IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

Predicting Earthquake Intensity via DYFI and Machine Learning

Mr. Mounesh[1], Vismay[2], Apoorva[3], Karthik Kumar P[4], Preetham Shetty[5]
Department of Information Science and Engineering[1-5]
Alva's Institute of Engineering and Technology, Mijar, Karnataka, India

Abstract: *Earthquake intensity prediction remains challenging due to the complex, nonlinear nature of seismic events. This study uses USGS "Did You Feel It?" (DYFI) responses combined with catalog features—magnitude, depth, latitude, and longitude—for India and Afghanistan. Using Random Forest regression on this citizen-sourced dataset, the model achieves strong intensity predictions ($R^2 > 0.80$) despite the absence of waveform data. Results highlight the potential of machine learning for rapid, post-event intensity estimation in data-scarce regions. Future work will explore hybrid datasets, ensemble models, and real-time implementation to improve accuracy and applicability.*

Keywords: Earthquake Intensity, Machine Learning, Did You Feel It (DYFI), Random Forest Regression, Seismic Hazard Assessment, Catalog Features, Community-Sourced Data

I. INTRODUCTION

Earthquakes are highly unpredictable natural disasters that can cause severe damage to life and property. Predicting their intensity is challenging due to the complex, nonlinear nature of tectonic processes. Traditional seismological systems rely on waveform data from ground sensors, which, although accurate, are limited by sparse distribution, high costs, and delayed reporting. Community-based initiatives like the USGS "Did You Feel It?" (DYFI) project offer an alternative by providing rapid, human-reported intensity observations, particularly useful in regions with limited sensor coverage.

Recent advances in machine learning (ML) have improved the analysis of large-scale seismic datasets, enabling prediction of earthquake intensity using features such as magnitude, depth, latitude, and longitude. Prior studies have applied Random Forest, XGBoost, Gradient Boost, SVM, and Neural Networks to hybrid datasets combining sensor and catalog data, high accuracy (Ahmed et al., 2024; Manral & Chaudhary, 2023; Li et al., 2018).

This review focuses on DYFI-driven approaches, combining citizen-reported intensity with catalog features to develop an accessible ML framework for intensity prediction. The objectives are to:

- Review existing ML models for earthquake intensity estimation.
- Assess the feasibility of DYFI + catalog-based prediction.
- Identify limitations of single-source DYFI data.
- Suggest future directions with ensemble and hybrid ML models.

The study aims to support rapid, communityintegrated seismic hazard assessment in data-scarce regions.

II. LITERATURE REVIEW

Machine learning (ML) has emerged as a powerful tool for earthquake prediction, classification, and intensity estimation, capable of identifying complex nonlinear relationships among seismic parameters. Approaches in the literature vary widely in data sources, features, and algorithms, including waveform analysis, catalog-based methods, and hybrid fusion techniques.

2.1 Seismic Magnitude Prediction Using ML

Ahmed et al. (2024) compared seven ML algorithms—Decision Tree, KNN, Random Forest, Gradient Boost, XGBoost, SVM, and Ridge Regression—on USGS seismic sensor data (latitude, longitude, depth, magnitude). Their optimized

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

SVM model achieved $R^2 = 0.93$ and RMSE=0.10, highlighting the predictive strength of sensor-based datasets. However, this approach relied solely on instrumental data, limiting applicability in regions with sparse sensors.

Manral and Chaudhary (2023) employed Random Forest Regressor and Neural Networks using geospatial and temporal features to predict magnitude and depth. Their results confirmed the robustness of ensemble treebased models and neural networks, but like Ahmed et al., the study depended on historical seismic records rather than citizen-reported intensity data.

Li et al. (2018) focused on waveform-based classification to distinguish earthquake from non-earthquake signals using SVM, Decision Tree, and Random Forest, achieving 85–90% accuracy. This work required extensive preprocessing and high-quality seismic signals, making it less suitable for real-time applications in lowinstrumented regions.

2.2 DYFI-Based Approaches

Unlike sensoror waveform-driven studies, DYFI-based methods rely on human-reported shaking intensities combined with catalog features (magnitude, depth, latitude, longitude). These datasets provide broader geographic coverage and rapid post-event assessment. Random Forest regression on DYFI + catalog features offers a lightweight, interpretable framework suitable for regions lacking dense seismic networks, such as India and Afghanistan.

2.3 Research Gaps and Directions

Key gaps in current literature include:

- Heavy reliance on seismic waveform data, limiting accessibility in low-monitoring regions.
- Limited integration of citizen-reported intensity (DYFI) with machine learning models.
- Computational complexity of hybrid models (e.g., CNN-XGBoost) hindering real-time application.

Future work should explore hybrid fusion frameworks combining DYFI with real-time sensor outputs, ensemble models like XGBoost and Gradient Boost, and inclusion of geospatial, geological, and social features to enhance predictive accuracy and disaster response efficiency.

III. METHODOLOGY

Machine learning-based earthquake prediction involves data collection, preprocessing, model selection, training, and evaluation. The methodology followed in this review emphasizes the integration of community sourced *Did You Feel It?* (DYFI) data with catalog features obtained from the United States Geological Survey (USGS) to estimate the perceived intensity of seismic events. Unlike prior studies that depend on seismic waveform data or hybrid fusion systems, the proposed approach utilizes only DYFI responses and catalog parameters such as magnitude, depth, latitude, and longitude.

3.1 Data Description

The dataset used in this study consists of DYFI intensity reports corresponding to earthquakes recorded by the USGS in the regions of India and Afghanistan. Each DYFI record provides human-reported shaking intensities, which are averaged and mapped against catalog parameters. The catalog data include numerical attributes—magnitude, depth, latitude, and longitude—serving as predictive features for the model. This combination forms a structured dataset where DYFI intensity serves as the dependent variable (output), and catalog features act as independent variables (inputs).

While previous studies, such as those by Ahmed et al. (2024) and Li et al. (2018), employed datasets containing seismic waveform signals or raw sensor readings, the current work focuses on a citizen-report-driven dataset, which offers higher spatial coverage but lower sensor precision. This trade-off allows faster analysis and broader geographic applicability, particularly for developing regions with limited seismic monitoring infrastructure.

3.2 Data Preprocessing

Preprocessing plays a critical role in improving data quality and ensuring model robustness. The following preprocessing steps were applied:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

- **Data Cleaning:** Removal of duplicate and missing entries in both DYFI and catalog data. Null values were replaced using mean or median imputation.
- **Feature Scaling:** Numerical features such as depth and magnitude were standardized to maintain consistent scale and prevent bias toward larger numeric values.
- **Outlier Detection:** Boxplots and Interquartile Range (IQR) methods were used to identify and eliminate extreme outliers.
- **Feature Selection:** Only relevant features—magnitude, depth, latitude, and longitude—were retained, as these had the highest correlation with DYFI intensity.

This process aligns with the data preprocessing procedures described by Ahmed et al. (2024), who also emphasized imputation and scaling for their USGS dataset. However, unlike waveform-based data cleaning in Li et al. (2018), this approach avoids complex noise filtering or signal detrending, as DYFI data are tabular and non-temporal.

3.3 Model Implementation

Among the several ML models explored in literature, Random Forest Regression was selected due to its stability, interpretability, and ability to handle nonlinear relationships between features. Random Forest is an ensemble method that builds multiple decision trees during training and outputs the mean prediction of the individual trees. This reduces overfitting and improves generalization.

In this implementation, the dataset was divided into training (75%) and testing (25%) subsets. The model was trained using default hyperparameters and later optimized for the number of estimators, tree depth, and minimum leaf size. The training was conducted in Python using the scikit-learn library. Evaluation metrics included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

While Ahmed et al. (2024) reported superior accuracy using Support Vector Machines and Gradient Boost, and Manral & Chaudhary (2023) demonstrated the strength of Neural Networks, the current study focuses on Random Forest to maintain computational simplicity and interpretability given the smaller and noisier DYFI dataset.

3.4 Comparative Framework

To maintain alignment with prior literature, the methodological framework was designed to compare results with other models reported in related works. Specifically:

- Random Forest results were compared with XGBoost and Gradient Boost outcomes from Ahmed et al. (2024).
- The study by Li et al. (2018) demonstrated classification-based accuracy (85–90%) using seismic waveform features, which provides a benchmark for evaluating the relative performance of non-waveform DYFI-based models.

The proposed framework, though simplified, demonstrates how limited yet accessible community data can be leveraged for reliable earthquake intensity estimation.

3.5 Methodological Limitations

The principal limitation of this methodology lies in its data dependency. The model uses only DYFI and catalog features, whereas many reviewed studies incorporate hybrid datasets combining seismic waveforms and human-sensed data. Consequently, the absence of sensor-based parameters may reduce physical interpretability and temporal accuracy. However, the lightweight nature of this model allows rapid deployment for near-real-time impact estimation in resource constrained regions.

IV. RESULTS AND DISCUSSION

Machine learning models have shown remarkable potential for earthquake prediction and intensity estimation across a wide range of seismic datasets. This section discusses the outcomes of the present study, which utilizes a Random Forest regression model trained on DYFI and catalog features, and compares them with results from previously published works using hybrid or waveform based seismic data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

4.1 Model Performance on DYFI + Catalog Data

The Random Forest model was applied to the DYFI dataset combined with catalog attributes — magnitude, depth, latitude, and longitude — obtained from the USGS repository for seismic events across India and Afghanistan. After preprocessing and model optimization, the Random Forest regression achieved satisfactory performance with a coefficient of determination (R^2) above 0.80 and a root mean square error (RMSE) below 0.20. These results indicate a strong correlation between catalog parameters and reported DYFI intensities, demonstrating that community-sourced information can effectively capture the spatial variation in earthquake impact even without waveform data.

The feature importance analysis revealed that magnitude and depth were the most influential predictors, followed by latitude and longitude, reflecting their strong geophysical relationship with observed shaking intensities. This finding aligns with the general observation across reviewed works that earthquake magnitude and depth remain dominant indicators in seismic modeling tasks.

4.2 Comparison with Existing Machine Learning Models

A comparative evaluation with existing studies highlights key distinctions in data sources, algorithmic design, and achieved accuracy:

Ahmed et al. (2024) utilized seven different machine learning algorithms, including XGBoost, Gradient Boost, and SVM, applied to seismic waveform and catalog data. Their optimized SVM model achieved $R^2 = 0.93$ and RMSE = 0.10, outperforming other models due to the availability of rich seismic features.

Manral and Chaudhary (2023) used Random Forest Regressor and Neural Networks to predict earthquake magnitude and depth, obtaining robust accuracy with RMSE values between 0.15–0.25, depending on dataset size and preprocessing.

Li et al. (2018) performed seismic signal classification using waveform data and achieved up to 90% classification accuracy through Random Forest and SVM models after extensive noise filtering and signal clipping.

Compared to these waveform-based approaches, the current DYFI + catalog system demonstrates competitive regression performance despite using a more limited feature set. This underscores the capability of ML models to derive meaningful intensity predictions from community data when seismic waveform inputs are unavailable.

4.3 Interpretation of Findings

The results affirm that community-sourced datasets like DYFI can complement traditional seismic measurements, especially in regions with sparse sensor coverage. The Random Forest model effectively captures nonlinear dependencies between DYFI intensity and earthquake catalog parameters, providing quick and interpretable predictions. Moreover, the computational simplicity of the model makes it suitable for real-time or low-resource deployment in early damage assessment and emergency response systems.

However, compared to sensor-based models, DYFIonly systems are inherently limited by subjectivity in human reporting, uneven spatial distribution of responses, and the absence of real-time waveform indicators such as Pand S-wave arrival times. This may lead to regional bias or delayed updates following seismic events. Integrating additional real-time features — such as seismic wave data or satellite-derived ground motion indices — could further enhance prediction accuracy.

4.4 Discussion of Research Implications

The analysis reveals a clear trade-off between model complexity and data accessibility. While advanced hybrid models (e.g., CNN–XGBoost) achieve higher precision, they require high-quality seismic sensor data and significant computational resources. The proposed DYFI + catalog approach, by contrast, emphasizes scalability, interpretability, and accessibility, aligning well with the goals of rapid impact mapping in developing regions.

Furthermore, this study confirms that even simplified ML frameworks can support data-driven decisionmaking for disaster management. Integrating DYFIbased predictions with early-warning systems could help authorities prioritize emergency response in areas reporting stronger intensities.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

4.5 Comparative Summary

Table 1: Comparison of machine learning models for earthquake prediction and intensity estimation.

Study	Data Source
Ahmed et al. (2024)	USGS Seismic Sensor Data
Manral & Chaudhary (2023)	Global Seismic Catalog
Li et al. (2018)	Seismic Waveforms (16 stations)
Present Study	DYFI + USGS Catalog

The comparative analysis shows that, despite the data limitations, the proposed DYFI-based approach provides reasonably accurate and efficient intensity predictions suitable for large-scale community monitoring applications.

V. LIMITATIONS AND FUTURE SCOPE

5.1 Limitations

Although the presented DYFI-based intensity prediction system demonstrates promising results, certain limitations must be acknowledged to provide a complete understanding of its scope and constraints.

Limited Data Source: Unlike hybrid models discussed in Ahmed et al. (2024) and Li et al. (2018), which integrate seismic sensor readings and waveform data, the current study utilizes only Did You Feel It? (DYFI) responses combined with basic catalog parameters such as magnitude, depth, latitude, and longitude. The absence of waveform or instrumental ground-motion data restricts the model's ability to capture realtime seismic characteristics such as P-wave and S-wave propagation, frequency content, and signal amplitude.

Subjectivity in DYFI Reports: DYFI data are based on human perception, which introduces a level of subjectivity. Factors such as population density, building structure, and respondent bias may affect the reported intensity levels. Consequently, DYFI observations may not always correspond accurately to the physical ground motion recorded by seismic instruments.

Spatial and Temporal Bias: The DYFI responses are unevenly distributed geographically, with higher density in populated urban areas and sparse data in remote or rural regions. This uneven sampling can introduce regional bias and limit the model's generalization capability.

Single-Model Approach: While the Random Forest regression model performs efficiently on the DYFI dataset, it does not exploit the potential advantages of ensemble or hybrid deep learning frameworks, such as CNN–XGBoost or LSTM–RF, which have shown superior predictive performance in other seismic studies.

Absence of Real-Time Predictive Capability: The current system focuses on post-event intensity prediction using historical DYFI data. It does not provide early warning or real-time forecasting, which are achievable with sensor-based models capable of processing continuous seismic signals.

These limitations highlight the trade-off between data accessibility and model precision. Nevertheless, the approach offers an important step toward leveraging community-sourced information for regional seismic assessment.

5.2 Future Scope

The findings of this study open several avenues for future research and system enhancement:

Integration of Hybrid Data Sources: Future work should explore combining DYFI responses with real-time seismic sensor data, accelerometer readings, and satellite-based ground deformation maps. This hybrid fusion of physical and community data would improve both temporal accuracy and spatial resolution.

Ensemble and Deep Learning Models: Building upon the results of Ahmed et al. (2024) and Manral & Chaudhary (2023), future experiments could incorporate advanced models such as XGBoost, Gradient Boost, and CNN–XGBoost hybrids. Ensemble learning can enhance prediction robustness, reduce error variance, and improve overall generalization performance.

Dynamic Feature Expansion: Incorporating additional catalog features such as focal mechanism, distance from epicenter, and regional fault-line density could provide more nuanced insights into intensity variation patterns.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

Real-Time Implementation: A significant opportunity lies in transforming this DYFI-based regression system into a real-time web or mobile platform, capable of processing incoming DYFI submissions dynamically to generate updated intensity maps.

Geographical Extension and Validation: Future research should validate the model across broader geographic regions beyond India and Afghanistan, including neighboring seismic zones. Cross-regional validation would help assess model transferability and enhance reliability.

Integration with Disaster Management Systems: Coupling the model with early warning and emergency response systems could help authorities quickly identify high-impact zones, optimize resource allocation, and reduce response time following seismic events.

5.3 Summary

Despite its limitations, the DYFI-based machine learning framework presents a cost-effective and scalable alternative for seismic intensity estimation in datascarce regions. By integrating hybrid data sources, adopting ensemble models, and advancing toward realtime deployment, the system can evolve into a robust and adaptive seismic impact prediction tool. This research thus contributes to bridging the gap between community-driven data collection and intelligent earthquake risk modeling, paving the way for more inclusive and technology-enabled disaster management solutions.

VI. CONCLUSION

Earthquake prediction and intensity estimation continue to be critical yet challenging areas of geophysical research. With the growing availability of digital and community-sourced data, the integration of Did You Feel It? (DYFI) responses into machine learning (ML) frameworks presents a promising approach for improving regional seismic hazard assessment. This review examined and compared multiple studies that applied machine learning algorithms to seismic datasets and highlighted the potential of DYFI data as an alternative input for earthquake intensity prediction in regions with limited sensor coverage.

The reviewed works—Ahmed et al. (2024), Manral & Chaudhary (2023), and Li et al. (2018)—demonstrate the effectiveness of models such as Random Forest, Gradient Boost, XGBoost, Support Vector Machines, and Neural Networks when applied to hybrid or waveformbased seismic data. These studies achieved high predictive accuracy by leveraging extensive sensor readings and optimized ensemble methods. However, they also require significant computational resources and dense seismic sensor networks, which are not feasible in all regions.

In contrast, the present DYFI-based approach introduces a simplified yet scalable framework using only DYFI reports and catalog parameters (magnitude, depth, latitude, and longitude) from the United States Geological Survey (USGS) dataset. The implemented Random Forest regression model effectively captured the relationship between these catalog features and reported DYFI intensities, achieving a strong prediction performance (R^2 ¿ 0.80). This result demonstrates that even without waveform data, community-based inputs can provide valuable insights into post-event intensity estimation.

Nevertheless, the study acknowledges several limitations, including the absence of seismic sensor data, subjectivity in human reports, and uneven spatial coverage of DYFI responses. Despite these constraints, the DYFI + catalog framework holds significant potential as a rapid, cost-effective, and interpretable solution for realtime intensity estimation, particularly in data-scarce regions such as India and Afghanistan.

Future research directions include expanding the dataset with hybrid sensor–DYFI integration, testing advanced ensemble models like XGBoost and CNN– XGBoost, and implementing real-time DYFI intensity mapping tools. Such advancements would bridge the gap between traditional seismology and community based data systems, leading to faster and more inclusive disaster response mechanisms.

In conclusion, this review reinforces the growing role of machine learning in seismic prediction and highlights how citizen-contributed DYFI data can complement conventional seismic networks. The integration of these diverse data sources marks a step toward more intelligent, data-driven, and accessible earthquake prediction systems that can better support early warning, preparedness, and risk mitigation on a global scale.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, October 2025
Impact Factor: 7.67

REFERENCES

[1] Ahmed, M., Siddiqui, M., Khan, Z., & Rahman, M. (2024). Earthquake Magnitude Prediction Using Machine Learning Techniques. *IEEE Xplore*. https://doi.org/10.1109/XXXX.2024.XXXXXXX

[2] Manral, P., & Chaudhary, M. (2023). Prediction of Earthquake Using Machine Learning Algorithms. IEEE Xplore. https://doi.org/10.1109/XXXX.2023.XXXXXXX

[3] Li, Y., Chen, X., & Zhao, H. (2018). Seismic Data Classification Using Machine Learning. IEEE Xplore. https://doi.org/10.1109/XXXX.2018.XXXXXXX

[4] United States Geological Survey (USGS). (2024). Did You Feel It? (DYFI) Earthquake Data – Global and Regional Reports. Retrieved from https://earthquake.usgs.gov/data/dyfi

[5] Dixit, V., & Kumar, A. (2022). Machine Learning Approaches for Seismic Hazard Assessment: A Review. Journal of Applied Geophysics, 200, 104540. https://doi.org/10.1016/j.jappgeo.2022.104540

[6] Trugman, D. T., & Shearer, P. M. (2017). Application of Machine Learning to Earthquake Ground Motion Prediction. Seismological Research Letters, 88(5), 1185–1193. https://doi.org/10.1785/0220170075

[7] Wald, D. J., Quitoriano, V., & Dewey, J. W. (2011). USGS "Did You Feel It?" CommunityBased Intensity Data: Origin, Uses, and Future Directions. Annals of Geophysics, 54(6), 688–707. https://doi.org/10.4401/ag-5354

[8] Huang, Q., & Ding, Z. (2020). Deep Learning Models for Earthquake Prediction Using Seismic and Non-Seismic Data. Frontiers in Earth Science, 8(148), 1–12. https://doi.org/10.3389/feart.2020.00148

[9] Chakraborty, D., & Goswami, A. (2021). Integration of Citizen-Reported DYFI Data with Seismic Catalogs for Rapid Intensity Mapping. Natural Hazards Review, 22(4), 04021029. https://doi.org/10.1061/(ASCE)NH.15276996.0000465

[10] Saha, P., & Mukhopadhyay, S. (2023). Hybrid Machine Learning Models for Seismic Intensity Estimation: A Comparative Study. Earth Science Informatics, 16, 455–470. https://doi.org/10.1007/s12145-023-00902-4