applied sciences

MDPI

Article

Forecasting Future Earthquakes with Machine Learning Models Based on Seismic Prediction Zoning

Xiaolin Chen [1], Daicheng Peng [2,3,*] and Li Li [1]

[1] Key Laboratory of Earthquake Geodesy, Institute of Seismology, China Earthquake Administration, Wuhan 430071, China; cxlcug@126.com (X.C.); ynulili@foxmail.com (L.L.)
[2] Wuhan Exploration Branch, Sinopec Petroleum Engineering Geophysics Co., Ltd., Qianjiang 433100, China
[3] Key Laboratory of Exploration Technologies for Oil and Gas Resource, Yangtze University, Ministry of Education, Wuhan 430100, China
* Correspondence: pengdc_geo@126.com

Abstract

Predicting future seismic trends and occurrence of earthquakes remains a long-standing challenge in seismology. Despite substantial efforts to unravel the physical mechanisms underlying earthquake occurrence, currently, no well-defined physical or statistical model is capable of reliably predicting major earthquakes. However, machine learning methods have demonstrated exceptional proficiency in identifying patterns within large-scale datasets, offering a promising avenue for enhancing earthquake prediction performance. Within the framework of machine learning, this study has developed a feature extraction method based on seismic prediction zoning, improving the effectiveness of machine learning-based earthquake prediction. The research findings indicate that the ensemble learning Stacking method, which is based on seismic prediction zoning, exhibits superior performance and high robustness in predicting the annual maximum earthquake magnitude. Additionally, the long short-term memory (LSTM) method demonstrates commendable performance within specific tectonic zones (e.g., the southwestern Yunnan region), providing valuable guidance for analyzing seismic trends in these regions.

Keywords: machine learning; seismic indicators; earthquake prediction; seismic prediction zoning

check for updates

Academic Editor: Sergio Baselga

Received: 12 November 2025
Revised: 5 December 2025
Accepted: 9 December 2025
Published: 12 December 2025

Citation: Chen, X.; Peng, D.; Li, L. Forecasting Future Earthquakes with Machine Learning Models Based on Seismic Prediction Zoning. *Appl. Sci.* **2025**, *15*, 13116. https://doi.org/10.3390/app152413116

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

China has frequent continental earthquakes and severe earthquake disasters. The Chinese mainland represents one of the most seismically active regions within a plate interior and is characterized by frequent earthquakes, widespread distribution, high intensity, and severe disasters. Mitigating the disasters caused by earthquakes holds significant practical importance for the national economy, population, and society as a whole. Earthquake prediction has long been a formidable challenge, attributable to several factors. First, earthquakes result from the complex interactions between tectonic plates, faults, and other geological factors, making precise prediction of their occurrence time and magnitude extremely challenging. Second, the scarcity of long-term and extensive data poses a significant obstacle to earthquake prediction. The recurrence intervals of major earthquakes are often lengthy (spanning centuries to millennia) and identify trends and patterns over such extended timescales is difficult. Owing to complex geophysical interactions, variability in seismic activity, insufficient comprehensive data, and limitations in current technological approaches, earthquake prediction remains a complex and arduous research domain [1].

In seismological research, regional seismic activity may exhibit anomalies such as seismic gaps, seismic belts, and earthquake swarms, which reflect the regional stress state. Case studies have demonstrated that enhanced seismic activity commonly precedes major earthquakes, with seismic activity anomalies being widely used as indicators to judge future seismic trends in a region. In the study of the temporal, spatial, and intensity distribution characteristics of seismic activity, Gutenberg and Richter [2] first proposed the magnitude-frequency relationship, known as the Gutenberg–Richter (G-R) law.

With the widespread adoption of machine learning techniques as indispensable tools, seismologists have gradually explored the application of machine learning algorithms in seismic activity prediction. In spatial regional prediction, Aslam et al. utilized clustering hotspot analysis machine learning methods to identify seismic hotspot regions in northern Pakistan [3]. For time series prediction, Kaftan et al. validated the performance of different neural networks, including multilayer perceptron neural networks (MLPNN), radial basis function neural networks (RBFNN), and adaptive neuro-fuzzy inference systems (ANFIS), in seismic time series applications [4]. Wang et al. incorporated spatial information into time series data and utilized long short-term memory (LSTM) networks to discover spatiotemporal correlations among earthquakes [5]. Additionally, raw earthquake catalog data were extracted into seismic activity characteristic parameters, transforming them into input patterns recognizable by machine learning models (such as neural networks and support vector machines). Seismic activity characteristics refer to statistical parameters obtained through the analysis of seismological observational data, which are capable of quantitatively describing the spatiotemporal intensity characteristics of seismic activity within a certain spatiotemporal range. These include statistical parameters describing the temporal distribution of earthquake sequences, spatial distribution of earthquakes, average seismic activity intensity, rate of seismic strain release, and average recurrence period of earthquakes [6]. Panakkat and Adeli [7] first used these seismic activity characteristic parameters as inputs to construct a neural network model for predicting the maximum earthquake magnitude over a subsequent period. Then, they employed methods such as RBFNNs, recurrent neural networks (RNN), and probabilistic neural networks (PNN) to establish nonlinear models between earthquake magnitudes and seismic activity indicators [8]. This was later extended to using neural network methods to predict both the occurrence time and magnitude of earthquakes [9]. Other scholars have made improvements based on these seismic activity parameters. First, additional characteristic parameters were added, including calculating the b-value increase over certain earthquake intervals, using probability density functions to record the probability of magnitudes greater than or equal to the target magnitude, and recording the maximum magnitude in the most recent week [10,11]. Second, improvements have been made in terms of methodologies, including ensemble algorithms which integrate neural network-based algorithms, genetic programming (GP) algorithms, and Adaptive Boosting (AdaBoost) algorithms [12]. Additionally, the dendritic cell algorithm in the field of artificial immunity [13] and deep learning-based approaches [14] have been introduced. Asencio-Cortés et al. evaluated the application of generalized linear models (GLMs), gradient boosting machines (GBMs), and random forests (RFs) in earthquake magnitude prediction [15]. For earthquakes with magnitudes ranging from 3 to 7, the RF method demonstrated the best performance, achieving a mean absolute error (MAE) of 0.6. Mallouhy et al. employed RFs, naive Bayes, logistic regression, MLPNNs, AdaBoost, K-nearest neighbors, support vector machines, and classification and regression trees to predict major earthquakes [16]. Among these methods, RFs achieved the highest accuracy. However, a major limitation of RFS is that when the number of trees becomes excessively large, the algorithm may become too slow to meet the demands of real-time prediction.

Sadhukhan [17] attempted to predict the magnitude of the next impending earthquake by analyzing eight mathematically calculated seismicity indicators. The study utilized three widely applied deep neural network models, namely LSTMs, bidirectional LSTMs (Bi-LSTMs), and Transformer models, to predict the magnitude of impending earthquakes in a given seismic region based on eight seismicity indicators calculated from past significant earthquake events greater than a predetermined threshold magnitude.

The core of artificial intelligence lies in machine learning, which serves as the key technology driving this transformation. The primary advantage of machine learning is its ability to identify functional relationships between vast amounts of data and their corresponding labels, which may be high-dimensional and nonlinear. This is the case in earthquake prediction, where the complexity is beyond human comprehension. Traditionally, earthquake prediction has relied on the empirical judgment of experts, leading to randomness and uncertainty. However, the application of machine learning to earthquake prediction offers a promising solution for obtaining more accurate and reliable results.

Nevertheless, owing to the scarcity of historical small earthquake data, many studies [18] have only utilized earthquake catalogs with magnitudes of 3 or above, resulting in limited training samples. This makes it difficult to train robust predictive models that can effectively generalize event patterns and prior knowledge. In this study, we compiled an earthquake catalog for the Yunnan region beginning from 1985, including earthquakes with magnitudes of 2 or above to capture more foreshock information. Additionally, as traditional sample feature extraction methods have failed to adequately consider regional active tectonic characteristics. Therefore, our feature extraction approach based on seismic prediction zoning can enhance the performance of machine learning prediction methods by incorporating implicit geographical-geological features.

2. Data and Methods

The selected earthquake catalog (Figure 1) included earthquakes with magnitudes of 2 or above in Yunnan, mainland China (97° E–106° E, 21° N–29° N) since 1985. Yunnan is located on the southeastern margin of the Tibetan Plateau. Since the Cenozoic, it has been subjected to the combined effects of the eastward migration of crustal materials from the Tibetan Plateau and the wedge-shaped intrusion of the Assam promontory [19]. Consequently, this region has experienced intense neotectonic deformation and seismic activity, making it the most prominent strong earthquake-prone area in mainland China. Influenced by regional tectonics, seismic activity in Yunnan exhibits high frequency, large magnitudes, and widespread distribution (Figure 2). Specifically, for the Sichuan-Yunnan region, Jiang [20] delineated 15 seismic prediction zones. These seismic prediction zones are delineated based on the distribution and patterns of historical earthquakes, combined with regional geological structural conditions. From the 1996 Lijiang Ms7.0 earthquake to the present, Yunnan, a region with historically intense seismic activity, has been in a state of seismic quiescence for earthquakes with $M \geq 7.0$. As of 22 September 2025, the duration of seismic quiescence for earthquakes with $M \geq 7.0$ in Yunnan has reached 29.68 years, significantly exceeding the longest historical quiescence period since 1900 or 1887. Therefore, this study aimed to address the current phenomenon of prolonged seismic quiescence for strong earthquakes in Yunnan. Based on the characteristics of strong earthquake activity in the region, we employed new artificial intelligence methods to conduct an in-depth analysis and summary of the future maximum earthquake magnitude in Yunnan across spatiotemporal scales.

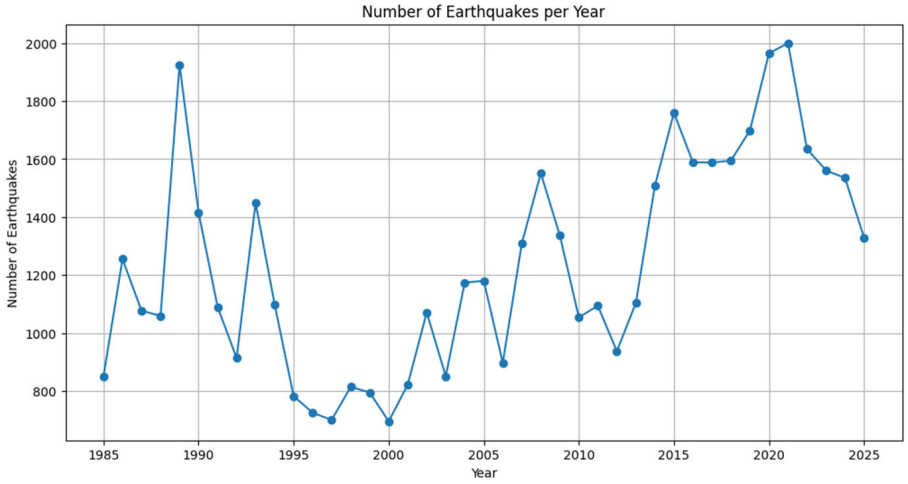

Figure 1. Number of Earthquakes by Year.

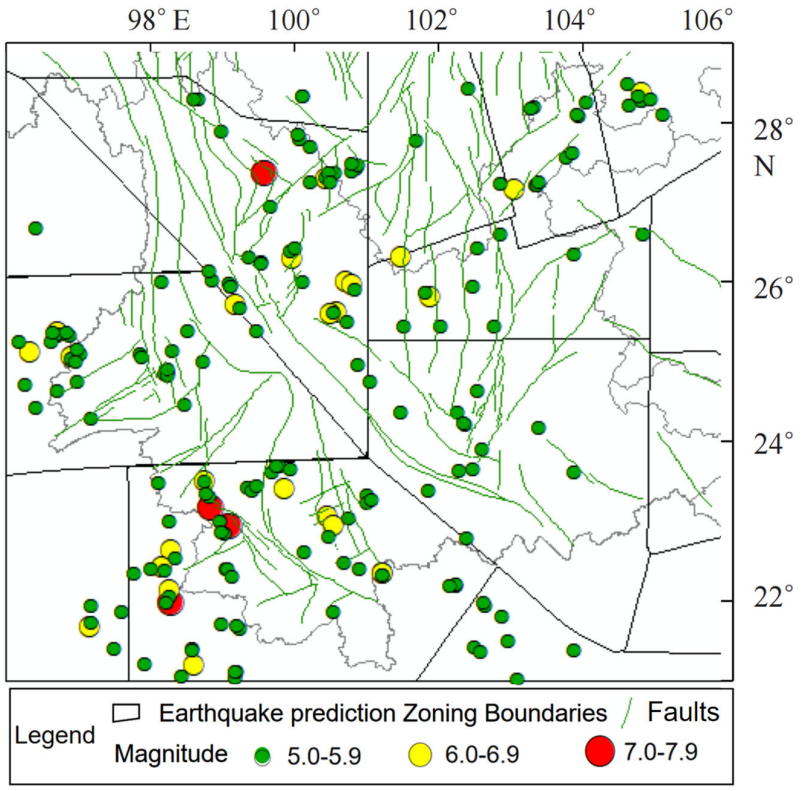

Figure 2. Distribution of M $\geq$ 5.0 earthquakes in Yunnan for the study period.

2.1. Estimating Completeness Magnitude

The earthquake catalog serves as the foundation for seismicity analysis. If it includes a large number of low-magnitude events that have not been fully recorded, biases can occur in the statistical results, such as overestimating the frequency of small earthquakes and underestimating the probability of large earthquakes. The magnitude of completeness (Mc) refers to the minimum magnitude at which an earthquake catalog can record events completely within a given spatiotemporal range. Earthquakes with magnitudes below this threshold may be overlooked or underestimated owing to insufficient monitoring capabilities. The purpose of conducting a completeness magnitude test is to evaluate the reliability of the earthquake monitoring system and ensure the data quality of subsequent seismicity analyses, such as frequency-magnitude distribution studies.

Three of the most commonly used approaches are the Maximum Curvature (MAXC) method [21], goodness-of-fit (GFT) test [22], and by b-value stability (MBS) method [23]. The MAXC method is the most straightforward approach, simply using the magnitude bin from the non-cumulative Frequency-Magnitude Distribution (FMD) (orange dots in Figure 3) containing the highest frequency of events However, this approach has the highest potential for underestimating when detection rates vary (e.g., due to changing station availability), typically giving the lowest estimate of all methods [24].

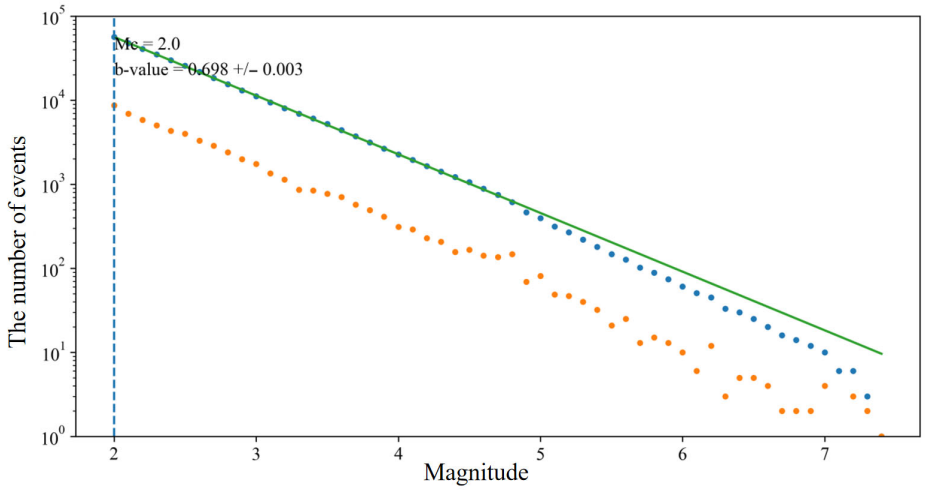

Figure 3. Estimating completeness magnitude (The *X*-axis represents magnitude. The *Y*-axis represents the number of events. The blue scatter points denote the cumulative events curve. The orange scatter points represent the interval events curve. The green curve is the theoretical fitting curve, plotted based on the Gutenberg-Richter (G-R) law).

Through MAXC analysis, we determined that the magnitude of completeness (Mc) for the study area is 2.0. Therefore, we utilized an earthquake catalog containing events with magnitudes $\geq$ 2.0 for our analysis. The relevant results are presented in Figure 3.

2.2. Seismic Indicators

This attempt represents an approach of "predicting earthquakes based on earthquake catalogs", that is, making predictive efforts based on the variations in seismicity characteristics reflected in the earthquake catalog. Seismicity primarily refers to the spatiotemporal distribution and intensity characteristics of seismic activity within a specific spatial domain (typically a seismic zone or belt) and time frame, as well as the differences in seismic activity across different spatial or temporal domains [25]. We selected eight seismic prediction factors for earthquake forecasting: seismic frequency within the spatiotemporal window (N), mean magnitude (M_{mean}), slope of the G-R relation (b), intercept of the G-R relation (a), magnitude difference (ΔM), release rate of the square root of seismic energy ($\mathrm{d}E^{1/2}$), root mean square deviation of the regression line (η), mean recurrence interval of characteristic events (μ). These eight characteristic factors describe the spatial distribution of earthquakes and reflect the spatial locations of seismic activity, clustering, zoning patterns of earthquakes, energy release centers. To a certain extent, these parameters are related to spatially distributed features observed in Chinese empirical earthquake prediction, such as seismic zones and gaps.

2.3. Analytical Methods and Metric

This study employed several classic machine learning methods for comparative analysis, including RF [26], Gradient Boosting Decision Tree (GBDT) [27], Stacking (an ensemble learning approach) [28], and LSTM networks [29].

RF is an ensemble learning method based on bootstrap aggregating (Bagging). It constructs multiple decision trees and determines the final output through voting (for classification tasks) or averaging (for regression tasks). During tree construction, each tree is trained on a bootstrap sample (i.e., a random subset of the original data with replacement). At each node split, a random subset of features is considered, and the split is determined using criteria such as the Gini index or information gain.

GBDT is an ensemble learning algorithm that falls under the Boosting category. It derives its final predictions by aggregating the predictions of multiple decision trees. The core idea is to train decision trees iteratively. For each iteration, the residuals (or the negative gradients of the loss function) between the current predictions and the true values are computed. A new decision tree is then trained to fit these residuals, enabling the new predictions to reduce the overall error.

Stacking is a meta-learning approach that combines predictions from multiple base models to train a meta-model for final predictions. In this study, a Stacking ensemble model was employed, using RF and GBDT as base learners.

The LSTM model is distinguished by its unique cell state and gating mechanisms, namely the forget, input, and output gates. These gates collectively regulate the flow and updating of information, enabling LSTM to effectively capture and process long-term dependencies in sequential data.

Eight seismic indicators outlined in Section 2.2 were selected as input features for the machine learning model, which was designed to predict the maximum earthquake magnitude within the subsequent year. The input data were derived using a spatiotemporal sliding window approach tailored to capture temporal trends and spatial distributions. For example, assuming the current date as 1 October 2022, the eight seismic indicators were computed based on earthquake catalogs spanning from 1 October 2020, to 1 October 2022. The corresponding target variable (label) was defined as the maximum earthquake magnitude observed between 1 October 2022, and 1 October 2023. This process was repeated monthly using a rolling window, with seismic indicators calculated for each region independently. The resulting dataset was subsequently partitioned into training, validation, and test sets to facilitate model development and performance evaluation. In this study, we employed the Mean Squared Error (MSE) [30] and R-squared Score (R^2) [31] to evaluate the model performance.

MSE is calculated as the average of the squared differences between the predicted and actual values, serving as a metric to quantify the accuracy of model predictions.

The R^2 quantifies the proportion of the variance in the target variable that is explained by the model, reflecting the goodness of fit of the model to the data.

MSE focuses on the magnitude of absolute errors, whereas R^2 emphasizes the proportion of relative error variance explained. Combining these two metrics provides a comprehensive evaluation of model performance: a smaller MSE and an R^2 closer to one indicate a better-performing model.

3. Results and Discussion

3.1. Comparison of Feature Extraction Based on Longitude-Latitude Grid Division and Seismic Prediction Zoning

Feature extraction based on longitude-latitude grid division refers to the following approach: Spatial Window: The samples are partitioned using a $2^\circ \times 2^\circ$ longitude-latitude grid (Figure 4). A sliding overlapping window with an overlap of 1° is employed. Temporal Window: For earthquake prediction, the maximum earthquake magnitude within the upcoming one-year period is forecasted using seismic data from the preceding two years. A sliding temporal window with a step of one year is utilized. Subsequently, seismicity

parameters are calculated for each spatial window. Based on these parameters, RF, GBDT, and Stacking ensemble models with RF and GBDT as base learners are constructed.

Feature extraction based on seismic prediction zoning operates on the premise that tectonic blocks exert control over strong earthquakes in the Chinese mainland. Therefore, this study categorized the spatial relationships between temporally adjacent strong earthquakes into two types: those occurring within the same tectonic block and those in different tectonic blocks, based on their tectonic associations. Specifically, for the Sichuan-Yunnan region, Jiang delineated 15 seismic prediction zones (Figure 5). Subsequently, seismicity parameters were calculated for each spatial window. Based on these parameters, RF, GBDT, and Stacking ensemble models with RF and GBDT as base learners were constructed.

Feature extraction was conducted to generate sample sets using two approaches: one based on longitude-latitude grid division and one considering seismic prediction zoning. Subsequently, RF, GBDT, and Stacking ensemble models were constructed based on these two sample sets.

From the results (Table 1), in the sample set divided by longitude-latitude grids, the RF Regressor exhibited an MSE of 0.2262 and R^2 value of 0.6561. The GBDT Regressor showed an MSE of 0.4694 and R^2 value of 0.2864. The Stacking Regressor performed the best, with an MSE of 0.2087 and R^2 value of 0.6827. The Stacking Regressor outperformed both individual base models in terms of both MSE and R^2.

From the perspective of the sample set partitioned by seismic prediction zones, the Stacking ensemble model consistently exhibits superior predictive performance across all evaluation metrics. However, across all three models, the sample set generated through seismic prediction zoning outperforms the one generated through longitude-latitude grid division. The sample set significantly enhances model performance: for all models trained on the active tectonics-derived sample set, MSE decreases by an average of 24.6%, and the R^2 value increases by an average of 12.3%.

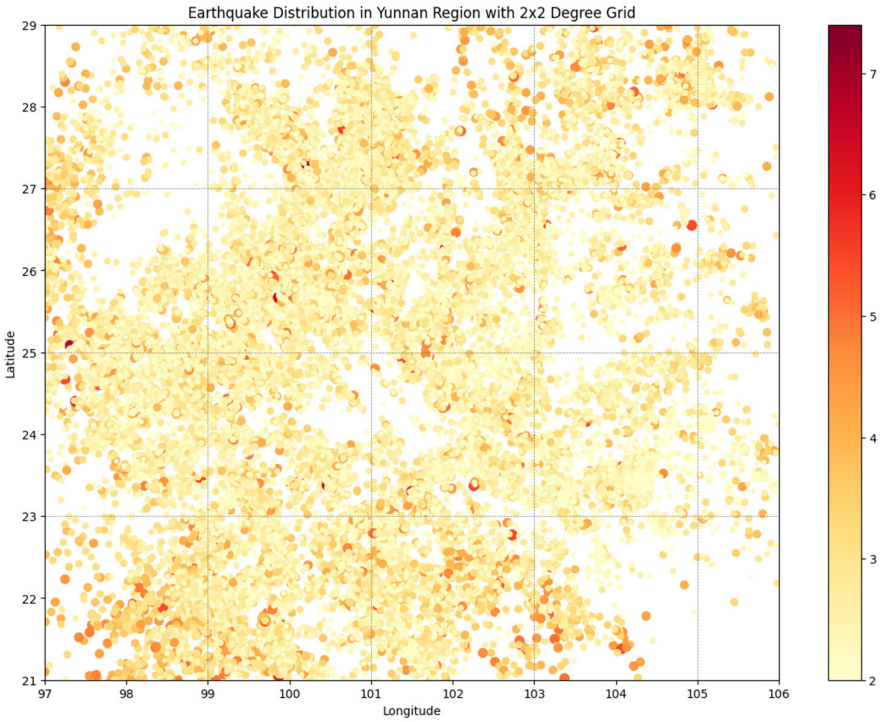

Figure 4. Earthquake distribution in Yunnan Region with 2° × 2°.

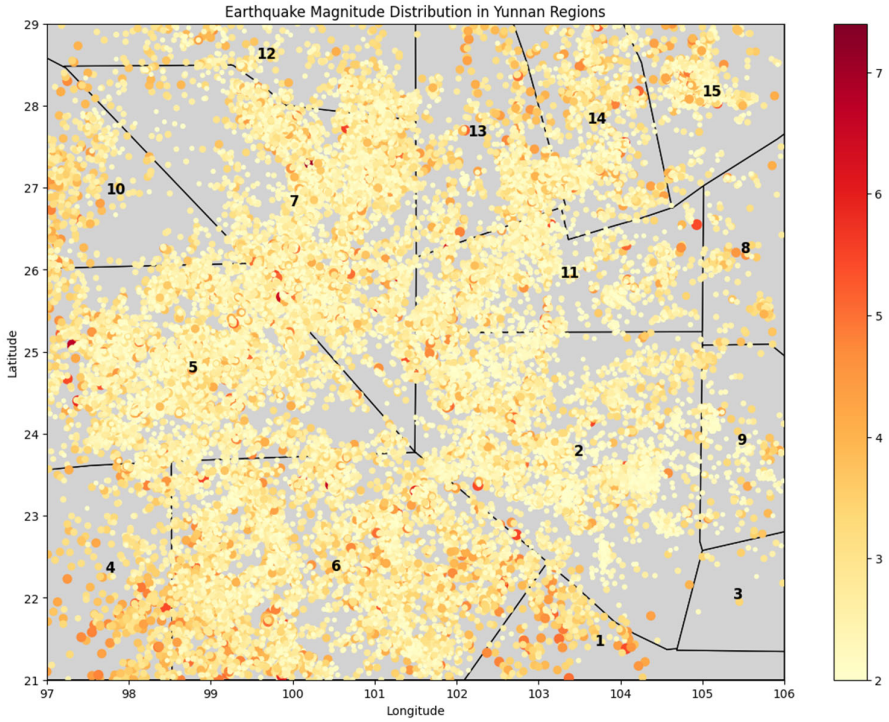

Figure 5. Earthquake distribution in Yunnan Region with seismic prediction zoning.

Table 1. Seismic indicators calculated from seismic catalog.

Indicators	Description	Mathematical Expression
N	Seismic frequency	
M_{mean}	Mean magnitude	$M_{mean} = \sum M_i / N$
b	Slope of the G-R relation	$b = \frac{1}{\ln(10)(M_{mean} - M_c)}$
a	Intercept of the G-R relation	$a = log_{10}(\frac{N(M)}{T}) + bM_c; M \geq M_c$
ΔM	Magnitude difference	$\Delta M = M_{\max,observed} - M_{\max,expected}$
$dE^{1/2}$	Release rate of the square root	$dE^{1/2} = \frac{\sum E^{1/2}}{T}$
η	Root mean square of the regression line	$\eta = \frac{\sum (\log_{10} N_i - (a - bM_i))^2}{(n-1)}$
μ	Mean recurrence interval	$\mu = \frac{\sum t_i}{n}$

The experimental results (Table 1) indicate that the spatial construction method of the sample set may exert a greater influence on model performance than the algorithm itself. In regions with pronounced geological activity, the active tectonics-based sample set can improve the prediction accuracy of tree-based models such as RF by approximately 33.8% by incorporating implicit geospatial-geological features. Enhanced spatial correlation: Active tectonic features (geological units) may introduce implicit geospatial-geological characteristics, enabling the model to capture spatiotemporal dependencies that traditional longitude-latitude grids fail to represent. If the study area exhibits distinct geological activity (e.g., seismic belts and mineralized zones), priority should be given to constructing sample sets based on active tectonics, as this can significantly enhance prediction accuracy.

In Figure 6, the dashed black line represents where the predicted value equals the actual value. Ideally, for a perfect model, all the points would fall directly on this line. The points in the Stacking plot appear to be the most tightly clustered around the diagonal line, indicating the best overall predictive performance among the three models. This aligns with the lower MSE and higher R-squared values observed for the Stacking Regressor.

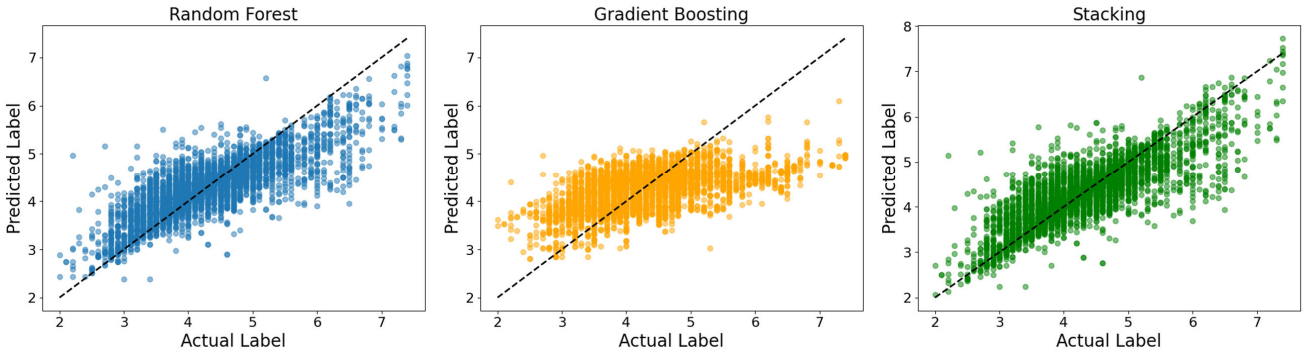

Figure 6. Predicted versus actual values for each model on the test set. (Each plot shows the actual 'Label' values on the *x*-axis and the predicted 'Label' values on the *y*-axis for one of the models).

3.2. Comparative Analysis of Machine Learning Methods Based on Different Seismic Prediction Regions

Based on the aforementioned analysis, the Stacking model demonstrates superior performance. Therefore, this study employed Stacking to compare with LSTM for analyzing the impacts of different seismic prediction regions.

Separate Stacking Regressor and LSTM models were trained for each of the 14 regions. Owing to the insufficient number of earthquakes, the models cannot be established in region No. 3. Validation loss (MSE) was calculated for both models in each region. A comparison table was generated displaying the validation loss for each model per region (Table 2).

Table 2. Comparison of machine learning results based on different feature extraction methods.

Methods	**Longitude-Latitude Grid**		**Seismotectonic Zoning**	
Metrics	**MSE**	**R-Squared**	**MSE**	**R-Squared**
Random Forest Regressor	0.2262	0.6561	0.167303	0.787097
Gradient Boosting Regressor	0.4694	0.2864	0.362549	0.538635
Stacking Regressor	0.2087	0.6827	0.202732	0.742012

The MSE and R^2 of the LSTM model exhibit significant fluctuations across different regions (Table 3). The LSTM model for region No. 5 demonstrates the best performance, with the lowest MSE and relatively small prediction errors compared to other regions. The LSTM models for regions No. 5, 6, and 7 also perform well. In contrast, the LSTM models for other regions perform poorly, with extremely high MSE values, indicating highly unsatisfactory prediction effects in these regions.

Regions No. 1 and 15 are located at plate boundaries, characterized by a relatively small number of minor earthquake catalogs and susceptibility to interference from external plates. This results in sample data that fail to comprehensively reflect the regional seismic activity patterns. During model training, there is a lack of sufficient effective information, making it difficult to capture the complex mechanisms of earthquake occurrence. Consequently, the prediction effects are extremely poor (e.g., both the LSTM and Stacking models exhibit high MSE values). This demonstrates that data quality (integrity and representativeness) is fundamental to model performance. Even with complex model structures, achieving ideal results is challenging if the samples exhibit systematic biases (e.g., external interference and data sparsity).

Regions No. 5, 6, and 7 account for 40% of the study area in Yunnan and are situated within the frequently active seismic zone in the interior. These regions contain 60% of the historical minor earthquake catalogs and are less influenced by external plates. Their data more accurately reflect the regional seismic activity characteristics, providing a more stable and reliable training basis for the models. Therefore, even with the same model structure,

the prediction effects in these regions are significantly superior to those in boundary regions (e.g., the MSE of the LSTM model for region No. 5 is 0.15, significantly lower than that in other regions). This validates the critical impact of local representativeness and data volume of samples on the generalizability of the model.

Table 3. Results comparison among 14 seismic predictioregions.

Region No	LSTM_MSE	Stacking_MSE	SumN
1	3.460	1.074	4583
2	0.616	0.606	128,796
4	0.897	1.074	14,845
5	*0.150*	*0.345*	*241,582*
6	*0.866*	*1.102*	*219,911*
7	*0.852*	*1.403*	*222,303*
8	1.380	0.999	3159
9	0.239	0.255	5377
10	0.793	0.608	11,270
11	0.499	0.501	68,456
12	0.553	0.524	34,621
13	0.540	0.210	45,436
14	0.451	0.484	55,671
15	1.915	0.810	85,196

A visual assessment of the model's spatial accuracy was conducted for the first time step (1 August 2020) in the test set. One year after this time window, on 7 May 2021, the Yangbi earthquake occurred in the region No. 7. From the perspective of spatial predictions, this region exhibited a relatively higher trend of seismic activity compared to other regions. From the visual comparison between predicted values and actual values (Figure 7), it is evident that the model can capture the general overall spatial distribution pattern, though notable local discrepancies exist.

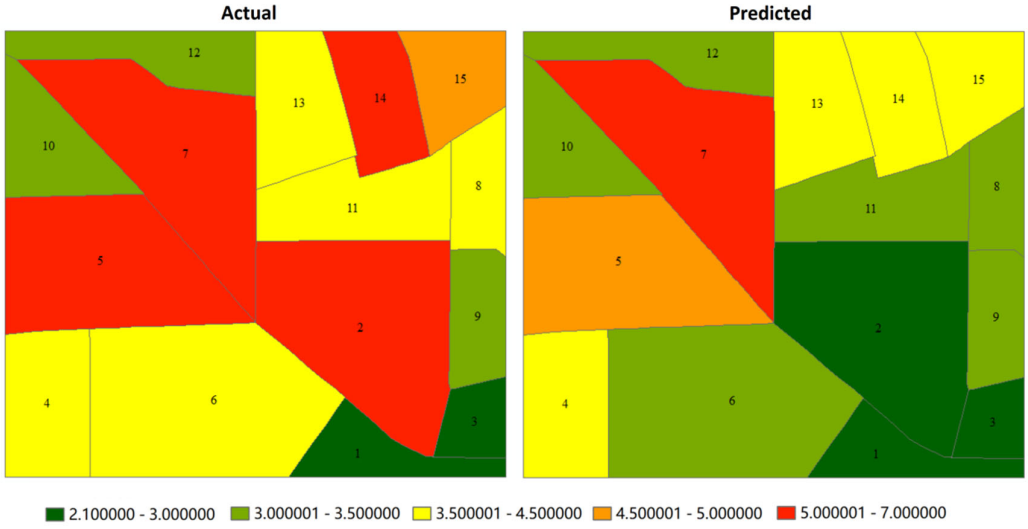

Figure 7. Visualization of predicted vs. actual values in space.

The model demonstrated high accuracy in certain regions. For example, in region No. 7 and region No. 5, the predicted values closely aligned with the observed actual values. However, significant discrepancies were also observed, including cases where predictions were consistently lower than actual values in certain regions, such as region No. 2 and region No.14. In summary, although the model provides reasonable spatial approximations for earthquake magnitudes, it faces challenges in accurately predicting

regions with lower seismic activity and exhibits varying degrees of underestimation across different regions.

The above analysis shows that the matching between regional geological characteristics and model applicability is crucial. Seismic activity at plate boundaries is controlled by the interactions of multiple plates, resulting in complex and highly nonlinear mechanisms. Existing models (LSTM and Stacking) may not be able to fully capture such cross-plate dynamic processes, leading to large prediction errors. In contrast, seismic activity in the interior regions of Yunnan may be more controlled by local tectonics, with more consistent data characteristics. Models can learn the underlying patterns of earthquake occurrence more accurately in such regions, indicating that models are more applicable in regions with simple geological conditions and homogeneous data characteristics. The complexity of seismic activity is closely related to the differences in regional geological structures. Seismic mechanisms, data characteristics, and prediction challenges vary significantly across different tectonic regions. Therefore, conducting research based on seismic prediction zoning is not only key to improving earthquake prediction accuracy but also an inevitable choice for scientifically understanding seismic activity patterns.

4. Conclusions

This study explored the enhance prediction performance using machine learning methods and proposed a seismotectonic zoning-based approach for earthquake prediction. Compared to the traditional feature extraction method that solely relies on latitude-longitude grids, the proposed approach, which is based on seismic prediction zoning, more effectively captures the spatiotemporal dependencies associated with earthquake occurrence by incorporating implicit geographical and geological features. The study finds that earthquake prediction based on seismic prediction zoning significantly outperforms the traditional latitude-longitude grid-based approach. On average, the prediction errors (MSE) of all tested models decreased by 24.6%, whereas the explained variance (R^2) increased by 12.3%. Notably, tree-based models (e.g., RF) demonstrated a remarkable improvement in prediction accuracy, with an approximately 33.8% increase. The LSTM approach exhibits satisfactory effectiveness within particular seismotectonic zones. Therefore, integrating geological structural knowledge into the machine learning framework and constructing sample sets based on seismic prediction zoning represent effective strategies for substantially improving the accuracy and reliability of earthquake prediction models. For regions with significant geological activity, this method should be prioritized.

Author Contributions: Conceptualization, X.C. and D.P.; Methodology, X.C. and D.P.; Software, X.C. and D.P.; Validation, X.C.; Formal analysis, X.C. and D.P.; Investigation, D.P. and L.L.; Resources, X.C.; Data curation, X.C. and L.L.; Writing—original draft, X.C.; Writing—review and editing, X.C., D.P. and L.L.; Visualization, X.C.; Supervision, D.P.; Project administration, D.P.; Funding acquisition, X.C. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science for Earthquake Resilience, China Earthquake Administration (No. XH24026YC, No. XH25019C) and the Sinopec Petroleum Engineering Geophysics Co., Ltd. Wuhan Exploration Branch 2025 Research Project (No. CIBG250096).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Daicheng Peng was employed by the company Sinopec Petroleum Engineering Geophysics Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Meng, L.Y.; Hu, F.; Zang, Y.; Zhang, Q.; Li, M. Introduction to the Artificial Intelligence Earthquake Prediction Models in Sichuan–Yunnan Region. *J. Seismol. Res.* **2026**, *49*, 43–50.
2. Gutenberg, B.; Richter, C.F. Earthquake Magnitude, Intensity, Energy, and Acceleration. *Bull. Seismol. Soc. Am.* **1942**, *32*, 163–191. [CrossRef]
3. Aslam, B.; Zafar, A.; Qureshi, U.A.; Ahmad, R.; Ali, M.; Khan, S. Seismic Investigation of the Northern Part of Pakistan Using Statistical and Neural Network Algorithms. *Environ. Earth Sci.* **2021**, *80*, 59. [CrossRef]
4. Kaftan, I.; Şalk, M.; Şenol, Y. Processing of Earthquake Catalog Data of Western Turkey with Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System. *Arab. J. Geosci.* **2017**, *10*, 243. [CrossRef]
5. Wang, Q.; Guo, Y.; Yu, L.; Zhang, X.; Chen, J. Earthquake Prediction Based on Spatio-Temporal Data Mining: An LSTM Network Approach. *IEEE Trans. Emerg. Top. Comput.* **2017**, *8*, 148–158. [CrossRef]
6. Shi, X.Y. Research on Earthquake Prediction Based on Machine Learning Regression Algorithms and Its Application in the China Earthquake Scientific Experimentation Site. Ph.D. Thesis, Institute of Earthquake Prediction, China Earthquake Administration, Beijing, China, 2021.
7. Panakkat, A.; Adeli, H. Neural Network Models for Earthquake Magnitude Prediction Using Multiple Seismicity Indicators. *Int. J. Neural Syst.* **2007**, *17*, 13–33. [CrossRef]
8. Adeli, H.; Panakkat, A. A Probabilistic Neural Network for Earthquake Magnitude Prediction. *Neural Netw.* **2009**, *22*, 1018–1024. [CrossRef]
9. Panakkat, A.; Adeli, H. Recurrent Neural Network for Approximate Earthquake Time and Location Prediction Using Multiple Seismicity Indicators. *Comput.-Aided Civ. Infrastruct. Eng.* **2009**, *24*, 280–292. [CrossRef]
10. Asencio-Cortés, G.; Scitovski, S.; Scitovski, R.; Štambuk, M. Temporal Analysis of Croatian Seismogenic Zones to Improve Earthquake Magnitude Prediction. *Earth Sci. Inform.* **2017**, *10*, 303–320. [CrossRef]
11. Asim, K.M.; Awais, M.; Martínez-Álvarez, F.; Hussain, A.; Ali, S. Seismic Activity Prediction Using Computational Intelligence Techniques in Northern Pakistan. *Acta Geophys.* **2017**, *65*, 919–930. [CrossRef]
12. Asim, K.M.; Idris, A.; Iqbal, T.; Martínez-Álvarez, F.; Hussain, A. Seismic Indicators-Based Earthquake Predictor System Using Genetic Programming and AdaBoost Classification. *Soil Dyn. Earthq. Eng.* **2018**, *111*, 1–7. [CrossRef]
13. Zhou, W.; Dong, H.; Liang, Y. The Deterministic Dendritic Cell Algorithm with Haskell in Earthquake Magnitude Prediction. *Earth Sci. Inform.* **2020**, *13*, 447–457. [CrossRef]
14. Li, L.F.; Shi, Y.L.; Cheng, S. Exploration of Long Short-Term Memory Neural Network in Intermediate Earthquake Forecast: A Case Study in Sichuan–Yunnan Region. *Chin. J. Geophys.* **2022**, *65*, 12–25.
15. Asencio-Cortés, G.; Morales-Esteban, A.; Shang, X.; Martínez-Álvarez, F. Earthquake Prediction in California Using Regression Algorithms and Cloud-Based Big Data Infrastructure. *Comput. Geosci.* **2018**, *115*, 198–210. [CrossRef]
16. Mallouhy, R.; Abou Jaoude, C.; Guyeux, C.; Makhoul, A. Major Earthquake Event Prediction Using Various Machine Learning Algorithms. In Proceedings of the International Conference on Information and Communication Technologies for Disaster Management, Paris, France, 18–20 December 2019; pp. 1–6.
17. Sadhukhan, B.; Chakraborty, S.; Mukherjee, S. Predicting the Magnitude of an Impending Earthquake Using Deep Learning Techniques. *Earth Sci. Inform.* **2025**, *16*, 803–823. [CrossRef]
18. Florido, E.; Asencio-Cortés, G.; Aznarte, J.L.; Rubio-Escudero, C.; Martínez-Álvarez, F. A Novel Tree-Based Algorithm to Discover Seismic Patterns in Earthquake Catalogs. *Comput. Geosci.* **2018**, *115*, 96–104. [CrossRef]
19. Wang, X.; Zhong, Z.Y.; Yao, Y.C.; Li, Z.; Zhou, S.H.; Jiang, C.C.; Jia, K. Small Earthquakes Can Help Predict Large Earthquakes: A Machine Learning Perspective. *Appl. Sci.* **2023**, *13*, 6424. [CrossRef]
20. Jiang, H.K.; Li, Y.L.; Qu, Y.J.; Zhao, X.W.; Chen, P.; Zhang, M. Spatial Distribution Characteristics of Sequence Types for Moderate–Strong Earthquakes in the Chinese Mainland. *Acta Seismol. Sin.* **2006**, *28*, 389–398.
21. Wiemer, S.; Wyss, M. Minimum Magnitude of Completeness in Earthquake Catalogs: Examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Am.* **2000**, *90*, 859–869. [CrossRef]
22. Wiemer, S. A Software Package to Analyze Seismicity: ZMAP. *Seismol. Res. Lett.* **2001**, *72*, 373–382. [CrossRef]
23. Su, Y.J.; Li, Y.L.; Li, Z.H.; Yi, G.X.; Liu, L.F. Analysis of the Minimum Magnitude of Completeness for Regional Earthquake Catalogs in the Sichuan–Yunnan Region. *J. Seismol. Res.* **2003**, *26* (Suppl. S1), 7–15.
24. Mignan, A.; Woessner, J. *Estimating the Magnitude of Completeness for Earthquake Catalogs*; Community Online Resource for Statistical Seismicity Analysis (CORSSA). 2012. Available online: https://www.researchgate.net/publication/285715153_Estimating_the_magnitude_of_completeness_for_earthquake_catalogs (accessed on 20 October 2025).

25. Martínez-Álvarez, F.; Reyes, J.; Morales-Esteban, A.; Rubio-Escudero, C. Determining the Best Set of Seismicity Indicators to Predict Earthquakes: Two Case Studies—Chile and the Iberian Peninsula. *Knowl.-Based Syst.* **2013**, *50*, 198–210. [CrossRef]
26. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
28. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
29. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
30. Bickel, P.J.; Doksum, K.A. *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd ed.; Pearson Education: London, UK, 2015; Volume 1, p. 20.
31. Wright, S. Correlation and Causation. *J. Agric. Res.* **1921**, *20*, 557–585.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.