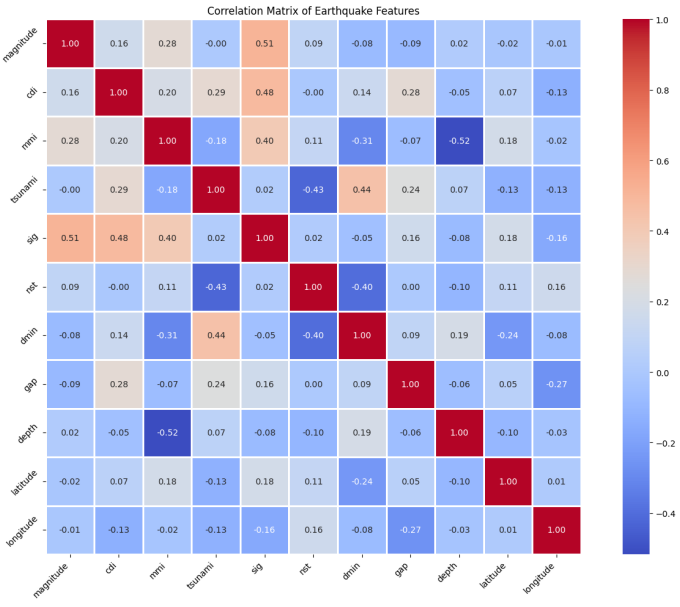

Fig. 1. Correlation Matrix of Earthquake Features

Overall, this dataset integrates quantitative seismic measurements with qualitative contextual information, offering a comprehensive representation of global earthquake events. Its balanced mix of physical, geographical, and observational

variables makes it suitable for both exploratory analysis and the development of predictive machine learning models aimed at estimating earthquake magnitude.

IV. DATA PROCESSING

The dataset used in this study was obtained from a publicly available earthquake record repository containing seismic parameters such as magnitude, depth, geographical coordinates, and other significance indicators. Before model development, several preprocessing steps were carried out to ensure quality and consistency. First, irrelevant columns were removed, and missing values in essential features including magnitude, depth, latitude, and longitude were eliminated. The date-time field was converted into a standard format, while categorical variables such as alert level and magnitude type were encoded using one-hot encoding. Outlier detection and removal were performed using two methods to minimize the effect of anomalies: first, by filtering data points outside 1.5 times the Interquartile Range (IQR), and second, by removing any remaining values with a Z-score greater than 3. Finally, the StandardScaler was applied to normalize the feature distributions.

Exploratory Data Analysis (EDA) was also conducted to better understand the dataset. Feature distributions were examined to observe statistical behavior, and a correlation heatmap was generated to reveal relationships between predictors and earthquake magnitude. Based on this analysis, the most relevant seismic parameters — signal strength (sig), number of stations reporting (nst), Modified Mercalli Intensity (mmi), depth, latitude, longitude, and Community Internet Intensity (cdi) — were selected as input features. The target variable for prediction was earthquake magnitude.

V. METHODOLOGY

To develop predictive models, several machine learning algorithms were implemented and compared. The models included Random Forest Regressor, Support Vector Regression (Linear SVR and SVR with an RBF kernel), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), and a Multilayer Perceptron (MLP) Neural Network. The data set was divided into the 80% training and 20% testing subsets. The hyperparameters were individually configured for each algorithm, such as the number of estimators in Random Forest, the type of kernel and the regularization constant in SVR, the learning rate in XGBoost and the hidden layer configurations in MLP.

Model evaluation was conducted using the coefficient of determination (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which together provided a measure of predictive precision and error magnitude. To support visual interpretation, scatter plots comparing predicted versus actual magnitudes were generated for each model, with a diagonal reference line indicating perfect prediction. Furthermore, a precision-tolerance analysis was conducted to evaluate the percentage of predictions falling within various error margins of the actual magnitude, offering deeper insight into model reliability.

Finally, a comparative analysis was carried out across all implemented models. This analysis highlighted differences in predictive accuracy, generalization capability, and robustness, allowing for the identification of the most suitable regression approach for earthquake magnitude prediction.

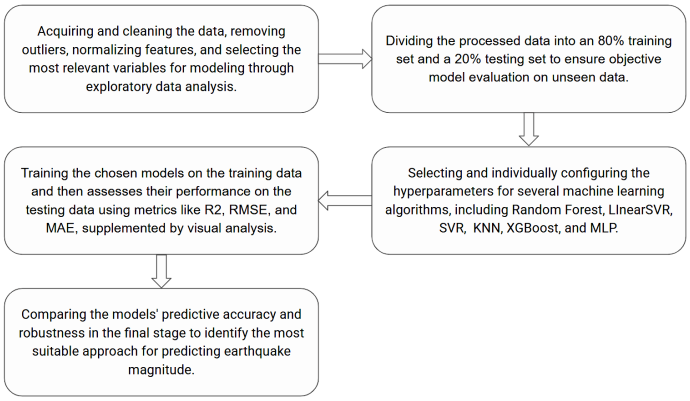

Fig. 2. Methodology

VI. MODEL EVALUATION

A. Random Forest Regressor Performance Analysis

Random Forest is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. The prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble [12]. Each tree is trained on a random subset of the dataset and features, and the final result is obtained by averaging the predictions of all trees. This process reduces overfitting and enhances predictive accuracy, making Random Forest highly effective for regression tasks involving complex, non-linear relationships.

The model was trained with 500 trees (n_estimators = 500), and its performance was assessed using standard evaluation metrics:

- R^2 Score: 0.788. This indicates that approximately 78.8% of the variance in earthquake magnitudes is explained by the model.
- Root Mean Squared Error (RMSE): 0.140. This represents the average prediction error in magnitude units, where lower values indicate better accuracy.
- Mean Absolute Error (MAE): 0.065. This shows that, on average, the predictions deviate from the actual values by only 0.065 magnitude units.

Overall, the Random Forest model achieved strong predictive performance, explaining most of the variability in earthquake magnitudes while maintaining low prediction errors. This demonstrates its suitability for accurately estimating earthquake magnitudes from the selected features.

B. Linear Support Vector Regression (LinearSVR) Performance Analysis

Support Vector Regression (SVR) is a machine learning technique that extends the principles of Support Vector Machines to regression tasks, aiming to find a function that approximates the relationship between input variables and continuous output values. SVR employs kernel functions to map input data into high-dimensional spaces, enabling it to model both linear and non-linear relationships effectively. The primary objective is to identify a hyperplane that best fits the data while maintaining a margin of tolerance for errors, thus ensuring robust predictions even in the presence of noise [13].

LinearSVR is a type of Support Vector Regression that models the relationship between input and output using a linear function while minimizing prediction errors within a specified margin [14]. A LinearSVR model assumes a linear relationship between the features and the target variable. The model was trained with a regularization parameter of $C = 5$ and a maximum of 10,000 iterations to ensure convergence.

The model performance was evaluated using three common regression metrics:

- R^2 Score: 0.102. This indicates that only 10.2% of the variance in earthquake magnitudes is explained by the model. This suggests weak predictive power.
- Root Mean Squared Error (RMSE): 0.289. This represents the average prediction error in terms of magnitude units, which is relatively higher compared to the Random Forest model.
- Mean Absolute Error (MAE): 0.130. This shows that, on average, the predictions deviate from the actual values by 0.130 magnitude units.

Overall, the Linear SVR model performed relatively poorly in this context. The low R^2 value and comparatively higher error metrics indicate that the linear assumption may not adequately capture the complex, non-linear relationships present in the earthquake data. This suggests that more flexible models, such as ensemble methods, may be better suited for this prediction task.

C. Support Vector Regression (SVR) Performance Analysis

The SVR model was trained with regularization parameter $C = 5$ and $\gamma = \texttt{auto}$, ensuring flexibility in learning non-linear patterns. The evaluation metrics were as follows:

- R^2 Score: 0.758. This indicates that the model explains 75.8% of the variance in earthquake magnitudes.
- Root Mean Squared Error (RMSE): 0.150. This represents the average magnitude prediction error, with lower values indicating better accuracy.
- Mean Absolute Error (MAE): 0.100. This shows that, on average, predictions deviate from the actual values by 0.100 magnitude units.

Overall, the SVR model with an RBF kernel achieved strong predictive performance, effectively capturing the non-linear relationships present in the earthquake data. While its performance was slightly lower compared to the Random Forest model, it still demonstrated high accuracy and reliability for magnitude prediction.

D. K-Nearest Neighbors Regressor (KNN) Performance Analysis

K-Nearest Neighbors Regressor (KNN) is a non-parametric, instance-based learning algorithm that makes predictions by considering the target values of the k closest training samples in the feature space. For regression tasks, the prediction is typically computed as the average of the neighbors' target values. This method does not assume a specific functional form of the data, making it flexible for capturing local patterns [15].

In this study, the model was trained with $k = 5$ nearest neighbors, meaning that each prediction was determined by averaging the magnitudes of the five most similar data points. The evaluation of the model produced the following results:

- R^2 Score: 0.695. This indicates that 69.5% of the variance in earthquake magnitudes was explained by the model.
- Root Mean Squared Error (RMSE): 0.168. This represents the average magnitude prediction error, where lower values indicate higher accuracy.
- Mean Absolute Error (MAE): 0.124. This shows that, on average, predictions deviated from the actual values by 0.124 magnitude units.

Overall, the KNN regression model achieved a moderate level of predictive performance. While it was less accurate compared to Random Forest and SVR, it still captured meaningful patterns in the data and provided reasonably accurate predictions.

E. Extreme Gradient Boosting (XGBoost) Regressor Performance Analysis

Extreme Gradient Boosting (XGBoost) Regressor is an advanced implementation of gradient boosting that builds an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones [16]. The algorithm minimizes a differentiable loss function using gradient descent, incorporates regularization to prevent overfitting, and supports efficient parallel computation. These characteristics make XGBoost highly effective for handling structured tabular data with complex feature interactions.

In this study, the model was trained with $n\_estimators = 100$ and a learning rate of 0.05. The performance of the model was evaluated using standard regression metrics:

- R^2 Score: 0.752. This indicates that the model explains 75.2% of the variance in earthquake magnitudes.
- Root Mean Squared Error (RMSE): 0.152. This represents the average prediction error in magnitude units, where lower values indicate higher accuracy.
- Mean Absolute Error (MAE): 0.066. This shows that, on average, predictions deviate from the actual values by only 0.066 magnitude units.

Overall, the XGBoost model demonstrated strong predictive performance, closely matching the accuracy of the Random Forest and SVR models. The relatively low error values and high explanatory power suggest that XGBoost is a robust and

reliable method for earthquake magnitude prediction in this context.

F. Multi-Layer Perceptron (MLP) Performance Analysis

Multi-Layer Perceptron (MLP) is a feed-forward artificial neural network consisting of multiple layers of neurons: an input layer, one or more hidden layers, and an output layer. Each neuron applies a weighted sum of its inputs followed by a non-linear activation function, enabling the network to capture complex, non-linear patterns within the data [17].

The implemented MLP architecture consisted of two hidden layers: the first with 64 neurons and the second with 32 neurons, both using the sigmoid activation function. The output layer contained a single neuron with a linear activation function to generate continuous-valued predictions. The model was trained for 400 epochs with a batch size of 32, using the Adam optimizer and mean squared error (MSE) as the loss function.

The model's predictive performance was evaluated using standard regression metrics:

- R^2 Score: 0.625. This indicates that 62.5% of the variance in earthquake magnitudes was explained by the model.
- Root Mean Squared Error (RMSE): 0.186. This represents the average prediction error in magnitude units.
- Mean Absolute Error (MAE): 0.116. This shows that, on average, predictions deviated from the actual values by 0.116 magnitude units.

Overall, the MLP model achieved a moderate level of predictive accuracy. While it captured non-linear relationships present in the dataset, its performance was lower compared to ensemble-based models such as Random Forest and XGBoost, as well as the SVR. This suggests that further optimization of the network architecture, activation functions, or training hyperparameters may be needed to enhance its predictive capability.

VII. RESULTS AND DISCUSSION

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	R^2 Score	RMSE	MAE
Random Forest Regressor	0.788	0.140	0.065
Linear SVR	0.102	0.289	0.130
RBF-SVR	0.758	0.150	0.100
KNN Regressor ($k = 5$)	0.695	0.168	0.124
XGBoost Regressor	0.752	0.152	0.066
MLP Regressor	0.625	0.186	0.116

From Table I, it is evident that the Random Forest Regressor achieved the best overall performance, with the highest R^2 score (0.788) and the lowest RMSE (0.140) and MAE (0.065). This indicates that the ensemble-based Random Forest model effectively captured both linear and non-linear relationships in the data. The XGBoost Regressor and SVR models also performed well, with R^2 scores above 0.75 and relatively low error metrics, demonstrating their ability to model complex patterns.

The scatter plot of predicted vs. actual magnitudes further support these results, showing that Random Forest prediction lies closest to the ideal diagonal line, indicating higher accuracy as seen in (Figure 3). XGBoost, and SVR show close results. In contrast, LinearSVR exhibits wide deviations from the diagonal, confirming its poor predictive performance. The KNN and MLP models show moderate alignment with the diagonal, consistent with their intermediate R^2 and error metrics.

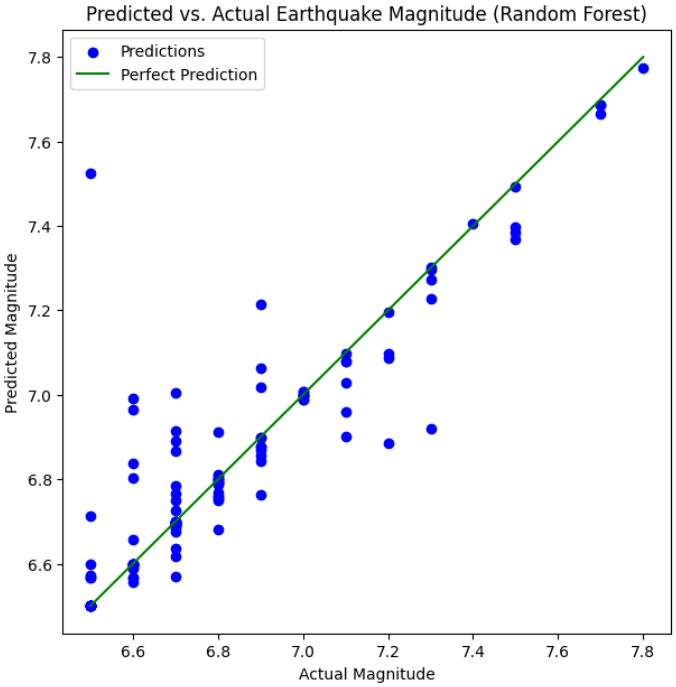

Fig. 3. Predicted vs. Actual Earthquake Magnitude (Random Forest)

Additionally, the precision vs. tolerance level curves illustrate the fraction of predictions within a certain magnitude tolerance. Random Forest maintains highest precision across all tolerance levels, as seen in (Figure 4), followed by XGBoost and SVR, while Linear SVR consistently underperforms. KNN and MLP show moderate precision, reflecting their limited ability to consistently predict close to the true values.

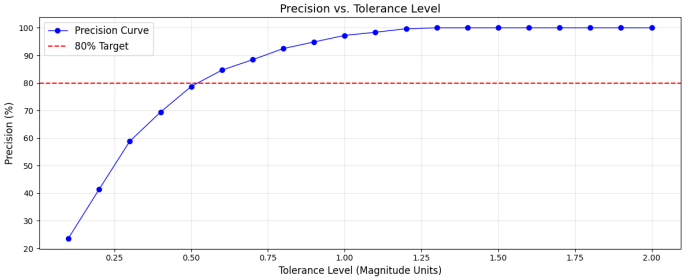

Fig. 4. Precision vs. Tolerance Level Curve (Random Forest)

Overall, ensemble-based models such as Random Forest and XGBoost demonstrated superior robustness and accuracy, as

confirmed by the table, scatter plots, and precision curves. SVR also performs well but slightly lower than the ensembles, whereas Linear SVR is inadequate for this non-linear regression task. KNN and MLP provide reasonable predictions but require careful tuning to match the accuracy of the top-performing models.

VIII. CONCLUSION

This study explored the prediction of earthquake magnitudes using various machine learning models, including Random Forest, Linear SVR, SVR, K-Nearest Neighbors, XGBoost, and Multi-Layer Perceptron (MLP). The objective was to evaluate the effectiveness of these models in capturing the complex relationships between seismic features and earthquake magnitudes.

The experimental results demonstrate that ensemble-based methods, particularly Random Forest and XGBoost, consistently outperformed other models across all evaluation metrics, achieving high R^2 scores and low RMSE and MAE values. Random Forest showcased the best result. This model effectively handled both linear and non-linear relationships in the data, as also confirmed by the scatter plot of predicted vs. actual magnitudes and the precision vs. tolerance level curve. SVR also delivered strong performance, highlighting the advantage of non-linear kernels in capturing intricate feature interactions.

Linear SVR, which assumes a strictly linear relationship between features and the target variable, exhibited the poorest performance, underscoring the non-linear nature of earthquake magnitude prediction. KNN and MLP models achieved moderate results, indicating their ability to capture local patterns and non-linear dependencies, respectively, but their predictive accuracy was limited compared to the top-performing ensemble models. The MLP's moderate performance suggests that further optimization of network architecture, activation functions, and training parameters could improve results.

Overall, the study confirms that ensemble methods are highly robust and accurate for earthquake magnitude prediction using seismic data, while non-linear models like SVR provide valuable alternatives. Instance-based methods (KNN) and neural networks (MLP) can be effective with careful tuning but generally require more sophisticated parameter optimization. Future work may involve incorporating additional geophysical features, exploring hybrid models, or applying advanced deep learning architectures such as LSTM or attention-based networks to further enhance predictive accuracy.

Our study was limited by the relatively small dataset, which contained only approximately a thousand rows which further reduced while data cleaning and precessing. This may have constrained the ability of complex models, such as neural networks and KNN, to fully capture intricate patterns in the data. Expanding the dataset in future work could improve model training, reduce prediction variance, and enhance generalization to unseen earthquake events

In conclusion, This study highlights the power of using data to understand and predict earthquakes. Even with a limited dataset, we were able to uncover patterns and make predictions that show promise for the future. While there is still work to be done, the findings from this study provide hope that, with continued effort and better data, we can develop tools that help communities prepare for and respond to earthquakes more effectively. This work is a step forward in making the world a safer place through knowledge and careful observation.

REFERENCES

[1] G. Gürsoy, A. Varol, and A. Nasab, "Importance of machine learning and deep learning algorithms in earthquake prediction: a review," in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–6, IEEE, 2023.

[2] E. Florido, J. L. Aznarte, A. Morales-Esteban, and F. Martínez-Álvarez, "Earthquake magnitude prediction based on artificial neural networks: A survey," *Croatian Operational Research Review*, pp. 159–169, 2016.

[3] F. Ahmed, J. B. Harez, S. Akter, A. Mubasira, S. M. Rahman, and R. Khan, "Earthquake magnitude prediction using machine learning techniques," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pp. 1–6, 2024.

[4] K. M. Asim, F. Martínez-Álvarez, A. Basit, and T. Iqbal, "Earthquake magnitude prediction in hindukush region using machine learning techniques," *Natural Hazards*, vol. 85, no. 1, pp. 471–486, 2017.

[5] S. Narayanakumar and K. Raja, "A bp artificial neural network model for earthquake magnitude prediction in himalayas, india," *Circuits and Systems*, vol. 7, no. 11, pp. 3456–3468, 2016.

[6] A. Galkina and N. Grafeeva, "Machine learning methods for earthquake prediction: a survey," *Proceedings of the Fourth Conference on Software Engineering and Information Management (SEIM-2019)*, pp. 13–25, 2019.

[7] M. Abdul Salam, L. Ibrahim, and D. S. Abdelminaam, "Earthquake prediction using hybrid machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 654–661, 2021.

[8] C. E. Yavas, L. Chen, C. Kadlec, and Y. Ji, "Improving earthquake prediction accuracy in los angeles with machine learning," *Scientific Reports*, vol. 14, no. 1, p. 24440, 2024.

[9] M. Mondol, "Analysis and prediction of earthquakes using different machine learning techniques." 35th Twente Student Conference on IT, University of Twente, Enschede, The Netherlands, 2021.

[10] N. Ridzwan and S. H. Md Yusoff, "Machine learning for earthquake prediction: a review (2017–2021)," *Earth Science Informatics*, vol. 16, pp. 1–17, 03 2023.

[11] R. Mallouhy, C. Abou Jaoude, C. Guyeux, and A. Makhoul, "Major earthquake event prediction using various machine learning algorithms," pp. 1–7, 12 2019.

[12] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[13] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[14] Q. Klopfenstein and S. Vaiter, "Linear support vector regression with linear constraints," *Machine Learning*, vol. 110, no. 7, pp. 1939–1974, 2021.

[15] T. Laloë, "A k-nearest neighbor approach for functional regression," *Statistics & probability letters*, vol. 78, no. 10, pp. 1189–1193, 2008.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[17] J. Gaudart, B. Giusiano, and L. Huiart, "Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data," *Computational statistics & data analysis*, vol. 44, no. 4, pp. 547–570, 2004.