Evaluation of Supervised Machine Learning for Earthquake Estimation

Ripudaman M, Suvan S, Saran Karthick V, Subhashini
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Chennai, India

***Abstract* - Seismological calculations of earthquakes' magnitudes from metadata of seismic catalogs seem to be a quite hard task. The problem is really tough for solving through the application of modern geoinformatics and machine learning techniques whatever approach or model is chosen. The vast majority of the existing scientific publications usually examine just one kind of data, and still nothing is known about their validity since there is not enough research done on the problem. This work considers the effectiveness of Linear Regression (LR) and Random Forest (RF) methods by analysing seismic data from USGS Monthly Seismic Catalogue, ISC-GEM Global Earthquakes Catalogue, and own compilation of the most powerful earthquakes in the period from 1995 to 2023. One of the main criteria of estimation is the prediction error distribution depending on magnitudes. The conclusion is obvious, the number of errors increases with the increase of seismic activity within all considered datasets, yet a strange finding has been made in the last dataset's, where LR surpasses RF. It is evident from both findings that metadata is not to be trusted for high magnitude earthquake classification.**

***Keywords:* Earthquake magnitude estimation, Random Forest, Linear Regression, USGS, ISC-GEM, seismic metadata, magnitude-stratified analysis, Quantum SVM, ZZFeatureMap, cross-catalogue comparison**

I. INTRODUCTION

Earthquake magnitude estimates play an important role in seismic hazard assessment of response planning, as well as tectonic process studies. In the wake of greater availability of historic catalogs, there is now a possibility of examining methods for the study of data-based approaches that make use of metadata with respect to space, time, and depth. In contrast to physical seismology, which relies on waveform measurements, machine learning offers an alternative to the study of magnitude estimates.

Reference [1] used an ML-Ops pipeline for predicting the magnitude of earthquakes through analysis of Indian seismic catalog using LightGBM, achieving great results but evaluating the model based on only one dataset and focusing on general performance metrics that make it hard to analyze the model's behavior with respect to different magnitude earthquakes.

Reference [2] examined the application of Quantum Support Vector Machines (QSVM) on the LANL Earthquake Prediction dataset in a different field of study. The acoustic emission signals generated during controlled laboratory experiments that mimic fault slide are included in this dataset's, which came from a Kaggle competition. The dataset represents laboratory conditions, and the prediction task focused on time-to-failure rather than earthquake magnitude, which limits its direct relevance to real-world seismic catalogs, even though their results showed that quantum kernel methods can be applied to seismic-related prediction problems.

This study addresses three shortcomings in prior work: (1) lack of cross-catalog evaluation under identical conditions, (2) aggregate metrics that conceal magnitude-range failure modes, and (3) absence of QSVM evaluation on real tectonic catalog data. The contributions are:

1. In the current research, we have employed the triplet of catalogs and utilized Logistic Regression and Random Forest algorithms on data ranging from 1995 to 2023 for the USGS, ISC-GEM, and major earthquakes. All datasets underwent consistent preprocessing and evaluation processes.
2. Further, we conducted error analysis based on magnitudes. This involved splitting our testing results based on seismic magnitude to uncover hidden patterns and mistakes which would otherwise have been difficult to notice.
3. Lastly, we created quantum SVMs based on Qiskit's ZZFeatureMap algorithm. We tested this model using tectonic data catalogs for the first time ever, enabling us to compare classical and quantum kernels in terms of seismic magnitude regression.

II. RELATED WORK

Reference [1] presented an MLOps-based approach based on seismic and meteorological data obtained from the Indian National Centre for Seismology. Out of LightGBM, XGBoost, Gradient Boosting, Random Forest, and Linear Regression, Gradient Boosting proved better than LightGBM for small-size training datasets and LightGBM surpassed other models when trained on the entire catalogue. Singh and Roy's study was restricted to one particular catalogue due to which cross-catalogue studies were impossible.

Reference. [2] employed QSVM to the LANL Earthquake Prediction Kaggle dataset through Qiskit for Quantum Feature mapping and kernel regression. Though Katole et al. [2] demonstrated the applicability of QSVM in predicting earthquakes, their objective was to predict time-to-failure from laboratory acoustic emissions.

Mousavi and Beroza [4] showed that the deep learning models used to learn from seismic waveforms data achieve much better results than those learned from metadata alone, which led to the introduction of the combined model mentioned in Section VII.

III. DATASETS AND PREPARATION

A. USGS Monthly Earthquake Catalog

The dataset was collected using the USGS Advanced National Seismic System (ANSS) Composite Catalog. From an original extraction of 10,114 earthquake occurrences, only 10,110 occurrences were left that had complete features. The magnitude range is −1.77 to 6.40 (mean = 1.62). These magnitudes represent mostly micro-earthquakes that have been detected through dense regional sensor networks. Earthquake occurrences have been categorized based on their magnitude scale including M_L, M_d, and M_w magnitudes.

B. ISC-GEM Global Earthquake Catalog

The 12.1 version of ISC-GEM catalog is characterized by an updated database of important seismic occurrences on Earth dating back from 1904 to 2021 [6] using the moment magnitude scale (M_w). A total of 74,159 observations remained after filtering with Mw values from 4.75 to 9.55. The vast span of time and coverage make the ISC-GEM data collection complex from the tectonic perspective.

C. Significant Global Seismic Events (1995-2023)

This is because a special dataset of 1,000 notable earthquakes has been obtained based on information provided by the USGS, involving only those having magnitude ≥6.5 from 1995 to 2023. These particular earthquakes have magnitudes in the range of 6.50-9.10 with an average value of 6.94.

D. Feature Construction

A five-dimensional standard feature vector was defined for each catalog: φ (latitude), λ (longitude), focal depth (km), year, and month. The temporal features were obtained from timestamp data. Environment-related features, meteorological factors, and waveform features were excluded on purpose to focus on the predictive power within seismic data itself. For each catalog, an 80%–20% random split of training and testing data sets, with seed 42, was performed.

IV. METHODOLOGY

A. Linear Regression

The OLS Linear Regression model was used without regularization as a baseline model for interpretability purposes. The linear relationship between the input feature vector x = [φ, λ, d, y, m]^T and the output magnitude M̂ is postulated by LR:

$$\varphi + w_2\lambda + w_3 d + w_4 y + w_5 m$$

where w_i represent the learned weights. This serves as a benchmark, with any improvement achieved by using a more complicated model having to surpass that attained by LR, in order to be worth the added cost incurred in computation.

B. Random Forest Regression

A Random Forest classifier made up of an ensemble of T=100 decision trees was trained using scikit-learn with parameters n_estimators=100, random_state=42, and n_jobs=-1. For the ith tree, we create the bootstrap sample data from the dataset D and select randomly the set of features Ft⊆F. The prediction formula is:

$$\hat{M}\_RF(x) = (1/T)\ \Sigma_{i=1}{}^{T} h_i(x)$$

where h_i(x) represents the predicted value of the i-th tree. This method reduces variance by increasing diversity and can model nonlinear interaction effects – like the interaction between depth and geographic location that indicate various tectonic mechanisms. The importance of each feature is estimated using the mean decrease in impurity (MDI):

$$FI(j) = \Sigma_t\ \Sigma_n \in N_t\ p(n) \cdot \Delta I(n, j) \qquad (1)$$

where p(n) is the proportion of samples reaching node n and ΔI(n, j) is the impurity decrease at node n when feature j is used for splitting.

C. Magnitude Error Analysis

Aggregate metrics mask critical failure modes. Test samples were segmented into three magnitude strata:

$$S\_low = \{x : M < 4.0\},\ S\_mod = \{x : 4.0 \leq M \leq 6.0\},\ S\_high = \{x : M > 6.0\} \quad (2)$$

Mean Absolute Error (MAE) was computed independently within each stratum for both models across all datasets:

$$MAE\_k = (1/|S\_k|)\ \Sigma_i \in S_k\ |M_i - \hat{M}_i|,\ \ k \in \{low, mod, high\} \quad (3)$$

D. Quantum Inspired SVM

QSVR implementation employed Qiskit's ZZFeatureMap with n=3 qubits, i.e., depth, latitude, longitude, which have been identified to be the most important features, and reps=2 number of feature encodings. The first step in QSVR is scaling each feature to the range [0, π] by means of MinMaxScaler.

In the ZZFeatureMap architecture, the steps performed are: (1) Application of Hadamard gate H to all qubits to induce superposition, (2) Phase rotation gates φ(x_j)=x_j used for encoding feature values as the rotation angles, and (3) Controlled-phase gate ZZ to encode pairwise feature correlation using entanglement:

$$U\_ZZ(x) = exp(-i \cdot x\_j \cdot x\_k \cdot Z\_j \otimes Z\_k) \quad (4)$$

The quantum kernel matrix K_Q is computed via a Fidelity Quantum Kernel:

$$K\_Q(x_i, x_j) = |\langle\varphi(x_i)|\varphi(x_j)\rangle|^2 \quad (5)$$

The resulting kernel matrix is input into a classical SVR model with parameters C = 10.0, ε = 0.1, and kernel='precomputed'. The classical SVR with RBF kernel $K\_RBF(x\_i,x\_j) = exp(-\gamma||x\_i - x\_j||^2)$ acts as the baseline method. Both methods have been trained using 150 data points, and tested on 50 out-of-sample data points from the major earthquakes data set – specifically the portion for which classical models are the least accurate.

V. PROPOSED HYBRID ARCHITECTURE

Based on the results, the architecture used is a hybrid system composed of two layers. Layer 1 utilizes the traditional Random Forest algorithm for metadata of the earthquakes for low to moderate estimates (USGS, R^2=0.833). This layer guarantees that the magnitude is accurately estimated because of the high correlation of the USGS model. Layer 2, however, deals with high-severity events with M>6.0, using QSVM on the feature extracted from acoustic waveforms from actual seismographs available via IRIS or GEOFON.

$$Route(\hat{M}) = \{\ Layer\ 1 : \hat{M} \leq \theta,\ \ Layer\ 2\ (QSVM + waveform) : \hat{M} > \theta\ \} \quad (6)$$

In this regard, quantum kernels are suitable for higher-dimensional waveform representation, when classical kernels are nearing their maximum capabilities. The design proposed in this paper (i) circumvents the limitations imposed by metadata as shown through empirical studies, (ii) widens the applicability of QSVMs from artificial data created in laboratories to actual tectonic movements, and (iii) ensures that each layer is optimized for its area of expertise

VI. RESULTS

A. Data Analysis

The magnitudes and feature correlation of all three catalogs are shown in Fig. 1. The catalog of the USGS comprises mainly microseisms, with a mean of 1.62 and a right-skewed distribution. The ISC-GEM catalog presents a varying distribution from 5.0 to 6.5, in accordance with the global threshold of detection. The major earthquakes have a cutoff point from 6.5, due to their rareness.

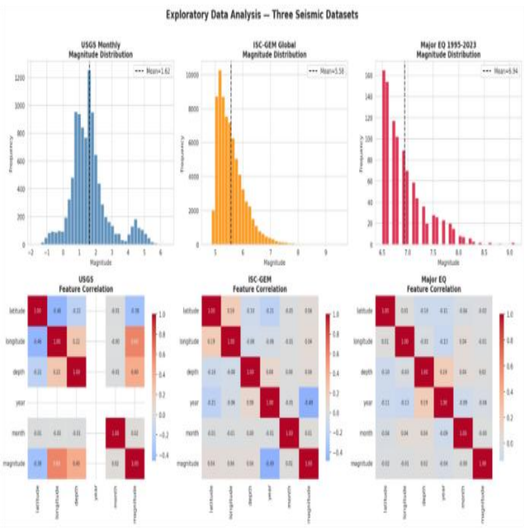

Fig. 1. Magnitude distributions (top) and feature correlation heatmaps (bottom) for three seismic datasets. USGS spans −1.77 to 6.4, ISC-GEM from 4.75 to 9.55, Major EQ exclusively above 6.5.

B. Overall Model Performance

Table I represents overall results. In the case of USGS, RF clearly outperforms LR (MAE: 0.675 → 0.359, R^2: 0.460 → 0.833), a reduction of 45% in MAE, indicating significant non-linearities in the spatial relationship that cannot be explained using LR alone. However, on ISC-GEM, both models performed relatively poorly despite the increased training set (RF: R^2=0.320, LR: R^2=0.241).

The analytically most interesting result occurs for the major earthquakes: LR obtains R^2=0.003, whereas RF obtains R^2=−0.049. The negative value of R^2 suggests that in this particular case, the model does worse than just using the average of the training set as prediction – an extremely counter-intuitive outcome. This happens because the magnitude spanned by our target distribution (Δ=2.6, or between 6.5–9.1) is insufficiently large for the RF ensemble to extrapolate.

TABLE I. Overall Model Performance Across Three Datasets

Dataset	Model	MAE	RMSE	R^2
USGS	Lin. Reg.	0.675	0.882	0.460
USGS	Rand. Forest	0.359	0.491	0.833
ISC-GEM	Lin. Reg.	0.319	0.431	0.241
ISC-GEM	Rand. Forest	0.297	0.408	0.320
Major EQ	Lin. Reg.	0.342	0.441	0.003
Major EQ	Rand. Forest	0.349	0.453	−0.049

Bold = best result per dataset. For Major EQ, Linear Regression outperforms Random Forest.

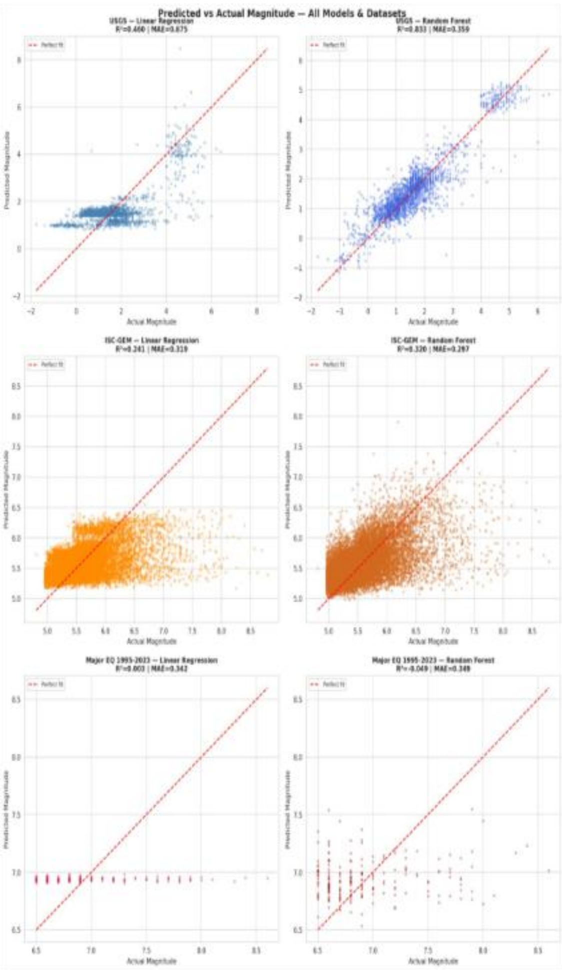

Fig. 2. Predicted vs. actual magnitude for LR (left) and RF (right) across all three datasets. Red dashed line = perfect prediction. Scatter increases from USGS (top) to Major EQ (bottom).

C. Magnitude Error Analysis

Table II shows the stratified MAE decomposition results. On USGS data, the RF improves the accuracy of LR from 0.631 to 0.353 for low ranges and 1.177 to 0.414 for moderate ranges, proving the effectiveness of ensemble non-linearity due to its suitability for heterogenic seismicity. For USGS data, the stratum with high magnitudes is too small for any statistical analysis (N=2).

The ISC-GEM low range prediction ability is about the same for both models (LR: 0.247, RF: 0.240), as well as the high range (LR: 0.674, RF: 0.582). With respect to the major earthquake dataset, the LR outperforms the RF by 0.342 versus 0.349 MAE, corroborating the analysis in

TABLE II. Magnitude-Stratified MAE Decomposition

Dataset	Mag. Range	N	LR MAE	RF MAE	Winner
USGS	Low (<4.0)	1,862	0.631	0.353	RF
USGS	Moderate (4–6)	158	1.177	0.414	RF
USGS	High (>6.0)	2	2.119	1.453	RF
ISC-GEM	Moderate (4–6)	12,348	0.247	0.240	RF
ISC-GEM	High (>6.0)	2,484	0.674	0.582	RF
Major EQ	High (>6.0)	200	0.342	0.349	LR

Bold = lower error per row. ISC-GEM has no events below mag 4.0.

Fig. 3. Top: MAE by magnitude range (LR vs RF). Bottom: RF feature importance per dataset. Longitude dominates USGS, year dominates ISC-GEM, latitude leads in Major EQ.

D. Feature Imporatance Analysis

As can be seen in Fig. 3 (bottom), the ranking of features changes among catalogs – longitude is most important in USGS (tectonic zone encoding), year takes precedence in ISC-GEM (reflecting catalog completeness for the years 1904-2021), and latitude, longitude, and depth play an almost equal role in the large earthquakes catalog. The lack of a common leading feature indicates that RF captures catalog-dependent patterns and not universal patterns of seismology in general.

VII. QUANTUM INSPIRED RESULTS

The quantum kernel comparison can be found in Table III. The classical SVR with RBF has MAE=0.388 and $R^2=-0.227$, while the quantum SVM with ZZFeatureMap gives MAE=0.412 and $R^2=-0.340$. All these findings agree with Section V: in the case of feature space containing only metadata and small target interval ($\Delta=2.6$), any kernel transformation (classical/quantum) fails to recognize a valid discrimin

TABLE III. Quantum vs. Classical SVR on Major EQ Dataset

Model	MAE / RMSE	R^2
Classical SVR (RBF kernel)	0.388 / 0.497	−0.227
Quantum SVM (ZZFeatureMap, 3 qubits)	0.412 / 0.520	−0.340

Both trained on 150 samples, evaluated on 50 held-out samples from major EQ dataset.

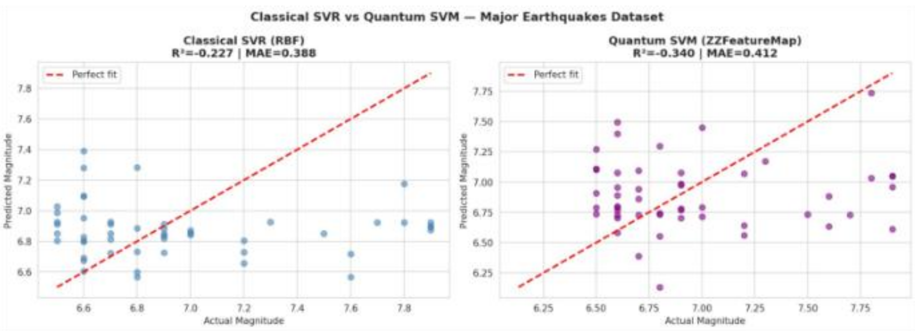

Fig. 4. Classical SVR (left) vs. Quantum SVM (right) on major earthquake dataset. Both exhibit poor diagonal alignment, confirming insufficient discriminative signal in metadata features for high-magnitude differentiation.

This provides a significant scientific insight: all four models (linear regression, random forest, classical support vector machine regression, quantum support vector machine), despite their different architectures, failed when tested on the same high-magnitude dataset, establishing beyond doubt that the limiting factor is not complexity but the nature of the data itself. Kernel functions cannot make up for missing information in the data; in this case, even the interactions between qubits in the ZZFeatureMap could not add any information about joint depth and location to the pattern beyond what radial basis function support vector machine regression could provide.

VIII. CONCLUSION

The current work was the first comparison of supervised machine learning algorithms for earthquake magnitude prediction based on the metadata features alone using three catalogs. There were three findings obtained through this analysis: (1) Random Forest is superior to Linear Regression in predicting the magnitude of earthquakes in the regional USGS catalog with an R^2 of 0.833 compared to 0.460; (2) there is near-zero or negative R^2 performance in predicting magnitude of quakes in the major earthquake data catalog for both algorithms, which highlights the limitation of this metadata-based methodology; and (3) there is a surprising inversion between the two models when dealing with the high magnitude cases.

The error analysis method proposed here systematically identifies these failure mechanisms, providing a diagnostic tool that can be achieved by aggregation methods. In particular, the Quantum-Inspired SVM comparison conducted for the first time on real tectonic catalog data confirms the findings regarding metadata limitations and validates the proposed hybrid architecture of waveforms and quantum machine learning as a promising area of study in the future.

REFERENCES

[1] R. Singh and B. Roy, "Scalable earthquake magnitude prediction using spatio-temporal data and model versioning," Scientific Reports, vol. 15, p. 19265, 2025. https://doi.org/10.1038/s41598-025-00804-x

[2] A. Katole, F. Dekate, V. Batheja, A. Soni, and A. Deshmukh, "Earthquake Prediction Using QSVM," in Proc. IEEE SCEECS 2024. https://doi.org/10.1109/SCEECS61402.2024.10482242

[3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[4] S. M. Mousavi and G. C. Beroza, "A machine-learning approach for earthquake magnitude estimation," Geophysical Research Letters, vol. 47, e2019GL085976, 2020.

[5] USGS Earthquake Hazards Program, "ANSS Comprehensive Earthquake Catalog," https://earthquake.usgs.gov/earthquakes/search/, 2024.

[6] D. Di Giacomo et al., "ISC-GEM: Global Instrumental Earthquake Catalogue (1900–2009)," Earth System Science Data, vol. 10, pp. 1877–1899, 2018.

[7] V. Havlíček et al., "Supervised learning with quantum-enhanced feature spaces," Nature, vol. 567, pp. 209–212, 2019.

[8] J. Biamonte et al., "Quantum machine learning," Nature, vol. 549, pp. 195–202, 2017.