Machine learning methods for predicting earthquake frequency in the geographical region surrounding northern Morocco

***My Ahmed* EL HAFIDI[1], *Abderrahim* BOULANOUAR[1], *Ahmad* EL ALLAOUI[2] and *Abdelaali* RAHMOUNI[3]**

[1]Laboratory of Applied Sciences (LSA), National School of Applied Sciences, Abdelmalek ESSAADI University, P.O Box 03, Ajdir Al-Houceïma, Morocco.
[2]Department of Computer Science, Faculty of Sciences and Techniques, Errachidia, Morocco.
[3]Laboratory of Solid-State Physics, Department of Physics, Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco.

Abstract. In order to reduce risks, we utilize machine learning/deep learning models to forecast the frequency of earthquakes in a specific geographic region, such as northern Morocco. Several types of studies have allowed the obtaining of acceptable results by integrating the ARIMA model with machine learning/deep learning models, such as LSTM, XGBoost, SVR and RF. Given that Morocco is situated in a moderately active seismic zone, our study examines how well machine learning and deep learning models predict the frequency of earthquakes in the area surrounding northern Morocco. Knowing that Morocco is located in a moderately active seismic zone, our study compares the effectiveness of machine learning and deep learning models in predicting the frequency of earthquakes in the region around northern Morocco. We employ a collection of hybrid models that integrate the ARIMA models with various machine learning/deep learning models to operate as a guiding core for future development challenges, particularly because it has allowed us to perform a large degree of prediction. We obtained very significant results regarding hybrid ARIMA models and machine learning models (RF, SVR, XGB), whilst the ARIMA model failed when used on its own or even when hybridised with the deep learning model LSTM.

Keywords: ARIMA, Random forest, LSTM, Support vector regression, XGBoost.

1 Introduction

Seismic activity has attracted widespread attention due to the constant danger it poses, particularly in urban areas that are frequently affected by it. Northern Morocco is a prime example of such an area, as it lies at the point where the African and Eurasian tectonic plates meet [1], thereby increasing the potential risks. For this reason, many scientists are focusing on developing earthquake prediction models to improve upon previous efforts in this field, despite the challenges and limitations they face, even with the current advancements in artificial intelligence and other technologies. For example, the scarcity of seismic events and the limited availability of data [2] pose a challenge to improving prediction models [3], necessitating innovation and the development of new approaches in the field of seismology.

The fields of scientific research into seismic activity remain diverse, and many researchers are focusing on utilising the capabilities of machine learning to identify and classify seismic events [4-6]. The use of machine learning is becoming increasingly widespread. By identifying patterns and changes in data over time, time series analysis enables the prediction of future trends and the detection of recurring changes. Time series analysis has been used in seismology to study the intervals between earthquakes [4, 7, 8]. The concept of 'natural time' has become a standard technique for analysing spiky time series, particularly seismic series. Time series models commonly used today include AR, MA and ARIMA models, and other models exist [9, 10]. Forecasting using seismic data is a highly significant area of research in earthquake engineering. In this context, many early efforts focused on applying this research approach using methods such as autoregressive (AR) models, moving average (MA) models, autoregressive integrated moving average (ARIMA) models, and others [11-13]. In 2020, Ma et al. employed a novel approach combining statistical models with machine learning models to predict water levels, and this hybrid model type has been utilised in other studies. Our research utilises various hybrid models (ARIMA-SVR, ARIMA-XGB, ARIMA-RF, ARIMA-LSTM, STACKING and ENSEMBLE) to predict earthquake frequency based on seismic data from a geographical area comprising northern Morocco, obtained from the Spanish National Institute of Geology and Mining (ING). To achieve this, three essential phases will be taken: presenting the methodology and seismic data, then presenting the hybrid models, and finally analyzing and evaluating the results to draw key conclusions. This study examines earthquake frequency forecasting in northern Morocco using hybrid models comprising ARIMA models and machine learning models; it is considered the first study of its kind to be conducted in Morocco and has yielded good results with these models.

2 Methodology

2-1 Visual representation of data

2.1.1 Data sources and collection

The Spanish National Geographic Institute (IGN) [14] provides seismic data recognised by researchers and scientific bodies, particularly in the region under study due to the proximity of its stations. This work focuses on analysing time series of earthquake frequency and date [15]. The seismic data used covers the period from 21 February 1960 to 28 December 2025 and includes 3,421 seismic events of magnitude 4 or greater in the following study area: "North-western Mediterranean and north-eastern Atlantic" (latitude: 30.15°N – 44.96°N, longitude: 19.97°W – 6.00°E), which includes northern Morocco. We will use the abbreviation NORDMA to refer to this region in this study. It should be noted that monthly frequencies will be our main focus in this research. Figure 1 illustrates the distribution of seismic data and their intensity for the study area over the period from February 1960 to December 2025.

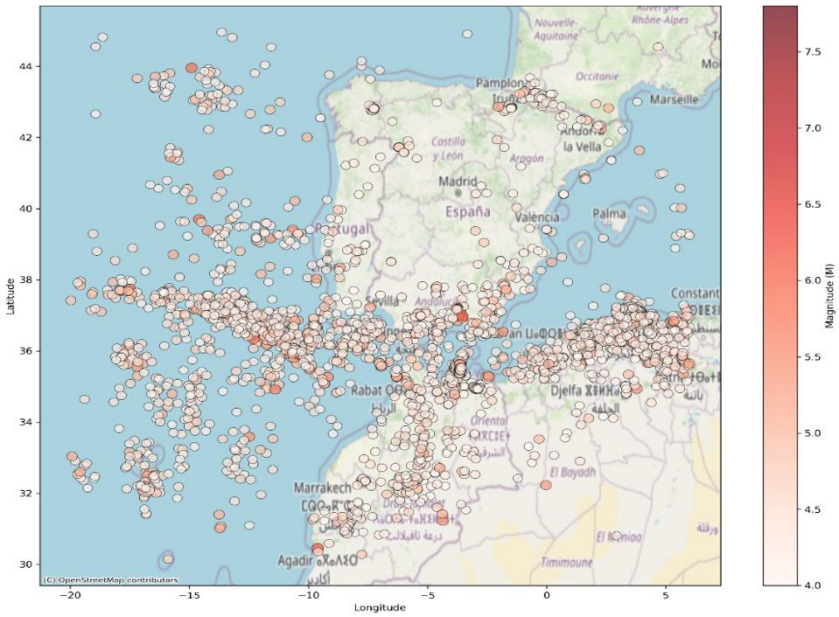

Figure 1. Seismic activity in the NORDMA region between February 1960 and December 2025.

2.1.2 Descriptive statistics and time trends

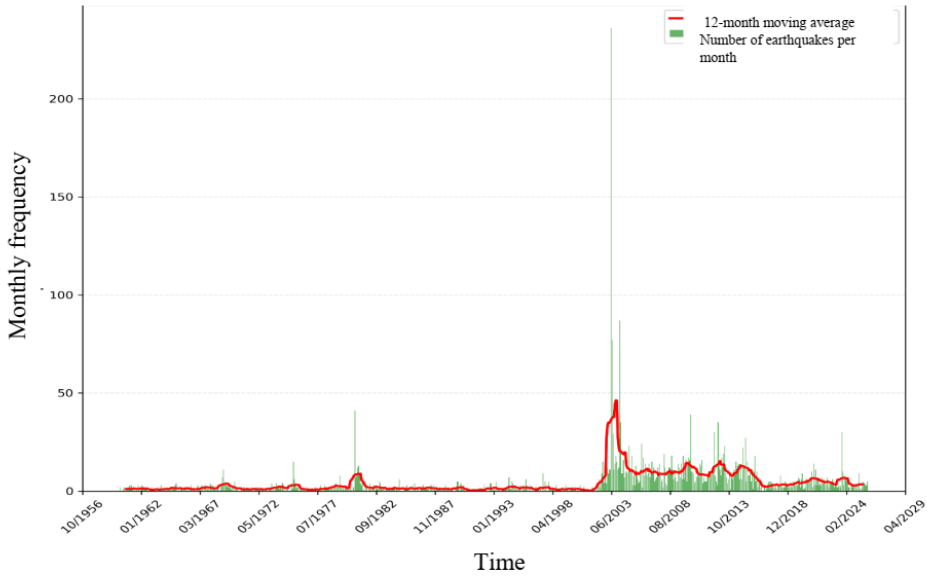

Figure 2. Shows the distribution of earthquakes in the NORDMA region over time on a monthly basis, as well as a 12-month moving average.

Monthly data reveal fluctuations in seismic activity over different time periods, whilst the moving average curve highlights cyclical trends and fluctuations in seismic activity,

smoothing out the short-term fluctuations present in the raw monthly data. Thus, many peaks in the moving average correspond to periods of increased seismic activity, which justifies the usefulness of this approach for identifying trends over time [15].

2.1.3 Checking the validity of data using the Gutenberg-Richter law

The Gutenberg–Richter law [16] was used to assess the quality and completeness of the seismic data used. Figure 3 shows a linear relationship between the logarithm of seismic frequency $log_{10}(N)$ and magnitude M, demonstrating that the data used comply with Gutenberg-Richter's law. This confirms the quality of the data, particularly as the coefficient of determination reached a high value of R^2=0.9498. Equation 1 represents the equation shown in Figure 3:

$$log_{10}(N) = 7.2 - 0.97M \quad (1)$$

N represents the cumulative number of earthquakes with a magnitude greater than or equal to M; the negative slope highlights the inverse relationship between earthquake magnitude and frequency.

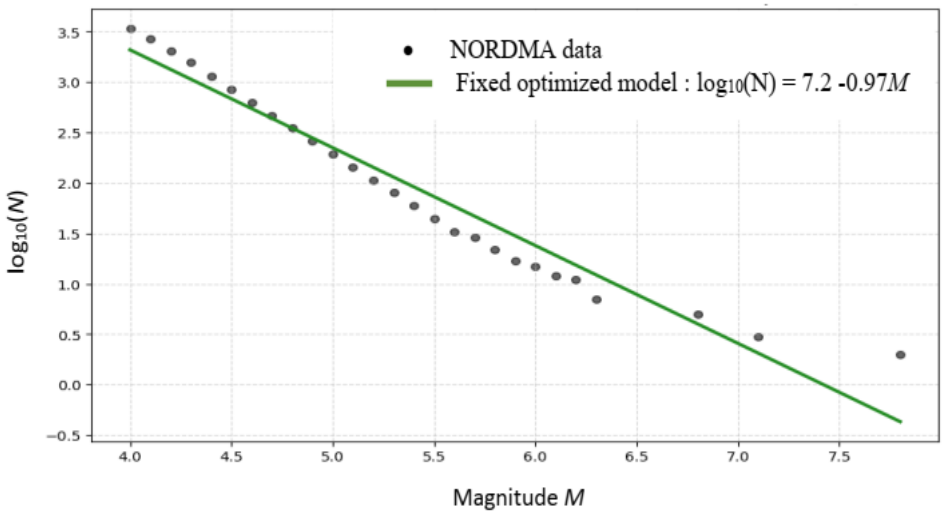

Figure 3. Validation of the Gutenberg-Richter law.

These results are similar to those obtained by Wenwen 2026 in his study of Indonesia (R^2 = 0.95; a = 7.75; b = -0.91).

2.2 Multi-stage prediction

2.2.1 Data segmentation for prediction

The data has been split into two sets: one covering the period from February 1960 to December 2024, which we will use to train the models, and another covering the period from January 2025 to December 2025, which will be used to evaluate the models.

This approach to data division enables the models to be trained over a long period to strengthen their learning, as well as to make predictions on recent data for evaluation [15].

2.2.2 The Support Vector Regression (SVR)

The main difference between Support Vector Regression (SVR) and most regression algorithms lies in the fact that it focuses on minimising the error [17], over a specific time period, rather than minimising the sum of all error squares. It should be noted that this model falls within the category of machine learning models [18]; it adopts the principle of Support Vector Machines (SVMs) with regard to regression, aiming to predict continuous target variables by constructing a hyperplane that maximises the margin around a subset of data designated for training, known as support vectors [15]. To address the limitations of non-linear regression, it also relies on kernel functions such as the Gaussian radial basis functions to capture complex patterns in the data [15].
The challenge of this algorithm lies in finding the appropriate hyperplane capable of minimising the error between predicted and true values whilst simultaneously maximising the margin [19].
For further details on the mathematical calculations underlying this model, please refer to our reference number 15 in this work.

2.2.3 The Random Forests (RF) model

Random Forests are part of the developments in classification and regression tree models, whose outputs are known as "decision trees", RF relies on constructing a large number of decision trees, and their results are aggregated into a single output using voting based on classification problems or averaging in the case of regression problems [20]. It is considered a type of ensemble machine learning model as it aggregates the final decisions of several trees [15].
For further details on the mathematical functions underlying this model, please refer to our reference number 15 in this work.

2.2.4 The XGBoost model

XGBoost is a machine learning model based on decision trees; it expands and enhances these trees through an iterative process, enabling it to achieve superior performance on a wide range of datasets [21]. The objective of the training phase in this algorithm is to minimise a specific loss function, measured by the difference between predicted and actual values [15]. Reference 15 provides further details regarding the loss function utilised by this algorithm to achieve the desired accuracy.

2.2.5 The Long Short-Term Memory (LSTM) model

The Long Short-Term Memory (LSTM) model is a deep learning model capable of handling error fading and overcoming time delays; it falls within the category of Recurrent Neural Networks (RNNs) specialised in learning based on long-term sequential data [22]. This model can resolve the issues that may arise in standard RNNs [23].
Please refer to reference 15 for further details and the mathematical relationships underlying this model.

2.2.6 The Autoregressive Integrated Moving Average (ARIMA) model

The Autoregressive Integrated Moving Average (ARIMA) model consists of three components [24], namely autoregression (AR), integration (I) and moving average (MA), and is used in many fields for forecasting purposes.
For further technical details and information on the mathematical equations underlying this model, please refer to references 15 and 25 of this study.

2.2.7 The ARIMA-machine learning hybrid model

2.2.7.1 The general principle of the hybrid approach

The approach adopted in this hybrid model involves first using ARIMA to generate forecasts, after which the differences between the forecasts and the actual values (known as residuals [15]) are calculated, Through machine learning models, the residuals are processed via a regression relationship between the residuals and the actual values, thereby obtaining the final values of the hybrid model [15, 26]. Figure 4 illustrates the flowchart of the approach used in this hybrid model.

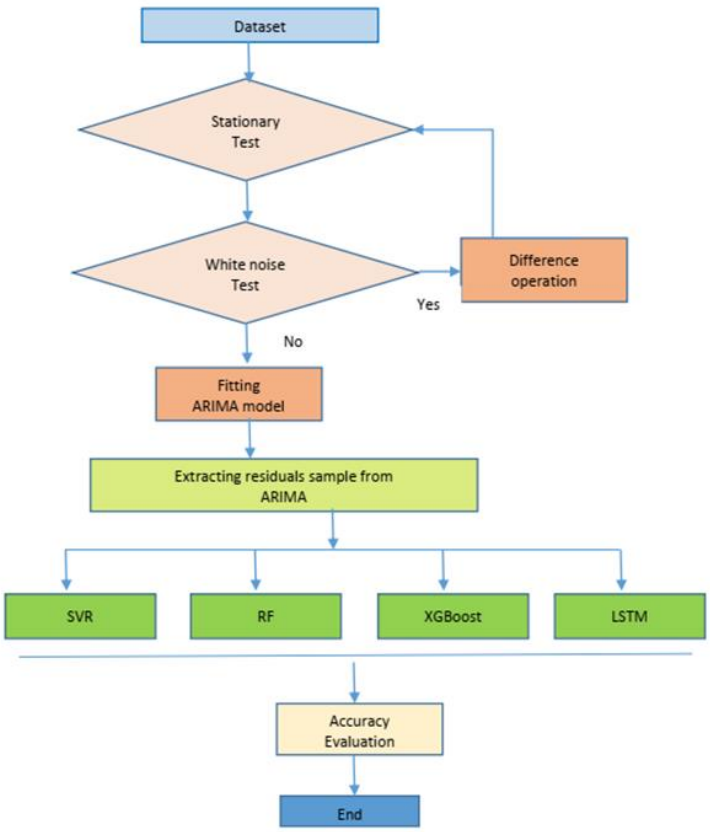

Figure 4. The diagram illustrating the structure of the approach adopted in the hybrid model.

2.2.7.2 ARIMA Model – Selection and Validation

The ARIMA model consists of three sub-components: autoregression AR(p), integration I(d) and moving average MA(q) [24, 25]. Stationarity was tested using the ADF test (statistic = -4.3478, p = 0.0004). We estimated a ARIMA(1, 1, 1) model using auto_arima without a constant term (AIC = 5483.16).

Table 1. Diagnostic criteria for the selected ARIMA(1,1,1) model.

Model	AIC	BIC	Ljung-Box p-val	Std Résidus
ARIMA(1,1,1)	5483,16	5496,99	0,0000	9,687

2.2.7.3 Support Vector Regression (ARIMA+SVR)

SVR minimises errors within a specific range using Gaussian RBF kernel functions [18, 19]. The model uses lags 1, 2, 3 and Fourier terms (periods of 5, 9 and 10 months). WF-CV over 30 combinations: C=100, ε=0.0005 (CV_R^2=0.6104).

2.2.7.4 Random Forest (ARIMA+RF)

A random forest combines several decision trees trained on random subsets of features [20]. Selection by importance (top 60%: 44/73 features). WF-CV on 27 combinations: n_estimators=800, max_features=0.5 (CV_R^2=0.7197, OOB=0.4251).

2.2.7.5 XGBoost (ARIMA+XGB)

The XGBoost algorithm relies on building sequential decision trees, with each tree correcting the errors of its predecessor using Omega(f) regularisation [21]. WF-CV was applied to 108 configurations: n_estimators=800, Ir=0.01, max_depth=4, subsample=0.8 (CV__R2=0.5716).

2.2.7.6 LSTM with Bayesian Optimisation (ARIMA+LSTM-BO)

The LSTM network manages long-term dependencies via gate mechanisms, resolving the vanishing gradient problem [22, 23]. Bayesian optimisation [27, 28] is used to select the hyperparameters. 7 features, look_back=12. Convergence at k=6, units_1=48, units_2=48, dropout=0.20, lr=0.0005. A Gaussian anomaly detection mechanism identifies two anomalous months in 2025 (July and August).

2.2.7.7 Ensemble and Stacking Ridge

Two combination strategies are explored. The R^2-weighted ensemble assigns proportional weights: SVR=33.4%, RF=33.4%, XGB=33.1%, LSTM=0%. Ridge Stacking (meta-learner) linearly optimises the predictions of the base models, an approach whose relevance has been demonstrated in the literature [26].

3 Presentation of results and evaluation

3.1 Performance comparison table

Table 2. A comparison of the performance of all models on the 12-month test set from 2025.

Model	R^2	MAE	RMSE	MAPE%	SMAPE%	MBE	MaxErr
STACKING	0,9997	0,035	0,04	1,31%	1,31%	0,000	0,07
ARIMA+SVR	0,9993	0,044	0,056	2,55%	2,45%	0,014	0,11
ARIMA+RF	0,9992	0,053	0,062	2,40%	2,44%	-0,014	0,12
ENSEMBLE	0,9983	0,078	0,089	3,80%	3,67%	0,038	0,18
ARIMA+XGB	0,9891	0,195	0,228	9,54%	8,68%	0,116	0,44
ARIMA seul	-0,1722	1,874	2,36	111,91%	54,43%	0,925	4,57
ARIMA+LSTM-BO	-0.3719	2,141	2.553	82,03%	71.40%	-1.119	5.94

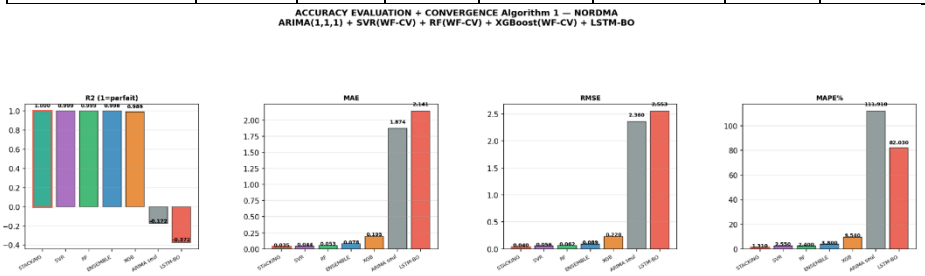

Figure 5. Visual comparison of the three key metrics (R^2, MAPE%, RMSE, MAPE %) for the seven models evaluated. The STACKING Ridge model outperforms the others on all criteria

3.2 Monthly predictions for 2025

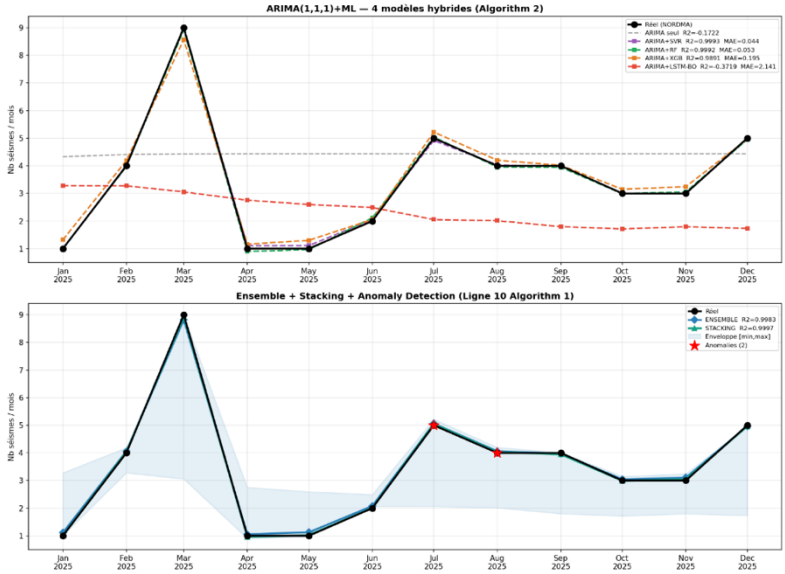

Figure 6. Monthly forecasts versus actual values (2025). Top: the models (ARIMA-XGB, ARIMA-SVR, ARIMA-RF, ARIMA alone, LSTM-BO). Bottom: the models (STACKING, ENSEMBLE).

4. Discussion

The results confirm the significant benefits of The ARIMA-machine learning hybrid model approach for multi-step forecasting of seismic frequency. The decomposition into a linear component (ARIMA) and a residual non-linear component (machine learning) enables remarkable levels of accuracy to be achieved, which are unattainable with ARIMA alone.

The superiority of stacking over individual models demonstrates the value of meta-learning: by combining the complementary prédictions of the three high-performing hybrid models, the Ridge meta-learner exploits their strengths and mitigates their individual weaknesses, resulting in a near-perfect forecast (R^2=0.9997).

The poor performance of the LSTM with Bayesian optimisation warrants attention. Several factors account for this: the inherent stochasticity of seismic processes, the relatively short duration of the time series (755 months), the high variability of the residuals, and the absence of regular patterns in the non-linear component. These conditions do not favour the generalisation of deep neural networks [2, 22]. Mignan and Broccardo (2020) highlighted that the limited availability of data constitutes a major obstacle to deep learning in seismology. This result is consistent with the observations of Wenwen, 2026 on Indonesian data, the same finding was recorded in our study. What is even more striking here is that ARIMA alone achieved a higher R^2 than ARIMA+LSTM, we favour the argument that attributes the failure of this model to a lack of seismic data.

The anomaly detection feature integrated into the LSTM-BO pipeline identified two anomalous months (July and August 2025: 5 and 4 earthquakes respectively). This Gaussian mechanism is a valuable complementary tool for seismic monitoring, regardless of the model's predictive performance.

Validation using the Gutenberg-Richter law (R^2 = 0.9498) [16] confirms the quality and representativeness of the NORDMA catalogue, an essential prerequisite for any statistical modelling of seismicity [4, 7].

The benefit of limiting the use of seismic data to the last 12 months for the region is also evident in the fact that no further data is lost during the training phase, given the problem of data scarcity which typically poses an obstacle to machine learning models; As for the evaluation, this was carried out using recent data that had not been included in the models' training set. It is worth noting that the models are only capable of predicting future events; it may not be possible to evaluate them due to the absence of actual values.

The effectiveness of this hybrid model and its impressive results are evident in the fact that it employed a well-structured, multi-stage hybrid approach, whereby forecasting was first carried out using an ARIMA model, whilst machine learning models were used to handle the residuals.

5. Conclusion

This study presented a comprehensive hybrid pipeline for multi-step forecasting of seismic frequency in NORDMA, building on recent work on ARIMA + machine learning approaches).

The main findings are: (1) the ARIMA+SVR and ARIMA+RF models significantly outperform ARIMA alone with $R^2 > 0.999$; (2) the Stacking meta-model is the best strategy with R^2=0.9997 and MAPE=1.26%; (3) LSTM-BayesOpt confirms the conclusions of Wenwen, 2026 regarding the limitations of deep learning for small-scale seismic time series.

6. References

1. CHERKAOUI et ASEBRIY, 2003. Le risque sismique dans le Nord du Maroc. Trav. Inst. Sci. Rabat, sér. Géol. & Géogr. phys., n° 21, 2003, p.225-232
2. Mignan A, Broccardo M (2020) Neural network applications in earthquake prediction (1994–2019): Metaanalytic and statistical insights on their limitations. Seismol Res Lett 91(4):2330–2342
3. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. Rev Modern Physics 66(2):657
4. Gulia L, Wiemer S (2019) Real-time discrimination of earthquake foreshocks and aftershocks. Nature 574(7777):193–199
5. Linville L, Pankow K, Draelos T (2019) Deep learning models augment analyst decisions for event discrimination. Geophys Res Lett 46(7):3643–3651
6. Meier M-A, Ross ZE, Ramachandran A, Balakrishna A, Nair S, Kundzicz P, Li Z, Andrews J, Hauksson E, Yue Y (2019) Reliable real-time seismic signal/noise discrimination with machine learning. J Geophys Res: Solid Earth 124(1):788–800
7. Li S, Xu W, Li Z (2022) Review of the SBAS InSAR Time-series algorithms, applications, and challenges. Geodesy Geodyn 13(2):114–126
8. Mousavi SM, Ellsworth WL, Zhu W, Chuang LY, Beroza GC (2020) Earthquake transformer–an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Commun 11(1):3952
9. Ambrosino F, Thinová L, Briestenský M, Šebela S, Sabbarese C (2020) Detecting time series anomalies using hybrid methods applied to Radon signals recorded in caves for possible correlation with earthquakes. Acta Geodaetica et Geophysica 55:405–420
10. Mandrikova O, Fetisova N, Polozov Y (2021) Hybrid model for time series of complex structure with ARIMA components. Mathematics 9(10):1122
11. Gao Y, Mosalam KM, Chen Y, Wang W, Chen Y (2021) Auto-regressive integrated moving-average machine
12. Amei A, Fu W, Ho C-H (2012) Time series analysis for predicting the occurrences of large scale earthquakes. Int J App Sci Tech 2(7):75
13. Musarat MA, Alaloul WS, Rabbani MBA, Ali M, Altaf M, Fediuk R, Vatin N, Klyuev S, Bukhari H, Sadiq A (2021) Kabul river flow prediction using automated ARIMA forecasting: A machine learning approach. Sustainability 13(19):10720
14. https://www.ign.es/web/en/ign/portal/sis-catalogo-terremotos
15. Wenwen Hou, 2026 Statistical and machine learning methods for multi-step earthquake frequency forecasting in indonesian regions https://doi.org/10.1007/s11069-025-07744-9
16. Kijko A, Smit A (2017) Estimation of the frequency-magnitude Gutenberg-Richter b-value without making assumptions on levels of completeness. Seismol Res Lett 88(2A):311–318
17. Al Banna MH, Taher KA, Kaiser MS, Mahmud M, Rahman MS, Hosen AS, Cho GH (2020) Application of artificial intelligence in predicting earthquakes: state-of-the-art and future challenges. IEEE Access 8:192880–192923
18. Aljarah Ibrahim and al (2018) Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm. Cognitive Comput 10(3):478–495
19. Ghosh R, Sinha N, Biswas SK (2019) Automated eye blink artefact removal from EEG using support vector machine and autoencoder. IET Signal Process 13(2):141–148

20. Rigatti SJ (2017) Random forest. Journal of Insurance Medicine 47(1):31–39. ISBN: 0743-6661 Publisher:American Academy of Insurance Medicine 1700 Magnavox Way. Fort Wayne, IN, p 46804
21. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
22. Staudemeyer RC, Morris ER (2019) Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. arXiv arXiv:1909.09586 American Academy of Insurance Medicine 1700 Magnavox Way. Fort Wayne, IN, p 46804
23. Sahoo BB, Jha R, Singh A, Kumar D (2019) Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophysica 67(5):1471–1481
24. Zhang M (2018) Time series: Autoregressive models ar, ma, arma, arima. University of Pittsburgh
25. Fattah J, Ezzine L, Aman Z, El Moussami H, Lachhab A (2018) Forecasting of demand using ARIMA model. Int J Engi Bus Manage 10:1847979018808673
26. Júnior DSdOS, Oliveira JF, Mattos Neto PS (2019) An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowl-Based Syst 175:72–86 learning for damage identification of steel frames. Appl Sci 11(13):6084
27. Wilson J, Hutter F, Deisenroth M (2018) Maximizing acquisition functions for Bayesian optimization. Advances in neural information processing systems 31
28. Meng H, Geng M, Han T (2023) Long short-term memory network with Bayesian optimization for health prognostics of lithium-ion batteries based on partial incremental capacity analysis. Reliab Engin & Syst Safety 236:109288