Cross-Regional Generalization Analysis of BiLSTM-Based Seismic Hazard Prediction

Preety Singh[1,3] and Rajiv Singh[2*]

[1,2] Department of Computer Science, Banasthali Vidyapith, Rajasthan, India
[3] Department of CSE (AIML & IoT), Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering & Technology, Hyderabad, India
preeti17singh@gmail.com, jkrajivsingh@gmail.com*

Abstract. Seismic hazard prediction is essential for disaster preparedness, yet the ability of deep learning models to generalize across geographic regions remains underexamined. This study introduces a cross-regional evaluation framework based on a BiLSTM network to assess the geographic transferability of seismic hazard prediction models, using two heterogeneous earthquake datasets: the IEB and Indian Subcontinent seismic catalogues. Binary hazard classification used five features, latitude, longitude, depth, month, and day, excluding the absolute year to avoid a temporal confound between the catalogues' differing collection periods. Four experiments evaluated within-region and cross-region performance. Within-region accuracy reached 82.7% (IEB) and 78.9% (Subcontinent), while cross-region accuracy reached 84.6% (IEB→Subcontinent) and 87.8% (Subcontinent→IEB). However, recall for high-hazard events fell from 20.6–32.9% within-region to 6.6–14.7% cross-region, showing that accuracy alone masks a severe loss of sensitivity to the events that matter most. A Generalization Retention Rate (GRR) is introduced to quantify this decline, yielding recall-based GRR of approximately 32% (IEB-anchored) and 45% (Subcontinent-anchored), compared with accuracy-based GRR near or above 100% for the same transfers. These results show that predictive accuracy alone cannot indicate a model's ability to detect high-risk seismic hazards across differing tectonic regions; robustness metrics such as GRR should be used instead.

Keywords: Seismic Hazard Prediction · BiLSTM · Cross-Regional Generalization · Deep Learning · Generalization Retention Rate.

1 Introduction

Natural disasters such as earthquakes are incredibly damaging, causing large amounts of death, destruction to infrastructure, and economic disruption on a global scale. Accurate prediction of seismic hazard is required for disaster preparedness, emergency response, and long-term urban planning. With developments in AI and deep learning, researchers can now produce data-driven models that analyse and identify patterns in seismic data and enhance prediction accuracy.

© The Author(s) 2026
A. Shastri et al. (eds.), *Proceedings of the Fourth Workshop on Applications of Computational Intelligence and Interactive Systems (WACIIS 2026)*, Advances in Intelligent Systems Research 217,
https://doi.org/10.2991/978-94-6239-832-0_6

The usefulness of such models is largely determined by the dataset used for training and testing. Several studies have used CNNs, LSTM, BiLSTM, and transformer networks for earthquake and seismic hazard prediction, achieving positive results when trained and tested on data from the same geographic location. However, many previous studies focus on region-specific datasets and report only accuracy, making no effort to determine how well their models generalise outside the training region. Each tectonic region has its own geological characteristics, fault systems, and earthquake statistics, so a model trained on one region will not necessarily perform well on another, particularly for low-probability, high-hazard events. Very little research has evaluated the geographic transferability of seismic hazard prediction systems systematically, leaving a significant gap.

This work provides a cross-regional evaluation framework using a Bidirectional Long Short-Term Memory (BiLSTM) network for seismic hazard prediction. Two heterogeneous datasets, the Indian Earthquake Bulletin (IEB) and the Indian Subcontinent seismic dataset, are used to evaluate within-region and cross-region performance across four experiments. A Generalization Retention Rate (GRR) is proposed to quantify performance degradation during geographic transfer.

The major contributions of this work are:

- A cross-regional evaluation framework assessing the geographical generalization of BiLSTM-based seismic hazard prediction.
- Within-region and cross-region experiments using the IEB and Indian Subcontinent datasets.
- A formally defined Generalization Retention Rate (GRR), computed across all evaluation metrics.
- Evidence that high overall accuracy coexists with a considerable drop in high-hazard detection under cross-regional transfer, motivating robustness-aware evaluation.

2 Related Work

Machine Learning for Seismic Hazard Prediction. Earlier work relied on probabilistic hazard assessment and statistical modeling. As more catalogs became available, ML techniques such as Random Forest, SVM, Gradient Boosting, and ANN ensembles were used to predict earthquake magnitude and hazard [10, 5, 7, 3]. These methods generally outperform traditional statistics but suffer from handcrafted features and poor adaptability across heterogeneous seismic environments; performance is known to drop when models trained on one dataset are applied elsewhere [7, 13].

Deep Learning Approaches. LSTM, BiLSTM, CNN, GRU, and Transformer networks have shown strong capability in magnitude prediction, event detection, and hazard classification [12, 9, 5, 6, 1]. Hybrid convolutional-recurrent architectures further improve spatio-temporal feature extraction, and generalized networks for real-

time early warning have recently been proposed [14]. Most of this work, however, remains confined to a single seismic region.

Generalization and Explainability. Physics-informed ML, generalized deep learning, and explainable AI have been used to improve accuracy and interpretability [2, 4, 8]. The POSEIDON system combines physics knowledge with deep learning for interpretable seismic prediction [8], and interpretable models have targeted the class-imbalance problem inherent to rare, large-magnitude events [4]. Several reviews highlight open challenges in imbalance, interpretability, real-time prediction, and geographic transferability [12, 14, 15].

Research Gap. Prior work largely (i) evaluates within a single geographic area, (ii) does not assess cross-region performance, (iii) reports accuracy but not high-risk-event detection, and (iv) rarely quantifies cross-region performance drop. Table 1 summarizes this gap. This paper addresses all four points using a cross-regional BiLSTM framework over the IEB and Indian Subcontinent datasets.

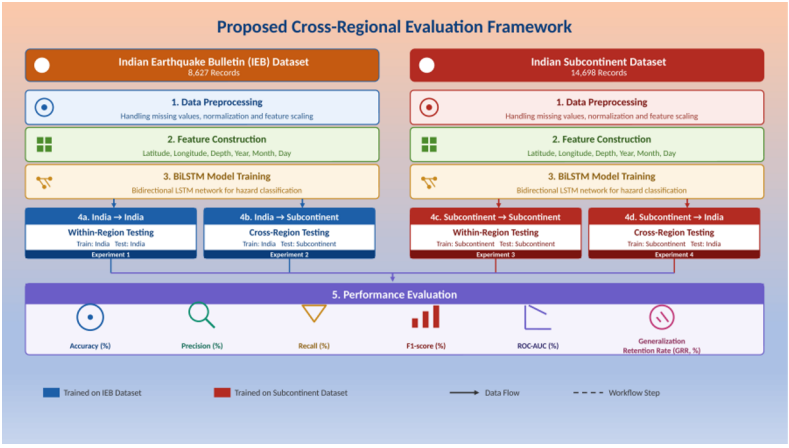

Fig. 1. Proposed cross-regional BiLSTM framework for within-region and cross-region seismic hazard prediction.

Table 1. Comparison of recent seismic hazard prediction studies.

Method	Region	Cross-Reg.	Contribution / Limitation
Gen. NN [14]	Global waveform	No	Early warning; no cross-region hazard eval.
LSTM, BiLSTM, Transf. [12]	Japan, Indonesia, HKKH	No	Magnitude prediction within selected regions
DNN [9]	Chile, Hindukush, S.Cal.	No	Limited transferability analysis
CNN–LSTM [5]	Waveform	No	Single regional dataset only

LSTM [1]	Nepal region	No	Generalization not investigated
LSTM-RNN, XAI [6]	Japan, Chile, Calif.	No	Cross-regional hazard not assessed
Ensemble, LSTM [3]	China	No	Region-specific only
LSTM, Transf., STTN [11]	Chile, Japan, Mexico	Partial	Transferability not quantified
Physics-Constr. ML [13]	Ground motion	No	Interpretability, not regional adapt.
POSEIDON [8]	Global	No	Cross-regional classification not studied
AI Review [4]	Multiple	No	No experimental validation
BiLSTM (ours)	IEB & Subcontinent	Yes	Cross-regional GRR framework

3 Proposed Cross-Regional BiLSTM Framework

The proposed framework (Fig. 1) comprises five steps: dataset collection, preprocessing, BiLSTM-based hazard prediction, cross-regional analysis, and performance evaluation, allowing both within-region and cross-region assessment.

3.1 Study Datasets

Two datasets are used: the Indian Earthquake Bulletin (IEB, 8,627 records: 8,459 low-hazard, 168 high-hazard) and the Indian Subcontinent dataset (14,698 records: 14,316 low-hazard, 382 high-hazard), summarized in Table 2. Both share latitude, longitude, depth, year, month, and day attributes, with the binary hazard label derived from magnitude thresholding.

Two data-quality characteristics are disclosed explicitly. First, a bounding-box overlap analysis found 90.4% of Indian Subcontinent points fall within the IEB bounding box (64.7% conversely), so the cross-regional evaluation tests generalization across differently curated, partially overlapping catalogues rather than fully disjoint tectonic regions (Fig. 2). Second, ~48% of depth values in both catalogues correspond to standard "fixed depth" convention values (10, 30, 33, 35 km) rather than precisely measured depths.

Table 2. Dataset description.

Dataset	Total	Low-Haz.	High-Haz.	Features
IEB India	8,627	8,459	168	Lat, Lon, Depth, Yr, Mo, Day
Ind. Subcontinent	14,698	14,316	382	Lat, Lon, Depth, Yr, Mo, Day

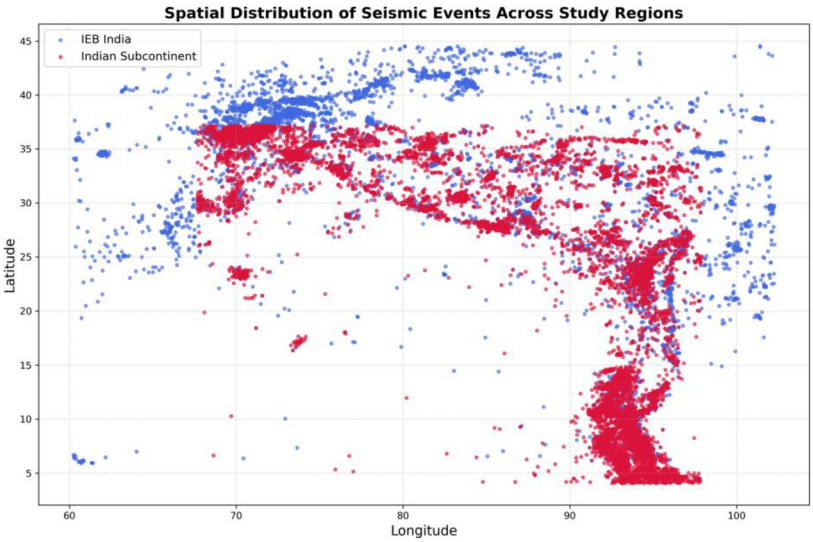

Fig. 2. Spatial distribution of earthquake events in the two datasets, showing substantial geographic overlap.

3.2 Data Preprocessing

Only attributes common to both catalogues were retained: latitude, longitude, depth, year, month, day, following the pipeline in Fig. 3. Because the IEB dataset spans 2015–2026 while the Subcontinent dataset spans 2000–2017 (only 2015–2017 shared), the absolute year was dropped to avoid the model learning catalogue/time-period identity rather than genuine seismic differences; latitude, longitude, depth, month, and day were retained. Magnitude ≥ 5.5 defines High Hazard (1), else Low Hazard (0). Features were standardized (scaler fit on training data only).

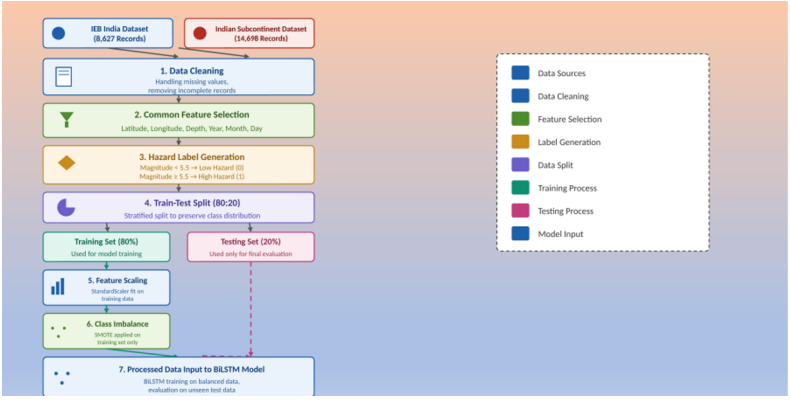

Fig. 3. Data preprocessing pipeline.

Class imbalance. Both datasets are severely imbalanced. SMOTE was applied to the training partition only, after the train–test split, leaving the test partition untouched. The severity is stated explicitly since it affects how Section 4 should be read: the IEB training imbalance is ~50:1 (6,767 vs. 134 events), and the Subcontinent training imbalance is ~37:1 (11,453 vs. 306 events) (Table 3). With only 134–306 real minority examples in five dimensions, SMOTE's k-NN interpolation operates on a sparse manifold, so the synthetic diversity added is smaller than post-SMOTE counts suggest — a plausible contributor to the low precision/recall reported later, alongside genuine class overlap in feature space.

Table 3. Class imbalance before/after SMOTE (training partitions; test left untouched).

Dataset (train)	Low-Haz.	High-Haz. (real)	Ratio (real)	High-Haz. (SMOTE)	Ratio (SMOTE)
IEB India	6,767	134	~50:1	6,767	~1:1
Ind. Subcontinent	11,453	306	~37:1	11,453	~1:1

3.3 Proposed BiLSTM Model

A Bidirectional LSTM processes the standardized 5-feature vector, followed by a dense (ReLU) layer, dropout (0.3), and a sigmoid output for binary hazard classification (Table 4).

Table 4. Architecture of the proposed BiLSTM model.

Layer	Output Shape	Parameters
Input	(5, 1)	–
Bidirectional LSTM	(None, 128)	33,792
Dense (ReLU)	(None, 64)	8,256
Dropout (0.3)	(None, 64)	0
Dense (Sigmoid)	(None, 1)	65
Total	–	42,113

Justification for the BiLSTM architecture. The five catalogue attributes are not a naturally ordered time series, so treating them as a length-5 pseudo-sequence merits justification. First, month/day encode seasonal periodicity, and lat/lon/depth jointly define a fault-relative position, so feature interactions are plausible without literal temporal order. Second, bidirectional gating is used here less for long-range temporal memory — the sequence is only 5 steps — and more as a data-driven feature-interaction mechanism, comparable to a structured alternative to explicit pairwise-interaction terms in a dense network. Third, this keeps the architecture comparable with the LSTM/BiLSTM backbones dominant in prior work [12, 9, 6, 1]. We do not claim this proves a BiLSTM is optimal for tabular features; dense/MLP and tree-based baselines on identical splits are named as future work (Sect. 5).

Training. TensorFlow-Keras; binary cross-entropy loss; Adam (lr=0.001); 30 epochs; batch size 32; class weights used to boost minority-class contribution during training.

3.4 Cross-Regional Evaluation Strategy

Four scenarios test within- and cross-region performance: (1) IEB → IEB and (3) Subcontinent → Subcontinent establish within-region baselines; (2) IEB → Subcontinent and (4) Subcontinent → IEB test forward and reverse cross-regional transfer.

3.5 Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC–AUC (threshold-independent ranking quality), alongside confusion matrices, where TP, TN, FP, FN denote true/false positives/negatives:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \qquad Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN}, \qquad F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \tag{2}$$

3.6 Generalization Retention Rate (GRR)

Accuracy is dominated by the majority (low-hazard) class and is therefore a weak indicator of cross-regional generalization for the minority (high-hazard) class. We formally define GRR as:

$$GRR_m = (M_cross\text{-}region / M_within\text{-}region) \times 100\% \tag{3}$$

where M is any metric from Sect. 3.5, M_within-region is Exp. 1 (IEB-anchored) or Exp. 3 (Subcontinent-anchored), and M_cross-region is the corresponding Exp. 2 or Exp. 4. GRR = 100% means full retention; < 100% means degradation.

Applying Eq. 2 to Tables 6–7 (Sect. 4) gives Table 5 and Fig. 4. Recall-based GRR the most operationally relevant, since it measures the fraction of real high-hazard events actually caught is ~32% (IEB-anchored) and ~45% (Subcontinent-anchored): roughly two-thirds and just over half of within-region high-hazard detection is lost on transfer. Accuracy-based GRR is 102% and 111% for the same transfers, exceeding 100% only because accuracy is computed over test sets that are >95% low-hazard in both catalogues. Precision/F1-based GRR (48–64%) fall in between. This confirms numerically that single-number accuracy reporting, common in prior work (Sect. 2).

Table 5. GRR for all four metrics and both transfer directions.

Metric	W-reg. (IEB)	X-reg. (IEB)	GRR (IEB)	W-reg. (Sub.)	X-reg. (Sub.)	GRR (Sub.)
Accuracy (%)	82.73	84.59	102.25%	78.88	87.83	111.35%
Precision (%)	2.52	1.29	51.19%	4.20	2.69	64.05%
Recall (%)	20.59	6.58	31.96%	32.89	14.71	44.72%
F1-score (%)	4.49	2.16	48.11%	7.45	4.55	61.07%

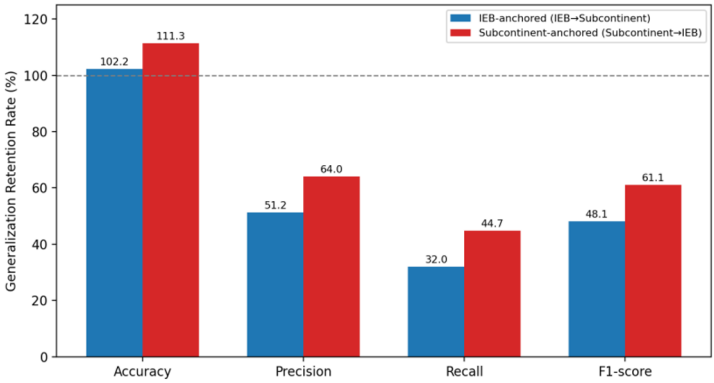

Fig. 4. Generalization Retention Rate (GRR) by metric and transfer direction. The dashed line marks 100% (full retention). Accuracy-based GRR exceeds 100% even as recall-based GRR falls to 32–45%.

4 Experimental Results and Discussion

The BiLSTM model was built in TensorFlow-Keras (Google Colab); both datasets used an 80/20 split, StandardScaler, and training-only SMOTE, with hyperparameters as in Table 4.

4.1 Within-Region and Cross-Regional Performance

Tables 6 and 7 report within-region baselines and cross-region transfer performance respectively, using the temporal-confound-corrected (year-dropped) feature set.

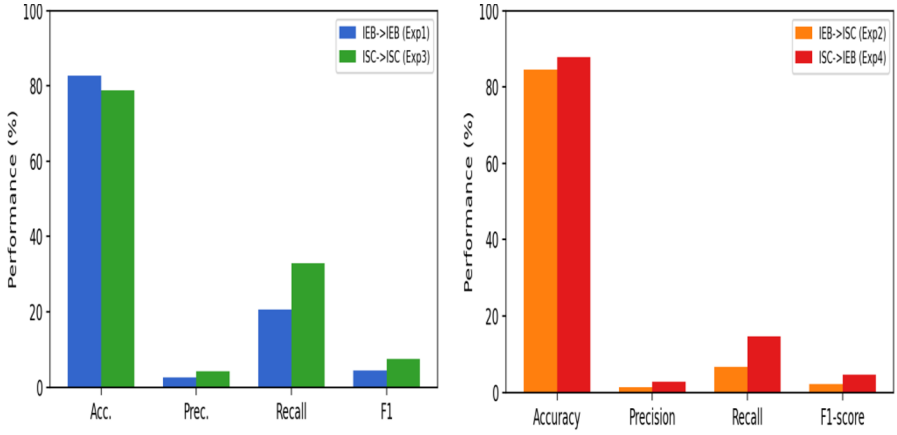

Fig. 5. Accuracy, precision, recall, and F1-score for the within-region (a, left) and cross-regional (b, right) experiments, using the temporal-confound-corrected (year-dropped) results.

Table 6. Within-region performance.

Exp.	Train	Test	Acc. %	Prec. %	Recall %	F1 %
1	IEB	IEB	82.73	2.52	20.59	4.49
3	Subcontinent	Subcontinent	78.88	4.20	32.89	7.45

Table 7. Cross-regional performance.

Exp.	Train	Test	Acc. %	Prec. %	Recall %	F1 %
2	IEB	Subcontinent	84.59	1.29	6.58	2.16
4	Subcontinent	IEB	87.83	2.69	14.71	4.55

As formalized in Sect. 3.6, the recall-based GRR (32%/45%) is a far more faithful summary of this discrepancy than accuracy alone, which is dominated by the majority class.

4.2 Confusion Matrix Analysis

Table 8 and Fig. 7 show per-experiment TP/TN/FP/FN. High-hazard recall is consistently low, and false negatives dominate the minority class across all four experiments, most severely under cross-regional transfer. Fig. 6 summarizes all four metrics across all four experiments as a single heatmap for quick visual comparison.

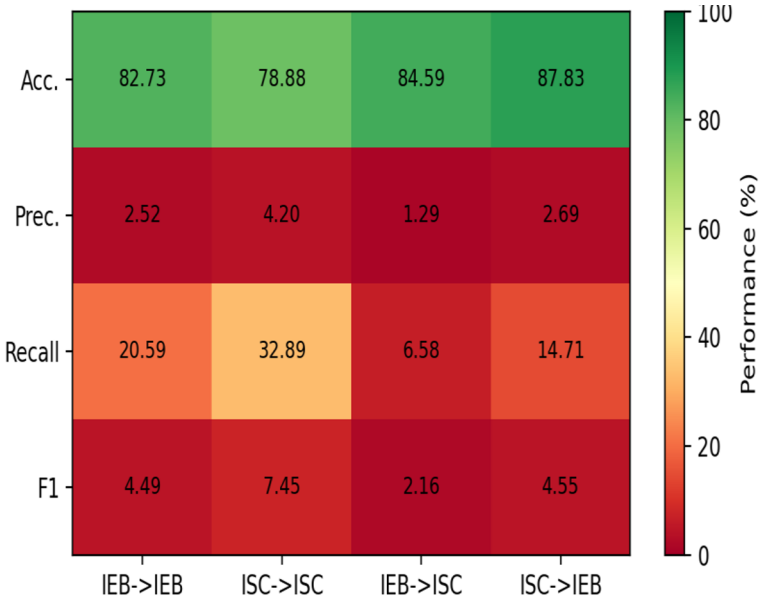

Fig. 6. Heatmap comparison of accuracy, precision, recall, and F1-score across all four within-region and cross-regional experiments.

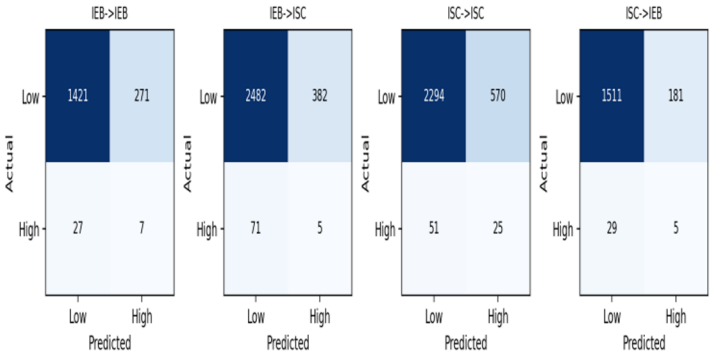

Fig. 7. Confusion matrices for within-region and cross-region experiments.

Table 8. Confusion matrix summary (TP, TN, FP, FN).

Experiment	TP	TN	FP	FN
IEB → IEB (Exp1)	7	1421	271	27
Subcontinent → Subcontinent (Exp3)	25	2294	570	51
IEB → Subcontinent (Exp2)	5	2482	382	71
Subcontinent → IEB (Exp4)	5	1511	181	29

4.3 Discussion and Practical Considerations

Cross-regional evaluation is essential: recall for high-hazard events collapses substantially under transfer even when accuracy remains stable or improves.

Geographic overlap. 90.4%/64.7% bounding-box overlap (Sect. 3.1) makes Exp2/Exp4 geographically weaker tests of generalization than transfer between fully disjoint tectonic regions would be. This makes the observed recall degradation a conservative estimate of the true cross-tectonic-region gap: if recall already collapses 55–68% between overlapping catalogues, a disjoint region would plausibly degrade at least as much, since the catalogues also differ in event density, curation practice, and depth-reporting convention. We frame this evaluation as generalization across differently curated, partially overlapping catalogues, and flag a genuinely disjoint external catalogue as a priority for future work.

Class imbalance. The severe imbalance (~50:1/~37:1, Table 3) compounds the overlap caveat: with only 134/306 real high-hazard examples, SMOTE interpolates within a small real sample rather than the true distribution, so the decision boundary is learned from a limited slice of seismicity — consistent with low precision (1.3–4.2%) even within-region. The recall collapse on transfer cannot be cleanly attributed to geography versus imbalance-driven instability under the current design; disentangling the two (via class-weighted loss, focal loss, or ensemble resampling) is named as future work.

Below-chance ROC-AUC. Experiment 2 (IEB → Subcontinent) produced ROC-AUC = 41.85%, below the 50% chance level, indicating an inverted rather than mere-

ly noisy probability ranking for this direction. Three non-exclusive hypotheses are offered: (i) sampling variance given only 76 true high-hazard test events; (ii) a decision function anti-correlated with hazard where Subcontinent high-hazard events cluster, due to IEB-specific fixed-depth/density conventions; (iii) local miscalibration of the SMOTE-interpolated boundary in feature-space regions rare in IEB training data. Bootstrapped AUC confidence intervals and feature-space visualization are needed to disambiguate these and are named as future work.

Taken together, these caveats mean the GRR values in Sect. 3.6 should be read as a lower-bound demonstration of the generalization gap rather than an isolated, precise measurement of geography alone. The central contribution — that accuracy alone is insufficient and a recall-sensitive metric such as GRR is needed — holds regardless, since the gap between accuracy-based GRR (102–111%) and recall-based GRR (32–45%) is large under any of the explanations above.

5 Conclusion and Future Work

A cross-region BiLSTM architecture was proposed for seismic hazard prediction using two heterogeneous earthquake datasets (IEB, Indian Subcontinent), with data pre-processing, SMOTE-based imbalance handling, and BiLSTM-based binary hazard classification. Unlike prior region-specific studies, this work systematically compares within-region and cross-region performance and introduces the Generalization Retention Rate (GRR, Sect. 3.6) to quantify degradation across all four evaluation metrics. While the model attains reasonable accuracy in all settings, cross-regional recall and F1 for high-hazard events drop substantially, showing that regional seismic differences materially affect deep learning generalizability and that accuracy alone is an insufficient robustness indicator.

Future work will incorporate additional, geographically disjoint catalogues to obtain a cleaner cross-tectonic-region measurement, and benchmark against transformer, dense/MLP, and tree-based baselines to isolate the contribution of the recurrent inductive bias. It will also compare SMOTE against class-weighted loss, focal loss, and ensemble resampling to disentangle imbalance from geographic shift, and conduct a dedicated diagnostic study of the below-chance ROC-AUC through bootstrapped confidence intervals and feature-space visualization. Finally, transfer learning, domain adaptation, and additional geophysical and tectonic features will be introduced to further improve cross-regional prediction of high-hazard earthquakes.

Acknowledgments. The authors declare that no funding, grants, or other support was received during the preparation of this manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data and Code Availability. The IEB India and Indian Subcontinent earthquake catalogues are derived from publicly accessible seismic bulletin sources; processed datasets and code are available from the corresponding author upon reasonable request.

References

1. Aryal, M., Tiwari, R.K., Paudyal, H.: Prediction of earthquakes in nepal and the adjoining regions using lstm. BMC Journal of Scientific Research 7 (2024)
2. Bondur, V.G., Voronova, O.S.: Identification of areas of probable seismic events using machine learning. Russian Journal of Earth Sciences (2025)
3. Chen, X., Peng, D., Li, L.: Forecasting future earthquakes with machine learning models based on seismic prediction zoning. Applied Sciences 15, 13116 (2025)
4. Deborah, W.J.W., Karishma, R., Pamela, D., et al.: Earthquake risk assessment using artificial intelligence: A review on traditional methods and artificial intelligence-based methods. In: Artificial Intelligence for Disaster Management (2026)
5. Elbes, M.W., AlZu'bi, S., Kanan, T.: Deep learning-based earthquake prediction technique using seismic data. In: Proc. IEEE MCNA (2023)
6. Fayaz, J., Astroza, R., Angione, C., et al.: Data-driven analysis of crustal and subduction seismic environments using interpretation of deep learning-based generalized ground motion models. Expert Systems with Applications 237, 121731 (2024)
7. Habek, G.C., Kahramanlı, H.: Regional analysis of earthquakes and earthquake magnitude estimation with machine learning techniques. Sinop University Journal of Natural Sciences (2024)
8. Kriuk, B., Kriuk, F.: Poseidon: Physics-optimized seismic energy inference and detection operating network. arXiv preprint (2026)
9. Muhammad, D., Ahmad, I., Khalil, M.I.K., et al.: A generalized deep learning approach to seismic activity prediction. Applied Sciences 13, 1598 (2023)
10. Nicolis, O., Carrasco, F., Peralta, B., et al.: Combining mra and lstm for predicting seismicity in chile. In: Proc. IEEE ICA-ACCA (2022)
11. Okoro, C.J., Hernández Ulloa, C., Peralta, B., et al.: Earthquake classification with deep learning across different subduction zones: Chile, japan, and mexico. In: Proc. IEEE ICPRS (2025)
12. Sadhukhan, B., Chakraborty, S., Mukherjee, S., et al.: Climatic and seismic data-driven deep learning model for earthquake magnitude prediction. Frontiers in Earth Science 11, 1082832 (2023)
13. Sreenath, V., Gatti, F., Jehel, P.: Breaking the black box: Inherently interpretable physics-constrained machine learning with weighted mixed-effects for imbalanced seismic data. Journal of Geophysical Research (2025)
14. Zhang, X., Zhang, M.: Universal neural networks for real-time earthquake early warning trained with generalized earthquakes. Communications Earth & Environment 5, 531 (2024)
15. Zhao, Y., Lv, S., Liu, P.: Advances in earthquake prevention and reduction based on machine learning: A scoping review. IEEE Access 13 (2025)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

