Natural Hazards (2024) 120:2283–2309
https://doi.org/10.1007/s11069-023-06221-5

ORIGINAL PAPER

Earthquake prediction from seismic indicators using tree-based ensemble learning

Yang Zhao[1] · **Denise Gorse**[1]

Received: 28 April 2023 / Accepted: 14 September 2023 / Published online: 27 January 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract
Earthquake prediction is a challenging research area, but the use of a variety of machine learning models, together with a range of seismic indicators as inputs, has over the last decade led to encouraging progress, though the variety of seismic indicator features within any given study has been generally quite small. Recently, however, a multistage, hybrid learning model has used a total of 60 seismic indicators, applying this to data from three well-studied regions, aiming to predict earthquakes of magnitude 5.0 or above, up to 15 days before the event. In order to determine whether the encouraging results of this prior work were due to its learning model or to its expanded feature set we apply a range of tree-based ensemble classifiers to the same three datasets, showing that all these classifiers outperform the original, more complex model (with CatBoost as the best-performing), and hence that the value of this prior approach likely lay mostly in its range of presented features. We then use feature rankings from Boruta-Shap to discover the most valuable of these 60 features for each of the three regions, and challenge our optimized models to predict earthquakes of larger magnitudes, demonstrating their resilience to imbalanced data. Notably, we also address the prevalent issue of inappropriate test data selection and data leakage in earthquake prediction studies, demonstrating our models can continue to deliver effective predictions when the possibility of data leakage is strictly controlled.

Keywords Earthquake prediction · Seismic indicators · Ensemble learning · Feature selection

1 Introduction

The desirability of earthquake prediction is obvious, an effective prediction system having a huge potential to both save human life and diminish financial losses. In last decade, a large amount of effort has been devoted to building the seismological infrastructure,

✉ Denise Gorse
d.gorse@cs.ucl.ac.uk

Yang Zhao
yang.zhao1215@gmail.com

[1] Computer Science Department, University College London, Gower Street, London WC1E 6BT, UK

due to rapid economic development and a related increase in the hazards posed by earthquakes; for example, the economic costs of the earthquake and resulting tsunami that struck Japan in 2011 were in excess of USD 300 billion (Noy et al. 2022). Increasing numbers of seismic monitoring stations have been established in countries such as Japan, and there has been a dramatic increase in the availability of seismic data (Bergen et al. 2019; Arrowsmith et al. 2022). Yet the goal of practical earthquake prediction, the ability to deliver warnings of heightened seismic risk that are timely, geographically precise, specific as to magnitude, and reliable, remains elusive; for example, in Ikram and Qamar (2015), a rule-based model was used to make accurate predictions of seismic activity for the next 12 h, but the predicted events could be located only within a single quadrant of the Earth's surface, hence far from being geographically precise.

As noted in Allen (1976), a prediction should ideally specify exactly when and where the event will occur, with what magnitude, and with what probability. It should also ideally be made by a model which generates an explanation. Machine learning (ML) has in recent years made substantial inroads into the Earth sciences (Bergen et al. 2019; Kong et al. 2019), including earthquake prediction (Galkina and Grafeeva 2019; Beroza et al. 2021). Deep neural networks (DNNs), while manifestly successful in geophysical applications (see, for example, the reviews of Yu and Ma (2021) Mousavi and Beroza (2022)), have been subject to some criticism on grounds of interpretability (Mignan and Broccardo 2020). There are additional problems in applying DNNs in most earthquake prediction scenarios due to the relatively small size of the available datasets (hundreds to thousands of examples), which if used to train deep networks can easily lead to overfitting.

To maximize interpretability (Doshi-Velez and Kim 2017), and make the best use of relatively scarce data, one might choose to adopt simpler models, such as logistic regression, with appropriately selected input features; however, while logistic regression has been used in the related task of earthquake detection, for example in Miranda et al. (2019) Waheed et al. (2020) Gorse and Goel (2022), for the more challenging task of prediction it is likely that more complex models will be required.

Models based on ensembles of decision trees, such as random forest (Breiman 2001) and XGBoost (Chen and Guestrin 2016), represent an intermediate level of complexity and transparency; while these are powerful models, they can also deliver some insight into the 'why' of their predictions. These models have seen wide use in the prediction of earthquakes in both natural (e.g., Asim et al. (2016), Asim et al. (2017a), Asim et al. (2017b), Asencio-Cortés et al. (2018), Asim et al. (2020), Yousefzadeh et al. (2021), Novick and Last (2023)) and laboratory settings (all three top-ranked models in the 2019 Kaggle 'labquake' prediction competition (Johnson et al. 2021) were of this type). Crucially, while being quite powerful models, tree-based ensembles can be used more easily with smaller datasets than DNNs; in an earthquake magnitude prediction study (Asencio-Cortés et al. 2018) based on datasets with between 717 and 5575 examples, random forest was the best-performing model, and a DNN the worst.

In order to maximize the value of a tree-based ensemble model, however, it is important to present it with appropriate inputs; these models, especially with the current relative scarcity of data, cannot be expected to learn effectively from raw seismic data. While it is possible to use generic time series analysis tools such as HCTSA (Fulcher and Jones 2014) to generate features, as was done for example in Sadhukhan et al. (2022), seismicity indicators (a brief introduction to which is given in Sect. 2.1) have been popular as inputs into ML earthquake prediction models, having the advantage of long use in seismology and an underpinning geophysical justification.

A recent trend in the use of seismic indicators in ML studies has been to gather together as many quantities as might be usefully indicative of subsurface conditions, and input all of these into a model, ideally with a feature selection tool to discover which inputs are most valuable, as in, for example (Asim et al. 2018a), and most recently in Novick and Last (2023). The work of this paper uses one such dataset, with 60 seismic indicator features, made available online (Asim et al. 2021) by the authors of Asim et al. (2018a). This study used these features to build hybrid neural network earthquake prediction models for each of three regions (Hindukush, Chile, and Southern California). It introduced the use, as promising input features, of earthquake recurrence times (Wiemer and Wyss 1997), and it was of interest to us to investigate the value of these, and of this 60-feature input set more broadly, when used with a tree-based ensemble model for these same three geographical regions.

Specifically, our paper will:

- Apply tree-based ensemble classifiers to the three datasets used in a promising prior work based on a hybrid, multi-stage learning model (Asim et al. 2018a), showing these outperform the original, more complex model, and hence that the value of this prior approach likely lay mostly in its expanded range of 60 input features.
- Use Boruta-Shap feature selection to improve the performance of our best-performing (CatBoost (Prokhorenkova et al. 2018)) models for each of the three datasets, and to discover the most valuable of the 60 features for earthquake prediction in each of the three considered geographical regions.
- Challenge our models to predict earthquakes of larger magnitudes, thereby demonstrating their resilience when faced with highly imbalanced data.
- Address the prevalent issue of inappropriate test data selection and data leakage in earthquake prediction studies, demonstrating our models can continue to deliver effective predictions when the possibility of data leakage is strictly controlled.

The remainder of this paper will be organized as follows. Section 2 will first discuss the range of available approaches to earthquake prediction using ML methods, before motivating the use of seismic indicators as inputs, and reviewing a set of works which use these input features. Section 3 will address the data source and overall methodology of the current work. Section 4 will then present the results of the study, beginning with a comparison of bagging and boosting models (random forest and XGBoost) with the hybrid ML model of Asim et al. (2018a), progressing to the introduction of further tree-based ensemble classifiers, demonstrating the effectiveness of Boruta-Shap feature selection in further improving the results of our best-performing models, with an investigation of the most important discovered features. Section 5 contains a discussion and suggestions for future work, and Sect. 6 summarizes our conclusions.

2 Background and related work

It has in the past been debated whether earthquakes are predictable at all; for example, Geller, in 1997, stated that effective earthquake prediction was essentially impossible (Geller 1997). Opinion has since become more optimistic, not only because of the availability of powerful machine learning tools, but also because of the large amounts of data now available (Bergen et al. 2019; Arrowsmith et al. 2022). However, the amount of data pertaining

to larger seismic events is still arguably insufficient to support deep learning models. Most of the work with raw data has been lab-based, attempting to make predictions for analog earthquake models, as for example in Pu et al. (2021) Blank and Morgan (2021) Laurenti et al. (2022), for which, by construction, as much training data as wished can be generated. Most earthquake prediction work using natural data has in contrast been based on seismic indicator features, which have long been used as proxies for information that cannot be obtained about conditions in the subsurface that would tend to trigger earthquakes, and can give some indication of the build-up of stress at a fault. Use of these calculated measures, critically in the context of the above discussion, also dramatically reduces the complexity of the models required, thereby allowing feasible learning from datasets with order of only hundreds to thousands of training examples.

The following section will give a brief survey of the seismicity indicators that have been most commonly used for building machine learning (ML) prediction models, and will be followed by a section that reviews past ML prediction work which has used such indicators, setting the scene for the description of our methodology and the experimental results that will follow.

2.1 A brief survey of seismicity indicators

Among seismic indicators, the slope parameter (b-value) of the Gutenberg-Richter law has been especially heavily investigated. The Gutenberg-Richter law (Gutenberg and Richter 1944)

$$\log N = a - bN \tag{1}$$

describes the frequency-magnitude distribution of earthquakes, where N is the total number of earthquakes with magnitudes $\geq M$, the intercept a measures the level of seismic activity in the region, and b describes the relative size distribution, such that a smaller than average b-value means proportionally more larger earthquakes have been experienced in a region, and a larger than average b-value the converse. There have been many studies of spatial and temporal variations in b-value (referenced for example in El-Isa and Eaton (2014)), establishing a potential predictive value for this parameter supported also by lab-based studies (Rivière et al. 2018).

The Gutenberg-Richter a-value has also been considered as an indicator of seismicity, as has the standard deviation of the b-value (Shi and Bolt 1982). Other studied parameters derived from the a- and b-values include deviation from the GR law (η value) (Panakkat and Adeli 2007), magnitude deficit (M_{def}) (Panakkat and Adeli 2007) (the difference between the maximum observed earthquake magnitude during a certain period and the maximum expected magnitude derived from the GR law), and probabilistic recurrence time (T_{rec}) (Wiemer and Wyss 1997) (the time between two earthquakes of magnitude greater than or equal to a specified magnitude).

A further set of possibilities for input features to ML prediction models derive from seismic indicators unconnected to the GR law, yet similarly aiming to detect a build-up of stress at a fault that might indicate an imminent seismic event. These include change in the rate of seismic energy release ($dE^{1/2}$) (Jaumé and Sykes 1999), and measures of changes in the rate of seismicity (Habermann's z-value (Wyss and Habermann 1988) and the β-measure of Matthews and Reasenberg (1988)). Habermann's z measures the difference between two means, the more recent for the possibly-anomalous period that might precede a seismic event, and the earlier, usually over a longer period, to establish a background level of

Table 1 Ma / Panakkat & Adeli (MPA) seismicity indicators. The first six were proposed by Ma et al. (1999); the remaining two were added by Panakkat and Adeli (2007)

Name	Description
b	b-value in Gutenberg-Richter (GR) law, calculated over last n events
η	Deviation from GR law during last n events
M_{def}	Magnitude deficit (difference between max observed earthquake magnitude during last n events, and max expected from GR law)
T	Time during which last n seismic events occurred
M_{mean}	Mean magnitude of last n seismic events
$dE^{1/2}$	Seismic energy release
μ	Mean time between last n characteristic events
c	Coefficient of variation of the mean time between last n characteristic events

seismicity. z will be positive when the seismicity rate decreases (seismic quiescence), negative in the converse case. The study of Oynakov and Botev (2021), for example, found that the epicenters of larger seismic events in the southern Balkans tended to fall in areas of relatively high z and relatively low GR b-value. The β-measure of Matthews and Reasenberg uses a different method to capture similar information; in this case a positive value indicates an increase in seismicity, and a negative value the converse. Of these two measures, the z-value is more widely known and used, but it is of potential value to consider both z and β as inputs to ML models, as was done in Asim et al. (2018b) Asim et al. (2018a) Asim et al. (2020), and will be done in the current work.

While some authors, such as Last et al. (2016) and Novick and Last (2023), have created new seismicity indicators for input to earthquake prediction models, most work in this area (including that of Last et al. (2016) and Novick and Last (2023)) has incorporated a substantial proportion of the indicators described in this section, ones which have had a long history of investigation as potentially informative of a forthcoming seismic event.

2.2 Past work using seismic indicators for earthquake prediction

This section will review relevant past works in which machine learning (ML) models have been used for earthquake prediction from seismic indicators. The scope of this review should be noted; we do not consider works that attempt to make predictions from raw seismic data or that use a feature set which has no features in common with those we make use of here. Nor, though we have tried to be as thorough as possible, and have searched widely in the literature, do we guarantee that all relevant works have been included. Table 2 lists those works we have reviewed.

The earliest work using seismic indicators for ML-based earthquake prediction was that of Ma et al. (1999), who introduced the first six of the eight indicators listed in Table 1. Panakkat and Adeli (2007) added a further two indicators, μ and c, to this list, using these also in Adeli and Panakkat (2009) and Panakkat and Adeli (2009). Indicators from this set, which we here refer to as the 'MPA indicators,' have been used by a substantial number of other authors, sometimes alone (Asim et al. 2017a, b; Rafiei and Adeli 2017; Zhang et al. 2019; Tehseen et al. 2020; Aslam et al. 2021; Salam et al. 2021), and sometimes along with other indicator features (Martínez-Álvarez et al. 2013; Asencio-Cortés et al. 2016; Last et al. 2016; Asim et al. 2016; Asencio-Cortés et al. 2017; Asim et al. 2018a, 2018b, 2020;

Table 2 Previous studies using seismic indicators for earthquake prediction, focusing on works that make substantial use of the MPA and / or Reyes et al. indicators sets

Refs	Year	Region(s)	Predicted magnitudes	Horizon	*n*	Model(s)
Ma / Panakkat & Adeli (MPA) indicators						
Ma et al. (1999)	1999	China	Max magnitude	One year	6	GA, ANN
Panakkat and Adeli (2007)	2007	SoCal, SF bay	≥ 4.5	Nine months	8	ANN, RBF, RNN
Adeli and Panakkat (2009)	2009	SoCal	in 4.5–−5.0,··,7.0–−7.5	15 days	8	PNN
Panakkat and Adeli (2009)	2009	SoCal	≥ 4.5, 5.0,··,7.0	7–30 days	8	RNN
Asim et al. (2017a)	2017	Hindukush	≥ 5.5	One month	8	ANN, RNN, RF, LPB
Asim et al. (2017b)	2017	Northern Pakistan	≥ 5.0	One month	8	ANN, RBF, SVM, RNN, RF
Rafiei and Adeli (2017)	2017	SoCal	in 4.5–−5.0	7,14,28 days	8	NDC, PNN, PNN+, SVM
Tehseen et al. (2020)	2020	Northern Pakistan	Magnitude	Next EQ	5/8	fuzzy expert system
Aslam et al. (2021)	2021	Hindukush	≥ 5.5	One month	8	SVR + HNN/EPSO
Hasan Al Banna et al. (2021)	2021	Bangladesh	> 4.7	One month	8	LSTM
Al Banna et al. (2021)	2021	Bangladesh	≥ 4.7	One month	8	LSTM with attention
Salam et al. (2021)	2021	SoCal	Magnitude	15 days	7/8	FPA-ELM, FPA-LS-SVM
Sadhukhan et al. (2022)	2022	Japan,Indonesia,Hindukush	Magnitude	Next EQ	8	LSTM, Bi-LSTM, transformer
Reyes et al. indicators						
Reyes et al. (2013)	2013	Chile	≥ mean+0.6×std dev	5 days	7	ANN
Morales-Esteban et al. (2013)	2013	Iberian Peninsula	≥ mean+0.6×std dev	7 days	7	ANN
Shodiq et al. (2018	2021	Indonesia	≥ 6.0	5 days	7	HK-m + ANN
Features from above sets combined together or with additional / novel indicators						
Martínez-Álvarez et al. (2013)	2013	Chile, Iberian Peninsula	≥ mean+0.6×std dev	5, 7 days	16	ANN
Zamani et al. (2013)	2013	Iran	2008 Qeshm EQ	N/A	8	RBF + ANFIS
Asencio-Cortés et al. (2016)	2016	Chile	≥ mean+0.6×std dev	5, 7 days	16	ANN, NB, SVM, KNN, DT
Last et al. (2016)	2016	Israel	> med yearly max mags	One year	26	ANN,M-IFN,SVM,AB,KNN,DT
Asim et al. (2016)	2016	Hindukush	≥= 5.0	15 days	51	DT, RF, RotF, RotB
Asencio-Cortés et al. (2017)	2017	Japan	≥ 5.0	7 days	16	ANN, NB, SVM, KNN, DT
Asencio-Cortés et al. (2018)	2018	California	Max magnitude	7 days	16	GLM, GBM, RF, DNN

Table 2 (continued)

Refs	Year	Region(s)	Predicted magnitudes	Horizon	*n*	Model(s)
Asim et al. (2018b)	2018	Hindukush, Chile, SoCal	≥ 5.0	15 days	60	GP + AB
Asim et al. (2018a)	2018	Hindukush, Chile, SoCal	≥ 5.0	15 days	60	SVR + HNN/EPSO
Florido et al. (2018)	2018	as in Morales-Esteban et al. (2010), Reyeset al. (2013), Asencio-Cortés et al. (2017)	(Morales-Esteban et al. 2010; Reyes et al. 2013; Asencio-Cortés et al. 2017)	(Morales-Esteban et al. 2010; Reyes et al. 2013; Asencio-Cortés et al. 2017)	16	novel tree-based clustering
Zhang et al. (2019)	2019	China	≥ 3.0	14 days	8	DT
Asim et al. (2020)	2020	Cyprus	≥ 3.5, 4.0, 4.5	5–15 days	60	ANN, SVM, RF
Yousefzadeh et al. (2021)	2021	Iran	Max magnitude	7 days	16	SNN, DNN, SVM, DT
Baveja and Singh (2023)	2023	Assam-Guwahati	≥ 4.5, days bet. EQs	Month	52	ELM, SVM
Rashidi and Ghassemieh (2023)	2023	Oklahoma	Magnitude (3 classes)	Half-month	10	SVM, PNN, AB
Novick and Last (2023)	2023	California, Japan, Israel	> Med yearly max mags	One year	94	C4.5,AB,XGB,IFN,KNN,RF, SVM,LR

AB—AdaBoost, *ANFIS*—adaptive network based fuzzy inference system, *ANN*—artificial neural network (NN), *Bi-LSTM*—bi-directional LSTM, *DNN*—deep NN, *DT*—decision tree, *ELM*—extreme learning machine, *EPSO*—enhanced particle swarm optimization, *FPA*—flower pollination algorithm, *GBM*—gradient boosting machine, *GLN*—generalized linear network, *GP*—genetic programming, *HK-m*—hierarchical K-means, *HNN*—hybrid NN, *KNN*—K nearest neighbors, *IFN*—info-fuzzy network, *LPB*—LPBoost, *LSTM*—long short term memory network, *LS-SVM*—least square SVM, *M-IFN*—multi-objective info-fuzzy network, *NB*—naïve Bayes, *NDC*—neural dynamic classification, *PNN*—probabilistic NN, *PNN+*—enhanced PNN, *RBF*—radial basis function NN, *RF*—random forest, *RNN*—recurrent NN, *RotB*—rotation forest + AdaBoost, *RotF*—rotation forest, *SNN*—single layer NN, *SVM*—support vector machine, *SVR*—support vector regression, *XGB*—XGBoost.

n—number of features used by a model, *EQ*—earthquake, *SF*—San Francisco, *SoCal*—Southern California

Yousefzadeh et al. 2021; Hasan Al Banna et al. 2021; Al Banna et al. 2021; Baveja and Singh 2023; Rashidi and Ghassemieh 2023). It should be noted, however, that the μ and c indicators require a substantial history of 'characteristic' (i.e., large) events in order to be well-defined. This in turn requires a long data history and / or a region with a high level of large-magnitude seismicity, and in fact these two indicators are often excluded on these grounds, as was made explicit in Last et al. (2016), and are not among the set of 60 features first proposed by in Asim et al. (2018b), used also in Asim et al. (2018a), and Asim et al. (2020) and in this current work.

Another influential set of indicators, termed here the 'Reyes et al. indicators', were contributed by the work of Reyes and collaborators (Reyes et al. 2013; Morales-Esteban et al. 2013). This set of seven indicators, denoted $x_1 \ldots x_7$, comprise of five differences of b-values that capture changes in the b-value during the last n events, together with x_6, the maximum earthquake magnitude during the last seven days (intended to allow the prediction model to learn the essence of the Omori-Utsu (Utsu 1961) and Bath's (Båth 1965) laws), and x_7, the probability, according to the GR law, of recording an earthquake with magnitude greater than or equal to a specified magnitude. While occasionally used alone and in its entirety, e.g., in Shodiq et al. (2018), most works aside from those of the originators have combined elements from the Reyes et al. set with other seismic indicators (Martínez-Álvarez et al. 2013; Asencio-Cortés et al. 2016, 2017; Asim et al. 2018b, a, 2020; Yousefzadeh et al. 2021), with a particular focus on the use of the x_6 and x_7 indicator features.

A number of previous works (Martínez-Álvarez et al. 2013; Asencio-Cortés et al. 2016, 2017, 2018; Florido et al. 2018; Yousefzadeh et al. 2021) have combined the MPA and Reyes et al. indicators into a single feature set. Other works have investigated the value of adding further seismicity indicators, sometimes (as in Last et al. (2016) and Novick and Last (2023)) entirely novel; or, as in Sadhukhan et al. (2022), derived from a standard time series analysis package such as HCTSA Fulcher and Jones (2014); or sometimes (as with the σ_b, T_{rec}, β, z first used in the work of Zamani et al. (2013), and the 'seismic frequency before main shock' parameter used in Zhang et al. (2019)), known in the geophysical literature but not previously used as inputs to earthquake prediction models.

Asim et al. (2018b) added a substantial number of new features to the MPA and Reyes et al. sets. They included the above-listed features from Zamani et al. (2013), multiple recurrence times T_{rec} (Wiemer and Wyss 1997) (estimates of the time between two earthquakes of magnitude greater than or equal to M'), and in addition, for the calculation of any parameter based on the GR a- and b-values, used both the methods of least squares and of maximum likelihood, its having being noted in Asencio-Cortés et al. (2017) and elsewhere that these methods may under certain circumstances be expected to produce considerably different results. This resulted in a total of 60 seismic indicator features, as listed in Tables 4 and 5. This is the feature set that will be used also in our work. While not the largest feature set used to date (the recent work of Novick and Last (Novick and Last 2023) having used 94 features in total), the Asim et al. set is among the largest, and it will be shown in Sect. 4.4.1 that the T_{rec} features, outside of the Asim et al. work previously used only in Baveja and Singh (2023), have considerable predictive potential.

It may be noted from Table 2 that there are few uses of deep learning. This is understandable given the relatively small size of the datasets (of the order of thousands of examples), which gives rise to a high risk of overfitting if complex models are used. Those studies we have found that use deep learning in a situation in which it is clear the test data are in future of the training data (a topic we will discuss further in this paper), such as Asencio-Cortés et al. (2018), are ones that interpret a 'deep' network as one with a moderate

Table 3 Dataset information for the three regions studied, Hindukush, Chile, and Southern California (SoCal). Data were obtained from Asim et al. (2021)

	Hindukush	Chile	SoCal
Number of data points	4351	7657	33544
Cut-off magnitude	4.0	3.4	2.6
Max magnitude	7.5	8	7.3

Table 4 Nonparametric seismic indicator features (six in total), as used, and further described, in Asim et al. (2018a)

Feature(s)	Description and / or source
T, M_{mean}, $dE^{1/2}$	Ma et al. (1999)
x_6	Reyes et al. (2013)
z	Seismic rate change (method of Habermann) (Habermann 1988)
β	Seismic rate change (method of Matthews & Reasenberg) (Matthews and Reasenberg 1988)

number of neurons and layers. Yet even in such cases the model may be struggling to learn from a limited amount of data; for example, in Asencio-Cortés et al. (2018), the DNN performed worst of the various modeling approaches considered.

3 Data and methods

3.1 Data description

In this study, three datasets (for the Hindukush, Chile, and Southern California regions) used in work by Asim et al. (2018a; 2018b), are used to build our earthquake prediction models. The datasets we used, which were based on earthquake catalog data provided by the United States Geological Survey (USGS) (USGS 2021), were made available by Asim et al. via an online platform (Asim et al. 2021). Properties of these datasets, from the time period January 1980 to December 2016, are given in Table 3. Catalogs from the three regions were evaluated by Asim et al. for cut-off magnitudes, these being given in the table. Only events from the complete (truncated) catalogs were used to compute the 60 seismic features. It was further ascertained from the authors of Asim et al. (2018a) that neither foreshocks nor aftershocks were removed from the provided datasets.

In addition to the seismic indicator features associated with each event, the datasets include the magnitudes of the events, from which binary classification targets can be derived, 'Yes' (1) meaning the magnitude of the next earthquake is greater than or equal to M, 'No' (0) meaning the magnitude of the next earthquake is less than M. The 60 features are of two types, classed, following the nomenclature of Asim et al. (2018a), as *nonparametric* or *parametric*, where the latter are features that depend on a parameter such as an earthquake magnitude or—because they depend upon the GR a- and / or b-values, both of which can be calculated using either *lsq* or *lik*—a mode of calculation. The nonparametric and parametric features are listed in Tables 4 and 5, respectively. Note that in contrast to Asim et al. (2018a) we additionally consider the GR a- and b-values themselves to be parametric, rather than nonparametric, features.

Table 5 Parametric seismic indicator features, as used, and further described, in Asim et al. (2018a). Note that each of these 27 parameters may be calculated by either least squares regression (*lsq*) or the maximum likelihood method (*lik*), which doubles the number of contributed parametric features from the initial 27 to a final 54

Feature(s)	Description and/or source
b, η, M_{def}	Ma et al. (1999)
x_7	Reyes et al. (2013), in this case measuring probability of occurrence of an earthquake with $M \geq 6$
a	y-intercept from GR law Gutenberg and Richter (1944) (level of seismic activity)
σ_b	standard deviation of b value, as used in Zamani et al. (2013)
T_{rec}	probabilistic recurrence time (Wiemer and Wyss 1997), for M in {4.0, 4.1, 4.2,..., 6.0}

The detailed definitions of all these features can be found in Asim et al. (2018a); we will here, in Sect. 4.4.1, present such details only for those features which were found, using Boruta-Shap feature selection, to be the most important in the predictions made by our final, optimized models.

3.2 Train/test split

In the prior work of Asim et al. (2018a), from which the data for this current investigation were derived, datasets were split into training and test data randomly, with training data accounting for 70 % and test data for 30 %. This practice is not uncommon in works in this area, as will be discussed further in Sect. 5. However, from a machine learning point of view, given the temporal nature of the data and the intended use (to predict seismic events in future), it is not ideal, especially due to the long windows of $n = 50$ seismic events used to compute many indicator features, which leads to a high risk of data leakage between train / validation and test sets. The preliminary results of Sect. 4.1 will evidence that a model tested on randomly selected data benefits very substantially from this data leakage. For this reason, beyond this aforementioned preliminary work, we do not use a randomly chosen test set. Instead, our 30 % test set is the last 30 % of the time-ordered data provided for each region. In addition, we introduce a gap of 50 events between the train / validation and test sets, in order to avoid data leakage associated with feature computation. (We use a similar gap between train and validation sets, as discussed in Sect. 3.5.) We note here that for those few initial experiments for which we do use a randomly chosen test set, the split is done a total of 20 times, the data being reshuffled before each new experiment, with the mean values of the chosen performance metric over these 20 separate experiments being presented, together with their standard deviations.

3.3 Creation of less balanced datasets

Most previous studies using seismic indicators for earthquake prediction over a given time-horizon (with the recent exception of Asim et al. (2020)) have focused on a single threshold value M, asking whether or not there would be an earthquake of magnitude $\geq M$ during this period. Our benchmark paper (Asim et al. 2018a) used $M = 5.0$, with a prediction horizon of 15 days. One difference between this work and Asim et al. (2018a) is that we also consider higher magnitude thresholds; specifically, magnitude thresholds of 5, 5.5, 6,

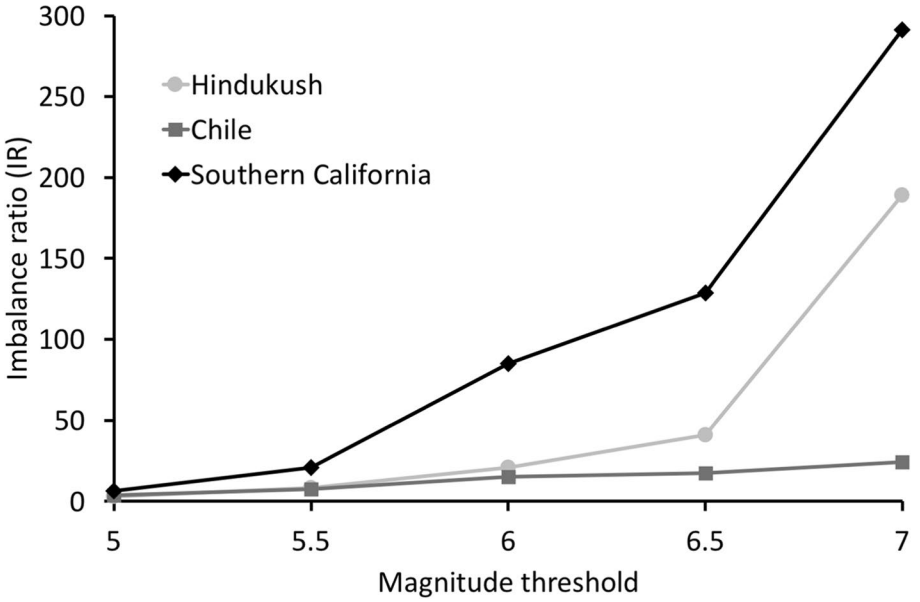

Fig. 1 Increase in imbalance ratio (IR) with prediction magnitude threshold, for each of the three regions considered

and 6.5. However, as higher thresholds are considered, the datasets inevitably become less balanced. The imbalance ratios (IRs) (ratios of positive to negative examples) show a rapid rise as the magnitude threshold increases, as shown in Fig. 1. These higher IRs pose a challenge for any machine learning classifier; however, as will be shown in Sect. 4, our best-performing ensemble models were able to cope well with even the most imbalanced datasets.

3.4 Classifier models

As previously motivated in the introductory Sect. 1, we consider (aside from preliminary benchmarking against the hybrid neural network model of Asim et al. (2018a)) only tree-based ensemble classifiers in this work. We implemented six such classifiers, subdivided into two categories, as outlined below.

3.4.1 Bagging models

Bagging (*bootstrap aggregating*) (Breiman 1996) involves fitting many decision tree models on different training data subsets and averaging the predictions; these subsets are randomly drawn from the whole training dataset with replacement, and each training data subset is used to train a different classifier of the same type, with the class chosen by the largest number of classifiers being, for a given instance, the ensemble decision. Base models can be trained in parallel, which means bagging algorithms can allocate computing resources efficiently. In this work we use the best known bagging model, random forest, and a more recent variant, rotation forest, both of these being briefly described below.

Random forest (Breiman 2001), a extension of the original bagging algorithm (Breiman 1996), includes another type of bagging scheme called 'feature bagging,' such that each new tree will use a random subset of features. In practice, for a classification problem with p features, $\sqrt{p}$ features are used in each new tree.

Rotation forest (Rodriguez et al. 2006) is a enhanced bagging algorithm which transforms the input features into sets of principle components (via PCA or sparse PCA) to obtain better performance. While the mixing of feature dimensions via PCA is not ideal for a model where interpretability is desired, as is the case here, it was of interest to include rotation forest due to the claims in the literature for its effectiveness; specifically, that it was a better classifier than random forest when the input features are not categorical (Bagnall et al. 2018).

3.4.2 Boosting models

Boosting, which as a concept in machine learning dates back to Schapire (1990) Kearns and Valiant (1994), follows a similar principle to bagging in that a set of weak learners are aggregated to obtain a strong learner; however, unlike in the case of bagging, each new weak learner is fitted on the basis of existing weak learners, such that more attention is given to those examples that were badly handled. Shallow decision trees are usually chosen as the base models, in order both to enhance the generalization ability of the final model, and to reduce computing time (as it could become too expensive to fit a sequence of complex models). Gradient boosting is a specific means of construction of such an ensemble, with three essential components: a suitable differentiable loss function; a weak learner (base model); and a gradient descent procedure to minimize the loss when fitting new trees. In this work we implement three of the best-known gradient boosting algorithms, XGBoost, LightGBM, and CatBoost. We also implement RUSBoost, since this is designed for imbalanced datasets and those for higher earthquake magnitude thresholds are substantially imbalanced, as can be seen in Fig. 1. Each of these algorithms will be briefly described below.

XGBoost (*extreme gradient boosting*) (Chen and Guestrin 2016) is a scalable, flexible, and portable gradient boosting library. It can solve many data science problems in a fast and accurate way. It uses regularization within its gradient boosting framework in order to avoid overfitting.

LightGBM (*light gradient boosting machine*) (Ke et al. 2017) is a distributed gradient boosting framework for machine learning, initially created by Microsoft, focusing on performance and scalability; it is free to use and open-source.

CatBoost (Prokhorenkova et al. 2018) is a fast, scalable, high performance gradient boosting library, used for many machine learning tasks. It is open-source and supports GPU acceleration, which can shorten training times dramatically.

RUSBoost (*random undersampled boosting*) (Seiffert et al. 2009) is a variant of gradient boosting that combines data sampling with boosting in order to enhance classification performance when training data are imbalanced.

3.5 Hyperparameter selection

Hyperparameter tuning is an essential task in machine learning, as an optimal choice of non-learnable parameters can result in a large improvement in a model's ability to generalize to out-of-sample data. In this work hyperparameter tuning needed to be done separately

for each of the three datasets, for each of the six ensemble models considered, and moreover, in initial experiments that replicated the random test data selection of Asim et al. (2018a), be repeated 20 times for each dataset, after each reshuffling and subsequent data partitioning. The optimal number of trees in an ensemble was established during preliminary experimentation, as running time itself was in this case one of the selection criteria considered, while the optimal values of all other hyperparameters were obtained in the case of randomly selected test data using five-fold cross-validation or, in the case of temporally split data, using validation on a 21 % dataset (30 % of the 70 % non-testing data) intermediate in time between the train and test data sets. Grid search was used for hyperparameters with a limited number of possible values, and random search, otherwise. We note that in the case of temporally split data, in addition to the gap of 50 events between the validation set and test set that was noted in Sect. 3.2, we use a gap of 50 events between the train and validation sets, to avoid data leakage also between these latter.

3.6 Feature selection

Our benchmark paper (Asim et al. 2018a) used mRMR (*minimum redundancy, maximum relevance*) (Ding and Peng 2005) to select, among the 60 originally proposed seismic indicator features, those most effective for prediction (these being potentially different in each of the considered geographical regions, during to differing underlying geological conditions). However, mRMR is not an ideal feature selector where interpretability is a concern, as it does not rank the retained features, and thus does not give optimal insight into the mechanism of prediction. Two methods of feature selection will be used in the work of this paper: mRMR (on the basis that it was used in Asim et al. (2018a)), together with Boruta-Shap .[1] Boruta-Shap is an extension of the Boruta feature selection algorithm (Kursa and Rudnicki 2010) that allows it to be wrapped around machine learning algorithms other than random forest, and which uses SHAP (SHapley Additive exPlanation) values (Lundberg and Lee 2017), rather than MDI [2], to calculate the average importance of each feature. Boruta-Shap, unlike mRMR, does allow for the ranking of selected features. It will be seen in Sect. 4.3 that the use of Boruta-Shap leads to considerable improvements compared with the use of mRMR, as well as allowing the identification of those features found most useful in prediction.

3.7 Performance measurement

In order to develop our models, which will involve contrasting the performance of many variants, we need a reliable single-number measure of classification success. Accuracy, though still widely popular as a metric, is unreliable and inappropriate for imbalanced classification problems, as it can give misleadingly high scores when there is over-assignment to the majority class. A large number of alternative, more suitable, evaluation metrics are available. We choose to use the area under the ROC curve (AUC), as used also in

[1] See https://github.com/Ekeany/Boruta-Shap for the official repository of Boruta-Shap. Last accessed 27 August 2023.

[2] MDI (*mean decrease in impurity*) sums the number of splits across all trees that involve a given feature, proportionally to the number of samples that the feature splits, and is a means of determining feature importance that can be used with any tree-based model.

the studies of Last et al. (2016) and Novick and Last (2023) as our primary performance measure. The ROC curve shows the trade-off between true positives,

$$TPR = \frac{TP}{(TP + FN)}, \tag{2}$$

and false positives,

$$FPR = \frac{FP}{(FP + TN)}, \tag{3}$$

(where TP is the number of true positives, FN the number of false negatives, TN the number of true negatives, and FP the number of false positives). The line $y = x$ would correspond to the random guessing of a class, resulting in an AUC of 0.5. The AUC has the attractive property of being independent of classification threshold, and thus very suitable for the assessment of a classifier on imbalanced data, for which the chosen threshold can otherwise make a large difference to assessed performance. However, we additionally use the Matthews correlation coefficient (MCC) (Matthews 1975), given by

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

due to its use in Asim et al. (2018a), making it valuable for comparison purpose, and because it has been argued to be the most informative single metric that can be derived from a confusion matrix (Chicco and Jurman 2020). MCC values fall in the range $[-1, +1]$, with a $+1$ value indicating a perfect classification, a -1 value a perfect misclassification (unlikely in practice), and a value of 0 (equivalent to an AUC of 0.5) meaning either that the classifier is no better than random, or that all examples are assigned to the majority class. It is this last property that makes the MCC highly useful for imbalanced datasets, though it should be noted that the MCC needs always to be calculated with respect to a chosen threshold, in our case a default threshold value of 0.5.

4 Results

4.1 Comparison of random forest and XGBoost with benchmark model

In this first set of experiments, results from the two best-known of our set of ensemble models, random forest (RF) (Breiman 2001) and XGBoost (XGB) (Chen and Guestrin 2016), as examples of bagging and boosting models, respectively, are compared to those of our benchmark model, (Asim et al. 2018a). Figure 2 compares the results for all three of these models, for each of the three regions, at a magnitude threshold of 5.0 (this threshold being picked for the initial comparison stage as it was the value used in our benchmark paper.

Following from the discussion in Sect. 3.2, comparison is done at this stage between results for temporal and random test splits, as well as between the three models. The most striking observation from Fig. 2 is the difference in observed test performances when the test data are selected temporally, to be the last 30 % in the data series and with a 50-event gap between validation and test sets, in this way avoiding data leakage. Test MCCs are in this case much lower than when the 30 % test data are selected randomly, as was done

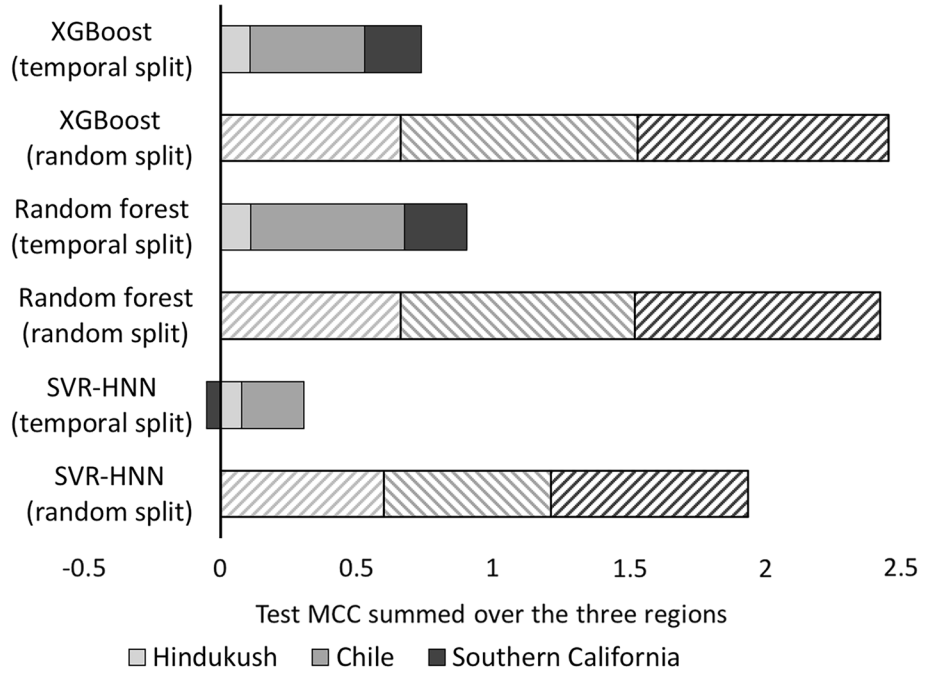

Fig. 2 Test performance for each region in terms of Matthews correlation coefficient (MCC). Comparison of random forest and XGBoost with the model of Asim et al. (2018a) (SVR-HNN) at magnitude threshold 5.0. Comparison of the three models is done for both temporal splits, in which the 30 % test data are taken from the end of the supplied data series, and random splits (as used in Asim et al. (2018a)), in which the 30 % test data are chosen randomly (mean of 20 runs is shown in these latter cases)

in our benchmark paper, and this can be seen to be especially true for SVR-HNN model used in our benchmark paper, Asim et al. (2018a). However, despite the lower test performances for the temporally split data, we from this point onward choose to work only with datasets for which the test period is the last 30 % of each series, on the grounds, as previously argued in Sect. 3.2, that these results give a more realistic and practically-useful picture of the performance of models trained on seismic indicator data.

Considering, therefore, only the temporal split results of Fig. 2, it is evident that both XGB and RF are considerably more effective than the SVR-HNN model, being around 41 % (average of XGB and RF results) better for Hindukush and around 116 % better for Chile. (In the case of Southern California it is difficult to compare as the MCC for the temporally-split SVR-HNN model returned a small negative value.) Given that the performance of the tree-based classifiers was additionally achieved without feature selection beyond that which is integral to the XGB and RF models (Sect. 4.3 will look at whether the performance of our best-performing model can be improved by the use of additional feature selection), and given that we intend to make predictions for magnitude thresholds M greater than 5.0, on the basis of this preliminary work we will not make further use of the SVR-HNN model as a comparison model. Instead, in the next section, we compare the results from RF and XGB, now at a range of values of M, with other tree-based ensemble classifiers, in both the bagging and boosting categories.

Table 6 Test AUCs for all ensemble models, for Hindukush, Chile, and Southern California, for a range of prediction magnitude thresholds *M*

	RF	RoF	XGB	LGBM	CatB	RUSB
M	Hindukush					
5.0	**0.5303**	0.4607	0.4661	0.4752	0.5256	0.4605
5.5	0.5356	0.5479	0.5785	0.5601	**0.5900**	0.5041
6.0	0.4574	0.4645	0.4993	0.4772	**0.5232**	0.5204
6.5	0.4934	0.5254	0.5961	0.6063	**0.6259**	0.5894
M	Chile					
5.0	0.7631	0.6852	**0.8459**	0.7775	0.8151	0.7199
5.5	0.8378	0.7660	0.8669	0.8808	**0.8970**	0.7421
6.0	0.8166	0.8411	0.8834	0.8808	**0.8960**	0.8283
6.5	0.7998	0.9044	0.8841	0.9051	**0.9138**	0.8734
M	Southern California					
5.0	0.8352	0.7924	0.7648	0.7425	**0.8100**	0.5701
5.5	**0.5251**	0.4937	0.5244	0.4967	0.4939	0.4795
6.0	0.5363	0.4422	0.4619	0.4640	**0.5762**	0.5291
6.5	0.6124	0.6829	0.4793	0.6191	**0.7301**	0.5428

RF—random forest (Breiman 2001), *RoF*—rotation forest (Rodriguez et al. 2006), *XGB*—XGBoost (Chen and Guestrin 2016), *LGBM*—LightGBM (Ke et al. 2017), *CatB*—CatBoost (Prokhorenkova et al. 2018), *RUSB*—RUSBoost (Seiffert et al. 2009)

The bold type indicates the best-performing model at a given threshold

4.2 Comparison of random forest and XGBoost with other ensemble models

This section will compare the results obtained from XGBoost and random forest, for each of the three regions, and for a range of magnitude thresholds, with prediction results from the other tree-based ensemble algorithms described in Sect. 3.4: rotation forest (Rodriguez et al. 2006) (bagging); and LightGBM (Ke et al. 2017), CatBoost (Prokhorenkova et al. 2018), and RUSBoost (Seiffert et al. 2009) (boosting). Table 6 summarizes the results, in terms of AUC, for each of the considered regions, for magnitude thresholds *M* from 5.0 (the single threshold considered in Asim et al. (2018a)) up to 6.5 (there being, as noted in Sect. 3.2, no earthquakes of magnitude 7.0 or above in any of the three regions during the test period). It is evident from Table 6 that boosting models in general perform better for these data than the two bagging models considered, as in only 2/12 cases, those of the Hindukush dataset at magnitude threshold 5.0 and the Southern California dataset at magnitude threshold 5.5, does a bagging model (random forest) perform best. Considering the various boosting models, while there was a single win here for XGBoost (Chile dataset at magnitude threshold 5.0), it can be seen that CatBoost is overall the most effective algorithm of those trialed here, being best-performing in 9/12 cases; this will be the model we take forward to the next section, which explores the benefits of feature selection.

Table 7 Full CatBoost (all 60 features) compared with Boruta-Shap and mRMR feature selection for each of the three regions, as a function of magnitude *M*. In the case of the latter two, the number of retained features is given, in brackets, for each region. The bold type indicates the best-performing model at a given threshold

	Hindukush		
M	CatBoost	Boruta-Shap (8/60)	mRMR (13/60)
5.0	0.5056	**0.5103**	0.4823
5.5	0.5900	**0.5967**	0.5345
6.0	0.5232	**0.5535**	0.4533
6.5	0.6259	**0.6910**	0.6165
	Chile		
M	CatBoost	Boruta-Shap (31/60)	mRMR (7/60)
5.0	0.8151	**0.8287**	0.8150
5.5	**0.8970**	0.8907	0.8680
6.0	0.8960	**0.9221**	0.8897
6.5	0.9138	**0.9178**	0.9161
	Southern California		
M	CatBoost	Boruta-Shap (48/60)	mRMR (10/60)
5.0	0.8109	**0.8822**	0.8258
5.5	0.4939	**0.5399**	0.5123
6.0	0.5762	**0.5917**	0.3565
6.5	0.7301	**0.8384**	0.5695

4.3 CatBoost results after feature selection

This section will investigate whether the CatBoost results can be further improved by the use of either mRMR (as used in Asim et al. (2018a)) or Boruta-Shap feature selection.

In the case of Boruta-Shap feature selection (these steps being inapplicable to a non-ranking method such as mRMR), the following procedure was used:

1. For each region, and for each magnitude threshold, SHAP feature importances are obtained from an all-features CatBoost model.
2. For each region, importances are then averaged over all thresholds for which the use of a restricted set of features is not damaging (because we want to discover which features are *most useful* for earthquake prediction), in order to have an identical feature subset for every magnitude threshold.
3. For each region, these averages are then used to select a top-*N* feature subset, the decision about the top-*N* being made on the basis of performance at the highest magnitude for a given region, as for lower magnitudes the performances are less easily distinguished. The procedure used is to progressively discard features, until this results in a statistically significant performance decrease (using a paired t-test).

This procedure resulted in the conclusion that the top 8, 31, and 48 features should be retained for Hindukush, Chile, and Southern California, respectively. The top-*N* optimized results from Boruta-Shap were then compared to the results obtained using mRMR feature selection, and also to the CatBoost model without feature selection. This was done for each of our three studied regions, as shown in Table 7. The top-*N*-optimized results are superior, compared to both the use of all 60 original features, in all but one case (Chile dataset at magnitude threshold 5.5). In contrast, mRMR feature selection underperforms

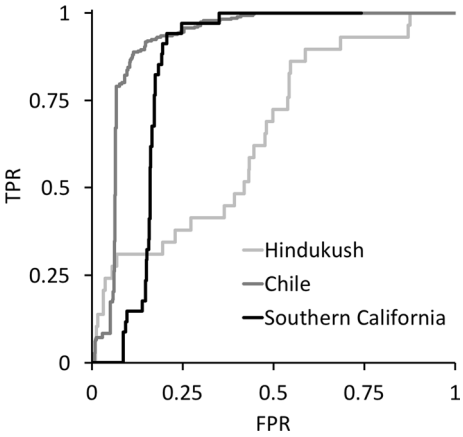

Fig. 3 ROC curves for the top-*N*-optimized CatBoost models, at magnitude threshold 6.5 (the highest used in this study) for the three considered regions

Boruta-Shap feature selection in every case, and underperforms CatBoost without feature selection in 9/12 cases. While mRMR was beneficial to the model of Asim et al. (2018a), it did not prove equally effective here.

Figure 3 shows ROC curves (AUC values given in Table 7) for our optimized models at the highest threshold of 6.5 considered here, giving rise to imbalance ratios ranging from 17.6:1 for Chile to 128.9 for Southern California. It is clear from both the figure and the table that the model trained on data from the Hindukush region performs considerably less well than those for Chile and Southern California; however, we note that the Hindukush dataset was the smallest of the three (4350 examples in total, compared to a total of 33,543 examples in the case of the largest dataset, that for Southern California), and this relative scarcity of data seems likely to have been a strong factor in the underperformance of the model for the Hindukush region.

4.4 Feature importances and interpretation

4.4.1 Ranking of the features

Figures 4, 5, and 6 show the rankings of the most important features for the cases of Hindukush, Chile, and Southern California, respectively, in which, for the parametric features (using the terminology of Asim et al. (2018a) and Sect. 3.1), the mode of calculation (least squares (*lsq*) and maximum likelihood (*lik*)), is denoted in brackets.

The most substantial commonality between the three regions is the large proportion of *probabilistic recurrence times* (T_{rec} values) (Wiemer and Wyss 1997) among the top-*N* features, T_{rec} being the time between two earthquakes of magnitude greater than or equal to a specified value M',

$$T_{rec} = \frac{T}{10^{a-bM'}}, \tag{5}$$

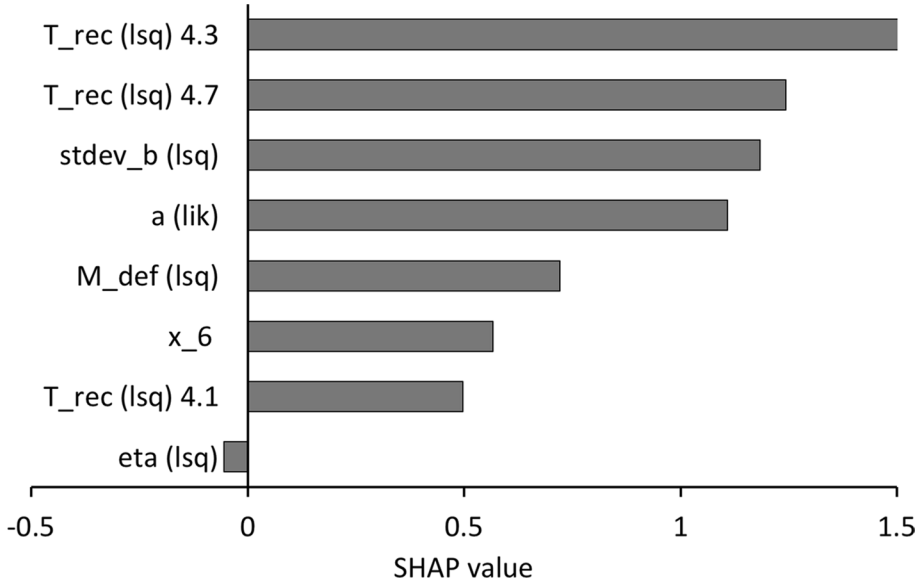

Fig. 4 Top-8 feature importances for Hindukush

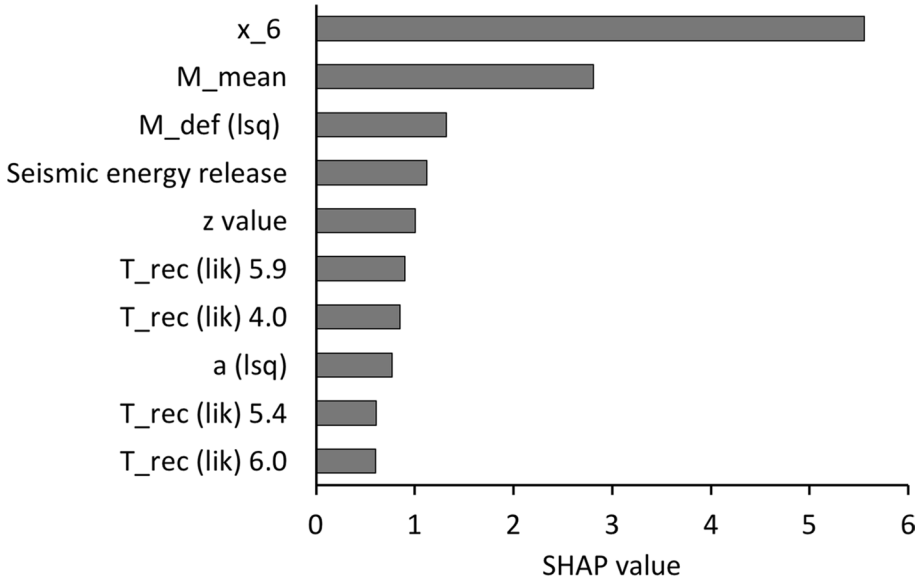

Fig. 5 Highest-ranked 10 (out of top-31 used) feature importances for Chile

T_{rec} features accounting for 37.5 % of the top-N features in the case of Hindukush (3/8), 51.6 % in the case of Chile (16/31), and 62.5 % in the case of Southern California (30/48). While none of the highest ranked 10 features for Southern California (Fig. 6) are recurrence times, the bulk of all those other features used by this model are T_{rec}s. These features, first introduced as inputs to an earthquake prediction model by in Asim et al. (2016), and

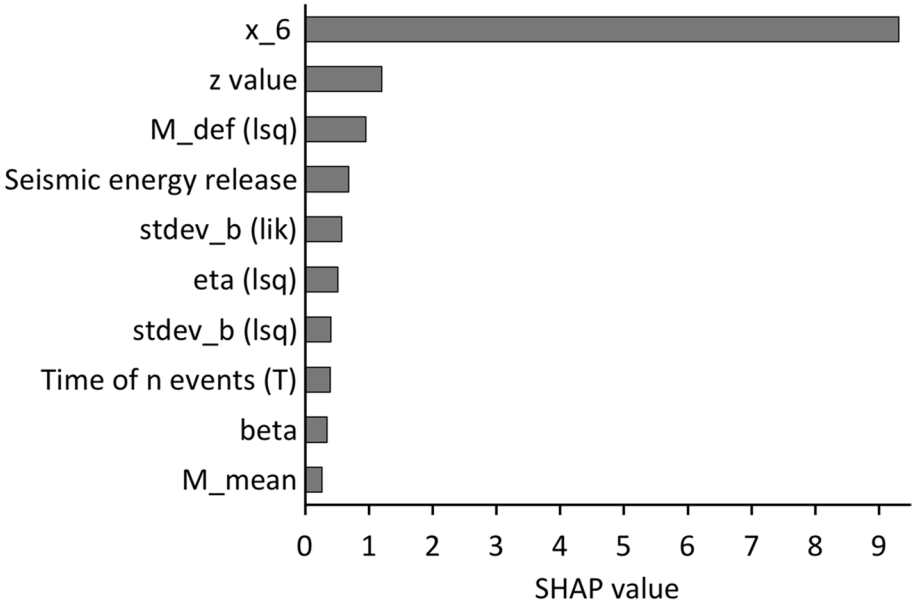

Fig. 6 Highest-ranked 10 (out of top-48 used) feature importances for Southern California

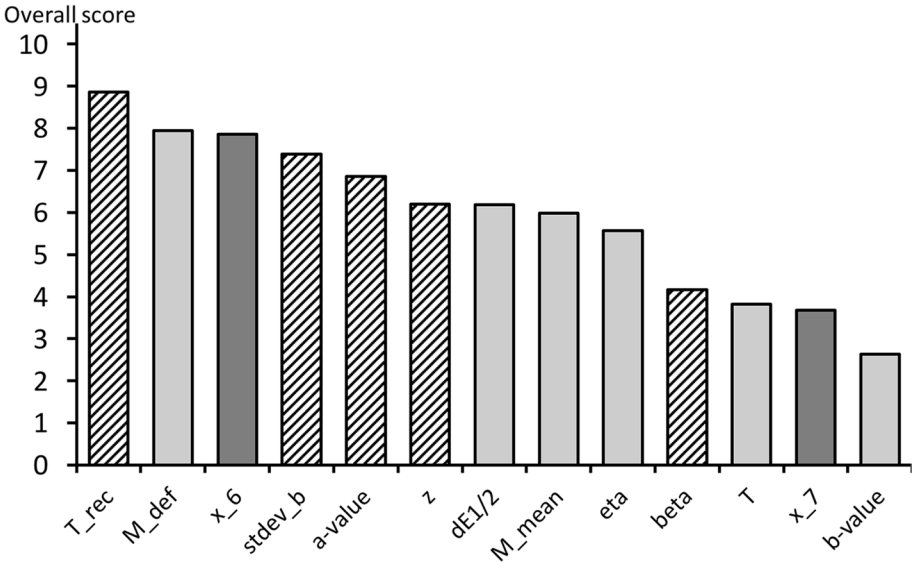

Fig. 7 Overall scores, a score of 10 corresponding to the highest-ranked feature, across all three regions, for the top 8, 31, and 48 features for Hindukush, Chile, and Southern California, respectively. Darker shading denotes Reyes et al. indicators, lighter shading denoted indicators from the MPA set (with dE1/2 being the seismic energy release parameter), and diagonal striping denotes additional indicators used in Asim et al. (2018a). (See text for details of scoring method)

subsequently used, aside from the work of the Asim group and this current work, only in a very recent work by Baveja and Singh (2023), thus appear to be ones of considerable potential value.

To discover further commonalities in the features used by the models, the following procedure was used to create a merged picture, in Fig. 7, of the value of the top-N parameters for the three regions.

1. For each top-N retained feature set ($N = 8, 31, 48$ for Hindukush, Chile, Southern California, respectively), create a score from 10 (most important) down to 1 (least important) for each rank $= 1..N$ feature:

$$\text{score} = (10N - 9 \times \text{rank} - 1)/(N - 1) \tag{6}$$

2. Within each list of scores, retain only the top-scoring instance of each a- and / or b-dependent parametric feature, independently of whether it was calculated using the *lsq* or *lik* method, and in the case of the T_{rec} features, retain only the instance with the highest score, independent of magnitude M'.
3. For every feature that occurs anywhere in the remaining lists, calculate its average score across the three regions, where any feature not in a retained list for a region contributes zero to this average. This average then constitutes the 'overall score' of Fig. 7.

It is seen from Fig. 7 that, aside from T_{rec} features, the overall top-ranked features are the MPA *magnitude deficit* (M_{def}) parameter, the difference between the maximum observed earthquake magnitude and the maximum expected earthquake magnitude, as calculated from the Gutenberg-Richter law,

$$M_{def} = M_{max,actual} - M_{max,expected}, \tag{7}$$

where

$$M_{max,expected} = a/b, \tag{8}$$

x_6, the maximum earthquake magnitude during the last seven days, and the standard deviation of the Gutenberg-Richter b-value. The b-value in itself is not, however, a highly ranked predictive feature, despite the many works discussed in El-Isa and Eaton (2014) that have studied the behavior of the b-value in the run-up to a seismic event.

4.4.2 Comparison with other ranking studies

Works which examine feature importances are relatively few, in proportion to the total number of prediction studies that use seismic indicators as model inputs, and only four previous papers, to the authors' knowledge, have ranked seismicity indicators in this way: Martínez-Álvarez et al. (2013), Last et al. (2016), Zhang et al. (2019), Yousefzadeh et al. (2021). (Novick and Last (2023) use feature selection but do not provide a ranking of their features). The comparison of these four works with our own, and comparison within this group of works, is not, however, straightforward, given that the five works use different feature sets (in the case of Last et al. (2016) considerably different, as the majority of features in this work are novel ones) and different methods of ranking. However, we will attempt to draw some broad conclusions below:

- Among those studies that used the Reyes et al. x_6 feature (our own and those of Yousefzadeh et al. (2021) and Martínez-Álvarez et al. (2013)), this feature was highly ranked by all (top-ranked here for Chile and Southern California, top-ranked in Yousefzadeh et al. (2021), and ranked second for both the Chile and Iberian Peninsula regions in Martínez-Álvarez et al. (2013)).
- There is substantial agreement about the lack of value, as a predictor, of the Gutenberg-Richter *b*-value; *b* is low-ranked by us and in both regions studied by Zhang et al. Zhang et al. (2019), and is lowest-ranked in Martínez-Álvarez et al. (2013) and also (among those features which this work shares with our models) in Last et al. (2016). As noted previously, this lack of predictive value of *b* is a surprise given the many studies of its pre-event behavior.

5 Discussion

This paper set out to investigate the value of tree-based ensembles for earthquake prediction when used with a recently-proposed large set of 60 seismic indicators, first used by Asim et al. in Asim et al. (2018b) to predict earthquakes in the Hindukush, Chile, and Southern California regions, and made available (Asim et al. 2021) by these authors in connection with a later work (Asim et al. 2018a) focusing on the same regions. This dataset was of additional interest due to its novel inclusion of earthquake recurrence times as features. The prior work of Asim et al. (2018b) and Asim et al. (2018a) considered a range of ML models, but tree-based ensembles, despite their demonstrated potential value for earthquake prediction (discussed in Sect. 1), were not among the classifiers investigated. We note that random forest was later used, among other models, by the same authors to make predictions for a Cyprus dataset with the same features (Asim et al. 2020), and had performed well, but because the test set in that case had, as in Asim et al. (2018a), been chosen randomly rather than strictly in future of the training period, we felt that this result, while positive, could only be indicative. We were additionally interested to discover whether our tree-based ensemble models could make effective predictions at higher magnitude thresholds than considered in Asim et al. (2018a), and if Boruta-Shap feature selection could be used to further improve the results and give insight into the value of features.

In Sects. 3.2 and 4.1 we addressed the issue of test data selection. Figure 2, in Sect. 4.1, shows that choosing the 30 % test set randomly, as was done in Asim et al. (2018a), leads to test results with a much higher performance metric than those for a temporal data split. (It should be noted that the use of randomly selected test data is not restricted to the work of Asim et al. (2018a). Of the 33 studies listed in Table 2, 7/33 either stated clearly that their test data selection was non-temporal, or we discovered this to be so when downloading available code, while a further 7/33 left the origin of the test data unclear.) However, only the selection of test data clearly in future of that used for training and validation allows for the development of a reliable model.

Having from this point onward chosen to work with temporally split data exclusively, Sect. 4.1 demonstrated that the use of either random forest or XGBoost (XGB) could lead to substantial performance improvements over the hybrid model of Asim et al. (2018a). It was subsequently shown in Sects. 4.2 and 4.3 that the results from RF and XGB could be improved upon first by considering alternative tree-based ensemble predictors, the best-performing of the investigated set being discovered to be CatBoost (Prokhorenkova et al. 2018), and second by then adding Boruta-Shap feature selection to the CatBoost model.

It was demonstrated that our best-performing models could perform well at magnitude thresholds beyond the threshold of $M = 5.0$ our work shared with that of Asim et al. (2018a), despite the rapidly increasing imbalance ratios of the datasets. The use of Boruta-Shap had the further advantage of providing feature rankings, and hence insight into the usage of features, as was detailed in Sect. 4.4.

Turning to future work, one minor improvement would be to move from using a single validation period (here, the last 30 % of the 70 % of the data used for training and validation) to the use of cross-validation. However, this would need to be of a form appropriate for time series data, which respected temporal ordering and ensured there could be no data leakage between any train set and its subsequent validation set. It would also be of interest to look at different prediction horizons. The Cyprus study (Asim et al. 2020) of Asim et al. considered not only a range of magnitude thresholds (3.0, 3.5, 4.0, and 4.5) but also a range of prediction horizons from 5 to 15 days, and showed that prediction at the shortest, five-day, horizon was much harder than at the longest. In this current work we have considered only a 15 day prediction horizon, as in Asim et al. (2018a), and it would clearly be valuable to challenge our model, for the Hindukush, Chile, and Southern California regions, to predict at the same set of magnitude thresholds, but for also for a set of shorter (and possibly, also, longer) horizons.

Finally, it would also be of value to use the methods of this paper in other geographical regions. However, the amount of available data should be considered, as should the imbalance ratio, if the intention is to predict larger earthquakes (of magnitude 5.0 or more). It is evident in Fig. 3 and Table 7 that the Hindukush model, for which the amount of training data was smallest, performed substantially less well than the models for Chile and Southern California. However, the best-performing model was not that for Southern California, which had been trained on by far the largest dataset, but that for Chile. This is due to the fact that Chile has both a dense seismic monitoring network (increasing the size of the dataset) and a high number of large earthquakes (reducing the imbalance ratio). The obvious candidate for further study is therefore Japan; in the recent work of Novick and Last (Novick and Last 2023), applying a range of ML models to seismic indicator data from California, Israel, and Japan, the highest test AUC values were obtained for the Japan dataset.

6 Conclusions

This study has confirmed the value of tree-based ensemble classifiers for the task of earthquake prediction from seismic indicators, and identified CatBoost (Prokhorenkova et al. 2018) as especially promising for this task. It has shown that the use of Boruta-Shap feature selection can further improve an already strong ensemble classifier, and that it is beneficial to begin with an expanded set of features, here those 60 features first proposed by in Asim et al. (2018b), incorporating seismicity indicators from multiple sources, with earthquake recurrence times, first proposed as predictive features by in Asim et al. (2016), emerging here as having particular potential value. This work has in addition adopted a methodology designed to eliminate data leakage, something not always present in machine learning studies within this topic area. It has shown that a lack of caution in the choice of test data can lead to possible overestimates of the capability of earthquake prediction models. It has also shown, more positively, that when test data are correctly separated in time from training and validation data, useful levels of predictive ability can even so be demonstrated by

tree-based ensemble models trained on seismic indicator data, under these more rigorous and realistic conditions.

Author contributions YZ contributed to conceptualization, methodology, software, investigation, validation, formal analysis, writing—review and editing; DG contributed to conceptualization, formal analysis, visualization, writing —original draft, writing—review and editing.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The data used in this study, associated originally with [ref 28], may be downloaded from https://figshare.com/articles/Earthquake Prediction using SVR and HNN/6406814, in order to emphasise that the data used in the study were made available in association with another work, but we cannot find a way to make the 'cite reference' function work.

Code availability The code used to generate the results of this study may be downloaded from https://github.com/sanmaoyang/Earthquake-prediction-from-seismic-indicators-using-tree-based-ensemble-learning.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Adeli H, Panakkat A (2009) A probabilistic neural network for earthquake magnitude prediction. Neural Netw 22(7):1018–1024. https://doi.org/10.1016/j.neunet.2009.05.003

Al Banna MH, Ghosh T, Al Nahian MJ, Taher KA, Kaiser MS, Mahmud M, Hossain MS, Andersson K (2021) Attention-based bi-directional long-short term memory network for earthquake prediction. IEEE Access 9:56589–56603. https://doi.org/10.1109/ACCESS.2021.3071400

Allen CR (1976) Responsibilities in earthquake prediction: to the Seismological Society of America, delivered in Edmonton, Alberta, May 12, 1976. Bull Seismol Soc Am 66(6):2069–2074

Arrowsmith S, Trugman D, MacCarthy J, Bergen K, Lumley D, Magnani M (2022) Big data seismology. Rev Geophys 60(2):000769. https://doi.org/10.1029/2021rg000769

Asencio-Cortés G, Martínez-Álvarez F, Morales-Esteban A, Reyes J (2016) A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction. Knowl Based Syst 101:15–30. https://doi.org/10.1016/j.knosys.2016.02.014

Asencio-Cortés G, Martínez-Álvarez F, Troncoso A, Morales-Esteban A (2017) Medium-large earthquake magnitude prediction in Tokyo with artificial neural networks. Neural Comput Appl 28(5):1043–1055. https://doi.org/10.1007/s00521-015-2121-7

Asencio-Cortés G, Morales-Esteban A, Shang X, Martínez-Álvarez F (2018) Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure. Comput Geosci 115:198–210. https://doi.org/10.1016/j.cageo.2017.10.011

Asim KM, Idris A, Martínez-Álvarez F, Iqbal T (2016) Short term earthquake prediction in Hindukush region using tree based ensemble learning. In: 2016 international conference on frontiers of information technology (FIT), 365–370. https://doi.org/10.1109/FIT.2016.073 . IEEE

Asim K, Martínez-Álvarez F, Basit A, Iqbal T (2017) Earthquake magnitude prediction in Hindukush region using machine learning techniques. Nat Hazards 85(1):471–486. https://doi.org/10.1007/s11069-016-2579-3

Asim KM, Awais M, Martínez-Álvarez F, Iqbal T (2017) Seismic activity prediction using computational intelligence techniques in northern Pakistan. Acta Geophysica 65(5):919–930. https://doi.org/10.1007/s11600-017-0082-1

Asim KM, Idris A, Iqbal T, Martínez-Álvarez F (2018a) Earthquake prediction model using support vector regressor and hybrid neural networks. PLOS One 13(7):0199004. https://doi.org/10.1371/journal.pone.0199004

Asim KM, Idris A, Iqbal T, Martínez-Álvarez F (2018b) Seismic indicators based earthquake predictor system using Genetic Programming and AdaBoost classification. Soil Dyn Earthq Eng 111:1–7. https://doi.org/10.1016/j.soildyn.2018.04.020
Asim KM, Moustafa SS, Niaz IA, Elawadi EA, Iqbal T, Martínez-Álvarez F (2020) Seismicity analysis and machine learning models for short-term low magnitude seismic activity predictions in cyprus. Soil Dyn Earthq Eng 130:105932. https://doi.org/10.1016/j.soildyn.2019.105932
Asim KM, Idris A, Iqbal T, Martínez-Álvarez F (2021) Earthquake prediction datasets. https://figshare.com/articles/dataset/Earthquake_Prediction_using_SVR_and_HNN/6406814
Aslam B, Zafar A, Khalil U, Azam U (2021) Seismic activity prediction of the northern part of Pakistan from novel machine learning technique. J Seismol 25(2):639–652. https://doi.org/10.1007/s10950-021-09982-3
Bagnall A, Flynn M, Large J, Line J, Bostrom A, Cawley G (2018) Is rotation forest the best classifier for problems with continuous features? arXiv preprint arXiv:1809.06705v2
Båth M (1965) Lateral inhomogeneities of the upper mantle. Tectonophysics 2(6):483–514. https://doi.org/10.1016/0040-1951(65)90003-x
Baveja GS, Singh J (2023) Earthquake magnitude and b value prediction model using extreme learning machine. arXiv preprint arXiv:2301.09756
Bergen KJ, Johnson PA, Hoop MV, Beroza GC (2019) Machine learning for data-driven discovery in solid Earth geoscience. Science 363(6433):0323
Beroza GC, Segou M, Mostafa Mousavi S (2021) Machine learning and earthquake forecasting-next steps. Nat Commun 12(1):1–3. https://doi.org/10.1038/s41467-021-24952-6
Blank D, Morgan J (2021) Can deep learning predict complete ruptures in numerical megathrust faults? Geophys Res Lett 48(18):2021–092607. https://doi.org/10.1029/2021gl092607
Breiman L (1996) Bagging predictors. Machine Learn 24(2):123–140. https://doi.org/10.1007/bf00058655
Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/a:1010933404324
Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom 21(1):1–13. https://doi.org/10.1186/s12864-019-6413-7
Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3(02):185–205. https://doi.org/10.1142/s0219720005001004
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
El-Isa ZH, Eaton DW (2014) Spatiotemporal variations in the b-value of earthquake magnitude-frequency distributions: classification and causes. Tectonophysics 615:1–11. https://doi.org/10.1016/j.tecto.2013.12.001
Florido E, Asencio-Cortés G, Aznarte JL, Rubio-Escudero C, Martínez-Álvarez F (2018) A novel tree-based algorithm to discover seismic patterns in earthquake catalogs. Comput Geosci 115:96–104. https://doi.org/10.1016/j.cageo.2018.03.005
Fulcher B, Jones N (2014) Highly comparative feature-based time-series classification. IEEE Transact Knowl Data Eng 26:3026–3037. https://doi.org/10.1109/tkde.2014.2316504
Galkina A, Grafeeva N (2019) Machine learning methods for earthquake prediction: a survey. In: Proceedings of the fourth conference on software engineering and information management (SEIM-2019), Saint Petersburg, Russia, 13, 25. https://ceur-ws.org/Vol-2372/SEIM_2019_paper_31.pdf
Geller RJ (1997) Earthquake prediction: a critical review. Geophys J Int 131(3):425–450. https://doi.org/10.1111/j.1365-246x.1997.tb06588.x
Gorse D, Goel A (2022) Deep vs. shallow learning: a benchmark study in low magnitude earthquake detection. In: 83rd eage annual conference & exhibition, pp 1–5. https://doi.org/10.3997/2214-4609.202210042 . European association of geoscientists & engineers
Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. Bull Seismol Soc Am 34(4):185–188. https://doi.org/10.1038/156371a0
Habermann R (1988) Precursory seismic quiescence: past, present, and future. Pure Appl Geophys 126(2):279–318. https://doi.org/10.1007/bf00879000
Hasan Al Banna M, Ghosh T, Taher KA, Kaiser MS, Mahmud M (2021) An earthquake prediction system for Bangladesh using deep long short-term memory architecture. Intell Syst Proc ICMIB 2020:465–476. https://doi.org/10.1007/978-981-33-6081-5_41
Ikram A, Qamar U (2015) Developing an expert system based on association rules and predicate logic for earthquake prediction. Knowl Based Syst 75:87–103. https://doi.org/10.1016/j.knosys.2014.11.024

Jaumé SC, Sykes LR (1999) Evolving towards a critical point: a review of accelerating seismic moment/energy release prior to large and great earthquakes. Seismicity patterns, their statistical significance and physical meaning, pp 279–305 https://doi.org/10.1007/s000240050266
Johnson PA, Rouet-Leduc B, Pyrak-Nolte LJ, Beroza GC, Marone CJ, Hulbert C, Howard A, Singer P, Gordeev D, Karaflos D (2021) Laboratory earthquake forecasting: a machine learning competition. Proc Natl Acad Sci 118(5):2011362118. https://doi.org/10.1073/pnas.2011362118
Kearns M, Valiant L (1994) Cryptographic limitations on learning boolean formulae and finite automata. J ACM (JACM) 41(1):67–95. https://doi.org/10.1145/174644.174647
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inform Process Syst, 30
Kong Q, Trugman DT, Ross ZE, Bianco MJ, Meade BJ, Gerstoft P (2019) Machine learning in seismology: turning data into insights. Seismol Res Lett 90(1):3–14. https://doi.org/10.1785/0220180259
Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Statist Softw 36:1–13. https://doi.org/10.18637/jss.v036.i11
Last M, Rabinowitz N, Leonard G (2016) Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. PLOS One 11(1):0146101. https://doi.org/10.1371/journal.pone.0146101
Laurenti L, Tinti E, Galasso F, Franco L, Marone C (2022) Deep learning for laboratory earthquake prediction and autoregressive forecasting of fault zone stress. Earth Planetary Sci Lett 598:117825. https://doi.org/10.1016/j.epsl.2022.117825
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inform Process Syst, 30 arXiv:1705.07874
Martínez-Álvarez F, Reyes J, Morales-Esteban A, Rubio-Escudero C (2013) Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. Knowl Based Syst 50:198–210. https://doi.org/10.1016/j.knosys.2013.06.011
Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)- Protein Struct 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9
Matthews MV, Reasenberg PA (1988) Statistical methods for investigating quiescence and other temporal seismicity patterns. Pure Appl Geophys 126(2):357–372. https://doi.org/10.1007/bf00879003
Ma L, Zhu L, Shi Y (1999) Attempts at using seismicity indicators for the prediction of large earthquakes by genetic algorithm-neural network method. Proc Asia-Pacific Econ Cooper Earthq Simul Brisbane Australia 31:483–489
Mignan A, Broccardo M (2020) Neural network applications in earthquake prediction (1994–2019): meta-analytic and statistical insights on their limitations. Seismol Res Lett 91(4):2330–2342. https://doi.org/10.1785/0220200021
Miranda JD, Gamboa CA, Flórez A, Altuve M (2019) Voting-based seismic data classification system using logistic regression models. In: 2019 XXII symposium on image, signal processing and artificial vision (STSIVA), pp 1–5. https://doi.org/10.1109/STSIVA.2019.8730280 . IEEE
Morales-Esteban A, Martínez-Álvarez F, Troncoso A, Justo J, Rubio-Escudero C (2010) Pattern recognition to forecast seismic time series. Exp Syst Appl 37(12):8333–8342. https://doi.org/10.1016/j.eswa.2010.05.050
Morales-Esteban A, Martínez-Álvarez F, Reyes J (2013) Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence. Tectonophysics 593:121–134. https://doi.org/10.1016/j.tecto.2013.02.036
Mousavi SM, Beroza GC (2022) Deep-learning seismology. Science 377(6607):4470. https://doi.org/10.1126/science.abm4470
Novick D, Last M (2023) Using machine learning models for earthquake magnitude prediction in California, Japan, and Israel. In: international symposium on cyber security, cryptology, and machine learning, pp 151–169 . https://doi.org/10.1007/978-3-031-34671-2_11 . Springer
Noy I, Okubo T, Strobl E, Tveit T (2022) The fiscal costs of earthquakes in Japan. Int Tax Publ Financ. https://doi.org/10.1007/s10797-022-09747-9
Oynakov EI, Botev EA (2021) Spatial and time variations of seismicity before strong earthquakes in the southern part of the Balkans. Ann Geophys 64(4):433–433
Panakkat A, Adeli H (2007) Neural network models for earthquake magnitude prediction using multiple seismicity indicators. Int J Neural Syst 17(01):13–33. https://doi.org/10.1142/s0129065707000890
Panakkat A, Adeli H (2009) Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators. Comput Aided Civ Infrastruct Eng 24(4):280–292. https://doi.org/10.1111/j.1467-8667.2009.00595.x

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features. Adv Neural Inform Process Syst, 31
Pu Y, Chen J, Apel DB (2021) Deep and confident prediction for a laboratory earthquake. Neural Comput Appl 33(18):11691–11701. https://doi.org/10.1007/s00521-021-05872-4
Rafiei MH, Adeli H (2017) NEEWS: a novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. Soil Dyn Earthq Eng 100:417–427. https://doi.org/10.1016/j.soildyn.2017.05.013
Rashidi JN, Ghassemieh M (2023) Predicting the magnitude of injection-induced earthquakes using machine learning techniques. Natural Hazards. https://doi.org/10.1007/s11069-023-06018-6
Reyes J, Morales-Esteban A, Martínez-Álvarez F (2013) Neural networks to predict earthquakes in Chile. Appl Soft Comput 13(2):1314–1328. https://doi.org/10.1016/j.asoc.2012.10.014
Rivière J, Lv Z, Johnson P, Marone C (2018) Evolution of b-value during the seismic cycle: insights from laboratory experiments on simulated faults. Earth Planet Sci Lett 482:407–413. https://doi.org/10.1016/j.epsl.2017.11.036
Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. IEEE Transact Pattern Anal Mach Intell 28(10):1619–1630. https://doi.org/10.1109/TPAMI.2006.211
Sadhukhan B, Chakraborty S, Mukherjee S (2022) Predicting the magnitude of an impending earthquake using deep learning techniques. Earth Sci Inform, pp 1–21 https://doi.org/10.1007/s12145-022-00916-2
Salam MA, Ibrahim L, Abdelminaam DS (2021) Earthquake prediction using hybrid machine learning techniques. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2021.0120578
Schapire RE (1990) The strength of weak learnability. Machine Learn 5(2):197–227. https://doi.org/10.1023/A:1022648800760
Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2009) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Transact Syst Man Cybern Part A Syst Humans 40(1):185–197. https://doi.org/10.1109/TSMCA.2009.2029559
Shi Y, Bolt BA (1982) The standard error of the magnitude-frequency b value. Bull Seismol Soc Am 72(5):1677–1687. https://doi.org/10.1785/bssa0720051677
Shodiq MN, Kusuma DH, Rifqi MG, Barakbah AR, Harsono T (2018) Neural network for earthquake prediction based on automatic clustering in Indonesia. JOIV Int J Inform Vis 2(1):37–43
Tehseen R, Farooq MS, Abid A (2020) Fuzzy expert system for earthquake prediction in western Himalayan range. Elektronika ir Elektrotechnika 26(3):4–12. https://doi.org/10.5755/j01.eie.26.3.25744
USGS (2021) Earthquake catalog. https://earthquake.usgs.gov/earthquakes/search/
Utsu T (1961) A statistical study on the occurrence of aftershocks. Geophys Mag 30:521–605. https://doi.org/10.1007/bf01592930
Waheed U, Afify A, Fehler M, Fulcher B (2020) Winning with simple learning models: detecting earthquakes in Groningen, the Netherlands. In: EAGE 2020 annual conference & exhibition online, 1–5. European association of geoscientists & engineers. arXiv:2007.03924
Wiemer S, Wyss M (1997) Mapping the frequency-magnitude distribution in asperities: an improved technique to calculate recurrence times? J Geophys Res Solid Earth 102(B7):15115–15128. https://doi.org/10.1029/97jb00726
Wyss M, Habermann RE (1988) Precursory seismic quiescence. Pure Appl Geophys 126(2):319–332. https://doi.org/10.1007/bf00879001
Yousefzadeh M, Hosseini SA, Farnaghi M (2021) Spatiotemporally explicit earthquake prediction using deep neural network. Soil Dyn Earthq Eng 144:106663. https://doi.org/10.1016/j.soildyn.2021.106663
Yu S, Ma J (2021) Deep learning for geophysics: Current and future trends. Rev Geophys 59(3):000742. https://doi.org/10.1029/2021rg000742
Zamani A, Sorbi MR, Safavi AA (2013) Application of neural network and ANFIS model for earthquake occurrence in Iran. Earth Sci Inform 6(2):71–85
Zhang L, Si L, Yang H, Hu Y, Qiu J (2019) Precursory pattern based feature extraction techniques for earthquake prediction. IEEE Access 7:30991–31001. https://doi.org/10.1109/ACCESS.2019.2902224

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.