Received: 13 December 2022 | Accepted: 21 May 2023

DOI: 10.1111/1365-2478.13386

Geophysical Prospecting EAGE EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS

ORIGINAL ARTICLE

A comparison of deep and shallow models for the detection of induced seismicity

Akshat Goel[1,2] | **Denise Gorse**[1]

[1]Department of Computer Science, University College London, UK

[2]Rocket Learning – Ekho Foundation, Delhi, India

Correspondence
Denise Gorse, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: d.gorse@cs.ucl.ac.uk

Abstract

Can an interpretable logistic regression model perform comparably to a deep learning model in the task of earthquake detection? In spite of the recent focus in academic seismological research on deep learning, we find there is hope that it can. Using data from the Groningen Gas Field in the Netherlands, relating to low-magnitude induced seismicity, we build on a recently presented four-input logistic regression model by adding to it four further statistically derived features. We evaluate the performance of our feature-enhanced model relative to both the original logistic regression model (shallow machine learning model) and a deep learning model proposed by the same research group. We discover that at the signal-to-noise ratio of this earlier work, our enhanced logistic regression model in fact overall outperforms the deep learning model and displays no false negative errors. At the lower signal-to-noise ratios also considered here, while the number of false positive errors made by the logistic regression model increases, the number of undetected earthquakes remains zero. Though the number of false positives is for the highest imbalance ratios currently prohibitive, the benefit of our four additional features, which increases as the signal-to-noise ratio decreases, suggests that an interpretable model might be made to perform comparably to a more complex deep learning model at real-world class imbalance ratios if further useful inputs could be identified.

KEYWORDS

benchmark study, earthquake detection, feature selection, induced seismicity, machine learning

INTRODUCTION

Machine learning algorithms are being increasingly adopted in a wide range of fields (Shinde & Shah, 2018). Newly available, relatively inexpensive computing power has made it possible to analyse datasets at a scale previously out of reach. The geosciences, too, have been transformed in recent years by a significant growth in the quantity and quality of available data (Bergen et al., 2019), which has spurred interest in machine learning methods for performing seismological tasks with minimal human intervention (Münchmeyer et al., 2022). In addition, the use of machine learning methods able to detect smaller magnitude seismic events than can easily be detected by classical algorithms itself generates yet more labelled data for analysis; this is both a boon and a challenge, as noted in Beroza et al. (2021).

Among available machine learning models, deep neural network models have seen particularly widespread uptake in recent academic work in the earth sciences (Reichstein et al., 2019). These models are appealing because they can be

powerful feature extractors. Given an unstructured dataset like an image collection or a text corpus, they can learn useful and relevant representations of this information to make highly accurate predictions. They are the current state of the art for many prominent benchmark tasks in computer vision, natural language processing, and in other areas (Alzubaidi et al., 2021). Motivated by the success of deep models in other fields, there has been much recent research on the application of deep learning to seismology (see, for instance, Zhu et al., 2022; van der Laat et al., 2021; Soto & Schurr, 2021; Saad et al., 2022). However, the increasing use of deep learning is a double-edged sword. Even though they can be highly accurate, these models are not explainable (i.e., there is no simple human-understandable account as to why the model is making a certain prediction), nor even interpretable (i.e., there is no direct or easy way to understand which input features are causing the model to return a certain output). In a risk-sensitive field such as medical diagnosis or seismology, these aspects of the problem are very important and predictive accuracy is usually only one among many criteria used to evaluate a prediction system (Doshi-Velez & Kim, 2017). In contrast to the recent academic interest in deep learning, seismological agencies around the world still currently use methods which are simple and interpretable by design, have been used for many years, and lend themselves to inspection by specialists for anomalies and errors (NORSAR, 2018).

This contrast between academic research focus and current seismological practice raises the following question: To what extent is the complexity of a deep learning model actually needed for seismological tasks? There is a level of concern in some quarters of the geosciences community (see, e.g., Waheed et al., 2020; Mignan & Broccardo, 2020) that deep learning models (neural networks with many auxiliary 'hidden' layers) are being developed unnecessarily for problems where less complex learning algorithms, such as shallow neural network models (networks with few, or no, hidden layers), have comparable performance and better interpretability or explainability characteristics. A notable recent example of such criticism was in the prediction of earthquake aftershocks, where it was demonstrated in Mignan and Broccardo (2019) that a logistic regression (LR) model with three trainable parameters – in other words, a single neuron with two inputs – performed as well as a 13,451 parameter neural network (DeVries et al., 2018) for this task. Such a case suggests it is possible there may be other seismological problems where the complexity of a deep neural network may be unneeded.

This study considers one such problem, that of the detection of induced seismicity in the Groningen Gas Field, located in the province of the same name in the Netherlands. The detection and characterization of microseismic events using machine learning have become a topic of increasing interest, as evidenced in the recent review of Anikiev et al. (2023). In Waheed et al. (2020), a simple LR model – essentially, a minimally shallow neural network without a hidden layer – with five trainable parameters was used for low-magnitude earthquake detection in data from this area, on the grounds that the interpretability of such a model would highly advantageous. The following year, a paper from the same research group (Shaheen et al., 2021) used a convolutional neural network (CNN), a type of deep learning model introduced originally for image recognition problems, for the same task, using a similar, but not identical, Groningen dataset. The results of these two studies were in apparent contradiction, as Waheed et al. (2020) implied that a simple LR model was more than adequate for this task and that there was no need for a complex CNN model, while Shaheen et al. (2021) seemed to conversely imply that deep learning was required. However, due to methodological and data differences, it was not possible to reach a definite conclusion from a comparison of the results of these works as to which model-simple, feature-based, or complex, based on raw seismic waveform data – is more appropriate.

We aim in this study to make a comparison between these models, the shallow LR model and the deep CNN model, with minor changes as appropriate, and as noted in the relevant sections of this paper, to ensure that the results from the LR and CNN models can be directly compared. We train both our own version of the LR model of Waheed et al. (2020) and an augmented version of this model that uses a further set of interpretable statistical input features. Our work in this paper builds on our earlier work in Gorse and Goel (2022), here re-training the CNN model of Shaheen et al. (2021) for use when the test data are segregated on the basis of event, as in Trani et al. (2020), rather than seismogram, as in Shaheen et al. (2021). Results from the LR models are then compared to results from the CNN model. (These two papers, Waheed et al., 2020, and Shaheen et al., 2021, will from this point onward be referred to, on occasion, as our LR and CNN benchmarks, respectively.)

Notably, in this work, while we train the models with the same high proportion of earthquake examples, relative to non-earthquake, 'noise' examples, as used in Waheed et al. (2020), we additionally challenge the models with test sets in which the proportion of earthquake examples is reduced, with the aim of discovering if the relative strengths of the LR and CNN models are affected by the proportion of earthquakes in the test data. Machine learning models trained on imbalanced data have a tendency to over-assign to the majority class during learning. One common way to address this problem is to train instead on balanced or close-to-balanced data (data, in this instance, in which the number of earthquake examples is close to that of non-earthquake examples). However, it is not always clear that models trained on balanced data will perform well for a test set in which the number of positive cases (here, earthquake examples) is proportionally

smaller, more typical of a natural scenario (in which most 30 s samples of seismic waveforms will not contain a seismic event). The problem potentially posed by imbalanced datasets is widespread in machine learning and would be expected to affect all the models considered in this work. In studying the effects of class imbalance on these models, we measure the degree of data imbalance in terms of imbalance ratio (IR), which is the inverse of the signal-to-noise ratio. IR is used in this work because it is the predominant measure of data class imbalance in machine learning applications.

We discover that, for each imbalance ratio considered, both LR models correctly detect every earthquake, while the CNN does not, and at the initial IR considered in both of our benchmark papers, our best-performing LR model, in fact, outperforms the CNN in relation to accuracy and Matthews correlation coefficient (Matthews, 1975). At higher IRs (proportionally fewer earthquake examples in the test set), the performance of the augmented LR model does deteriorate more rapidly than that of the CNN; however, we will argue that the use of further input features might be able to lift the performance of our interpretable LR model to a level at which it could become competitive with a CNN for practical use.

STUDY CONTEXT AND DATA USED

Induced seismicity in the Groningen Gas Field

The Groningen Gas Field is located in the province of the same name in the northern part of the Netherlands. It is the largest natural gas field in Europe and among the 10 largest in the world. The first significant discovery of gas in this area occurred in 1959 as a result of exploration by Nederlandse Aardolie Maatschappij (NAM), a joint partnership between private firms Shell and Esso. Soon after this discovery, it became clear to NAM that the volume of gas present in the reservoir was unprecedented. Initial estimates from exploration suggested 60 billion cubic metres (bcm), but this was quickly revised upwards to 150 bcm. The latest estimates suggest that both these numbers are significantly downward biased and place the correct estimated gas volume at 2900 bcm, of which 2070 bcm had already been extracted as of 2017 (van de Graaf et al., 2017).

Concerns about induced seismicity associated with gas extraction in the Groningen Gas Field were first publicly voiced in the late 1980s, as evidenced in Vlek (2019). However, as this paper goes on to explain, the watershed event which swayed public sentiment about Groningen gas from positive to negative was a magnitude 3.6 earthquake that occurred near Huizinge in 2012, which caused widespread damage to property, which led to the Dutch government introducing, and since maintaining, annual production caps for Groningen. It is currently planned to cease gas production by 1 October 2023 (Reuters, 2023). It has in addition initiated an enhanced monitoring process via an unusually dense network of seismic detection stations called the G-network (NORSAR, 2018), from which our data are derived.

The G-network

The G-network, which became operational in the Groningen region in 2016, was built upon a pre-existing seismic detection network, now known as the 'old borehole network', initiated in the early 1990s, with the aim of reducing the distance between stations to no more than 5 km in the new network. The sensor configuration of the G-network boreholes is the same as in the old borehole network, with four three-component geophones located at 50 m intervals (at 50, 100, 150 and 200 m) in each borehole. Seventy new borehole stations were set up between 2010 and 2015 (NORSAR, 2018), with Figure 1 showing the geographical extent of this seismic network and demonstrating the high density of stations.

Data from the G-network provide an excellent platform for the study of induced seismicity, these data being used not only in our benchmark papers, and in the current work, but also, for example, in Paolucci et al. (2021). The dense and evenly spaced detection stations of the network allow for granular monitoring of seismic activity at a range of magnitudes, with seismograms from each station being publicly available from a set of web services hosted by the Royal Netherlands Meteorological Institute (KNMI) website (Royal Netherlands Meteorological Institute (KNMI), 1993). In addition, the four-geophone structure of each station in the G-network leads to the possibility of using the moveout pattern as an additional indicator of example type (earthquake or noise), as in Shaheen et al. (2021).

Data sourcing, pre-processing, and partitioning

G-network data were obtained from the KNMI website referenced above, which makes available both raw seismic waveform data and meta-data (the latter allowing, for instance, the identification of the detection station that recorded the signal) for both event and non-event instances. The objectives were to obtain data as close to identical to those used by our convolutional neural network (CNN) benchmark (Shaheen et al., 2021) as was feasible and to partition the data as similarly as possible; where any adaptations needed to be made, on grounds of practical feasibility or good practice in machine learning, these will be noted in the discussion below. Following our CNN benchmark, we downloaded seismograms for a time window of 30 s; in the case of event data the window extended 15 s before and after the P-wave pick, and the case of noise data to 15 s before and after the (randomly) selected

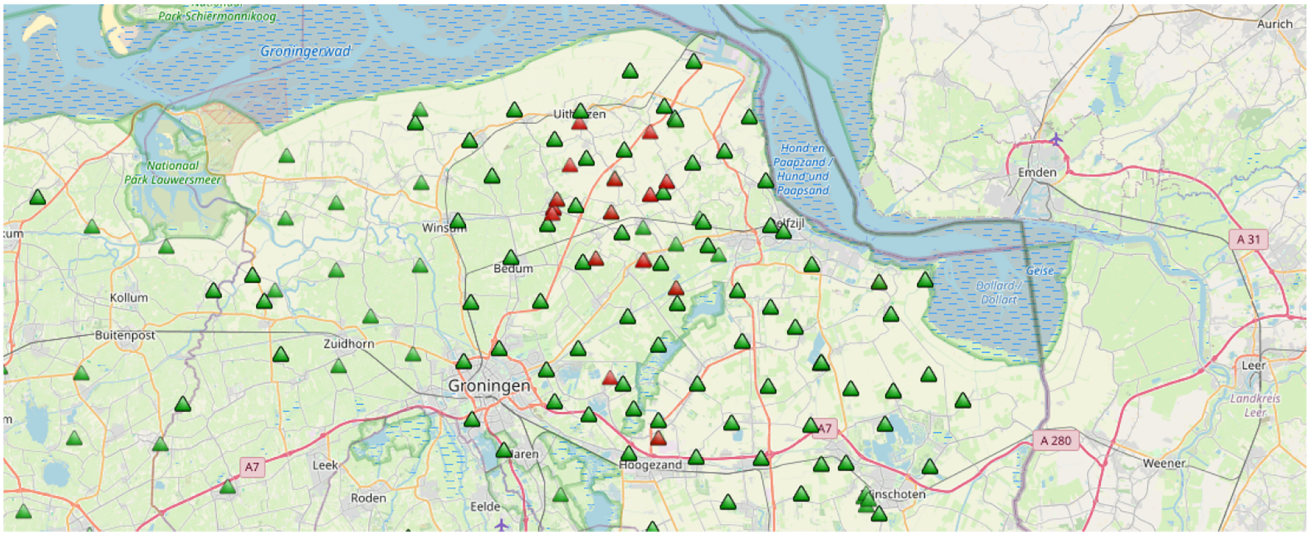

FIGURE 1 The Groningen area with functioning seismic stations shown in green triangles and closed stations shown in red triangles (reproduced from Royal Netherlands Meteorological Institute (KNMI) (1993), with permission).

origin time. All waveform data were detrended, demeaned and bandpass-filtered to frequencies of 5–25 Hz, again as in our CNN benchmark. In addition, in order to be compatible with this benchmark model, the resulting seismograms were then downsampled by a factor of two. Finally, we note that we gathered data from all four geophone levels of the G-network, as our CNN benchmark requires this. In the case of event data, we downloaded data satisfying the following criteria: magnitude ≥ 0.2; origin date between 1 October 2017 and 28 February 2018, inclusive of the end dates; depth between zero and 100 km. This query returned 56 events. However, both of our benchmark papers reference only 47 events during this period. The discrepancy was resolved by noting that nine of the 56 events were located in the Essen and Aarborg regions, considerably south of Groningen, only one of which, the closest to Groningen, was in fact detected by the G-network; these nine events were removed from our dataset. After discarding also seismograms that originated from the old borehole network (since both benchmark papers used data from the G-network stations only), we were left with 2300 positive-labelled seismograms, the same number as in the two comparison studies. We would ideally have picked non-event data from the same 2017–2018 period as event data. However, KNMI has limited retention of noise data, so the required number of 4000 non-event examples was obtained from the corresponding time period in 2020–2021. For those later experiments which examine the effect of the imbalance ratio on model performance, it was necessary to augment the test datasets with further noise examples. Additional negative examples required for the increasing noise-to-signal phase of this study were drawn from this same period, with 38,100 new samples collected.

We divide our data into the train (60%), validation (20%) and test (20%) sets, as was done in our CNN benchmark (Shaheen et al., 2021), though not in our logistic regression benchmark (Waheed et al., 2020), which latter had only a 70:30 train: test split. We split our data by event rather than seismogram (as was done in both benchmark papers) in order to prevent data leakage from test to train datasets. Event-based partitioning was also used in the KNMI study of Trani et al. (2020). Hence, for example, our test set may not contain 20% of the seismograms, as events are detected by varying numbers of G-network stations; however, this difference, as with the differing time period from which noise data were derived, is not problematic for the internal comparisons within this study, noting that in the current work the CNN is retrained specifically for use with event-segregated data.

LEARNING MODELS AND PERFORMANCE MEASUREMENT

Learning models

Logistic regression

The primary model used in this study is logistic regression (LR), as used in the first of our benchmark papers (Waheed et al., 2020). The major appeal of this linear model is that its fitted weights can be used to understand the role and importance of different features. An elastic net penalty

$$\lambda[\alpha \sum_{j=0}^{p} \left|\beta_j\right| + (1-\alpha) \sum_{j=0}^{p} \beta_j^2], \quad (1)$$

where the β_j are the $(p + 1)$ weights of the model, may be added to the LR loss function. The sum of weight magnitudes $\sum_{j=0}^{p} |\beta_j|$ in the above is known as the LASSO penalty. It encourages a sparse solution or, in other words, variable selection. The sum of squared weight values $\sum_{j=0}^{p} \beta_j^2$ is known as the ridge penalty. It functions to average the coefficients of highly correlated features but does not drive coefficient values to zero, so does not perform variable selection. The parameter α (which determines the balance between LASSO and ridge penalties) and the regularization parameter λ can be chosen via grid search on a validation dataset. It should be noted that in our work an elastic net penalty was used only in the preliminary feature selection phase, as a means to filter new candidate features according to their importance, our final model being a simple LR model as in Waheed et al. (2020).

Convolutional neural network

As stated previously, the major objective of this paper is to benchmark an interpretable LR model (based on that of Waheed et al., 2020, with a number of additional features, to be described below, that enhance its performance) against a convolutional neural network (CNN) devised for use on the same Groningen dataset. CNNs, a type of deep learning network based on the operation of mammalian visual systems, were first introduced in LeCun et al. (1998) and have since become the dominant paradigm for deep learning (Alzubaidi et al., 2021). They have become increasingly popular for seismic waveform analysis, being used, for example, in the work of Mousavi et al. (2020) and Zhu et al. (2022), and much has been claimed for the effectiveness of these models. However, in terms of interpretability, being deep learning models of substantial complexity they represent the polar opposite of LR models. For the safety-critical area of earthquake detection, it would therefore be reasonable to require concrete evidence that such complex models were the only feasible option for this task.

The CNN architecture used in this work was obtained from the authors of Shaheen et al. (2021). It was designed to take advantage of the multiple geophone levels used in the G-network, leveraging the potential of the moveout pattern of energy to distinguish between disturbances originating underground (more likely to be a seismic event) and ones originating at the surface (more likely to be noise). Because, as explained earlier, our datasets are not identical to those of our CNN benchmark (differing period for noise data extraction; use of event-based, rather than seismogram-based, stratification), this CNN was fully retrained for our purposes. The shape of the array input to the CNN is (4, 3001, 3), corresponding to four geophones, 3001 time points (after preprocessing), and three channels per geophone, respectively, with the CNN architecture as given in Shaheen et al. (2021). In addition, we used the same means of initialization of weights, the same optimizer and the same learning rate as in our CNN benchmark, also.

Performance measurement

One key sense in which the work of this paper differs from many other studies in academic seismology is that our models are challenged on test data with progressively higher imbalance ratios (IRs), representing more natural ratios of noise signals to earthquake signals, in order to determine the effect of IR on model performance. Assessing model performance in situations of high-class imbalance requires caution. Accuracy (the proportion of correct classifications relative to the total number of examples), despite still being in wide use in such situations, can be misleading: for example, if 90% of examples are negative, an accuracy of 90% can be achieved by assigning all examples to this majority class, despite the resulting model being entirely useless as a classifier. In our work, we quote accuracy due to its continuing wide use as a performance measure, independently of IR, but regard the Matthews correlation coefficient (MCC) (Matthews, 1975) as our primary performance measure, due to its robustness in situations of class imbalance (Chicco & Jurman, 2020). The MCC is defined by

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / [(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}))]^{\frac{1}{2}}, \quad (2)$$

in which TP is the number of true positives (earthquake examples correctly classified as such), TN is the number of true negatives (noise examples correctly classified as such), FP is the number of false positives (noise examples wrongly classified as earthquakes), and FN the number of false negatives (earthquake examples wrongly classified as noise). The MCC takes values between +1 and −1. A value of +1 indicates a perfect classifier, while an MCC of −1 indicates a classifier which predicts every example to be of the opposite class. An MCC of 0 indicates a classifier which performs no better than random or which, importantly for the case of imbalanced datasets, wrongly categorizes all examples as being of the majority class. In the example used in the discussion of accuracy, where a deceptively high accuracy of 90% could be obtained by assigning all examples to the majority class, the lack of utility of the model would be revealed in its MCC of 0.

FEATURE CONSTRUCTION

We derived our input features (listed below, with their names within the relevant packages, together with their designations

within this work) from two sources. The means of selection are described in the subsections below.

Initial choice of input features

Two sources of potential features were used, described below.

Highly comparative time series analysis features as used in logistic regression benchmark

Highly comparative time series analysis (HCTSA) (Fulcher & Jones, 2014) is a package that derives, for a given time series, up to 7700 statistical features that are known to perform well as descriptors in a wide range of domain areas. Using single (Z-) channel seismogram data, Waheed et al. (2020) used HCTSA's inbuilt features to first create a list of 50 high-performing features, then used HCTSA's correlation matrix functionality to choose the four features from this list that were closest to being uncorrelated with each other while at the same time separating the data well. These features, denoted here W1–W4, are among the final eight features used here, as all four were later found to be valuable in classification using the elastic net selection process. They have the following definitions, taken from the HCTSA documentation (Fulcher & Jones, 2014):

- DN_RemovePoints_min_05_fzcacrat (W1): It measures how time-series properties change as points are removed. Specifically, it computes the first zero-crossing of the normal linear autocorrelation function as 50% of the lowest values are removed.
- SY_SlidingWindow_s_s_5_1 (W2): Sliding window measure of stationarity. Specifically, it divides the time series into five windows and computes the standard deviation in each window followed by computing the standard deviation of the resulting five standard deviation values.
- ST_MomentCorr_002_02_mean_std_sqrt_mi (W3): It measures correlations between simple statistics in local windows of a time series. Specifically, it computes mutual information between two vectors formed by computing the mean and standard deviation of the time series in a sliding window. The length of the sliding window is 2% of the entire time series with a 20% overlap between consecutive windows.
- FC_Surprise_dist_100_5_q_500_tstat (W4): It measures the level of surprise due to the next data point given recent memory. Specifically, it coarse-grains the time series into five groups and computes a summary of information gain with 100 previous memory samples.

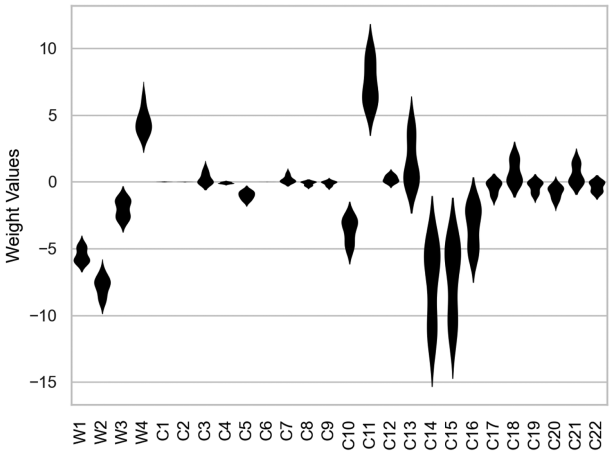

FIGURE 2 Distributions of weight values for the 72 best-performing LR models with an elastic net penalty.

Additional features from catch22

The catch22 MATLAB package (Lubba et al., 2019) was created by the authors of HCTSA as a computationally efficient package that uses only the 22 HCTSA features discovered to be 'best performing' over a wide range of different time series. It was found that none of the features in our LR benchmark were included in the catch22 set. This could be because their means of selection of the four HCTSA features substantially differed from ours, as noted in the conclusion of this work, when discussing the possibility of adding further input features to the LR model.

Feature selection using the elastic net

The combination of HCTSA, as used in Waheed et al. (2020), with catch22 thus provided us with four previously used features (from HCTSA), and 22 new, and potentially high-performing, ones (from catch22). We then used an LR model with an elastic penalty, as described in the section on learning models, to select the most important among these 26 features, using a grid search to select the λ and α parameters of Equation (1). We discovered many (72) models with identical performance in terms of the validation of the Matthews correlation coefficient (MCC), with the distribution of weight values for each feature for these models being plotted in Figure 2.

This figure confirms the value of the original four HCTSA features (denoted here W1–W4), and four of the catch22 features (denoted C10, C11, C14, and C15), with C11 appearing especially promising. These features have the following definitions, taken from the catch22 documentation (Lubba et al., 2019):

- PD_PeriodicityWang_th0_01 (C10): Time intervals between successive extreme events above the mean.
- CO_Embed2_Dist_tau_d_expfit_meandiff (C11): Time intervals between successive extreme events below the mean.
- DN_OutlierInclude_p_001_mdrmd (C14): Exponential fit to successive distances in two-dimensional (2D) embedding space.
- DN_OutlierInclude_n_001_mdrmd (C15): Periodicity measure of Wang et al. (2007).

These four new features were, therefore, added to the HCTSA-selected group from our LR benchmark. The 18 less-influential catch22 features, and the use of the elastic net penalty, were then discarded in order to have a simple, more easily interpretable LR model.

As examples of the interpretation of these features, we first consider, from the HCTSA set, the feature W2, which from Figure 2 appears the most potentially useful of this set, having its entire distribution of values most clearly separated from zero. It is calculated within HCTSA using a sliding window from which a mean and standard deviation of amplitude can be derived, and in this context may reflect that, while the mean remains constant, there is a sudden increase in the variance after the P-wave onset.

As a second example, from the catch22 set, we consider the feature C11, which from Figure 2 appears the most potentially useful of this set. This feature is the exponential fit to successive distances in a 2D embedding space and is constructed as follows:

- set a particular window size τ (we use the catch22 default value of 30); each point in the embedding space then becomes $X_t = (x_t, x_t + \tau)$,
- calculate 2D Euclidean distances between successive points so constructed to yield d_t,
- fit a 2D exponential distribution to these distances, and
- calculate the deviation from this fit and use this as the feature.

The plots in Figures 3 and 4 show the 2D distances from the above-calculated feature for a randomly picked negative (noise) example and a randomly picked positive (earthquake) example from the test set, respectively. The value of the calculated feature derived from these distances is likely to be very different in the negative and positive cases.

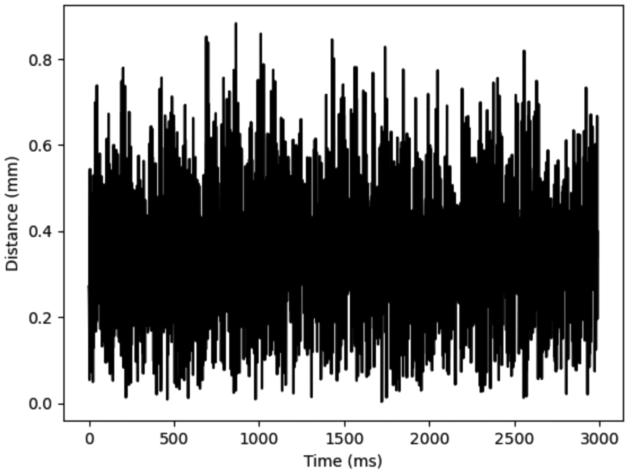

FIGURE 3 Feature C11: successive distances in 2D embedding space (see the text for details) for a randomly selected negative example from the training data.

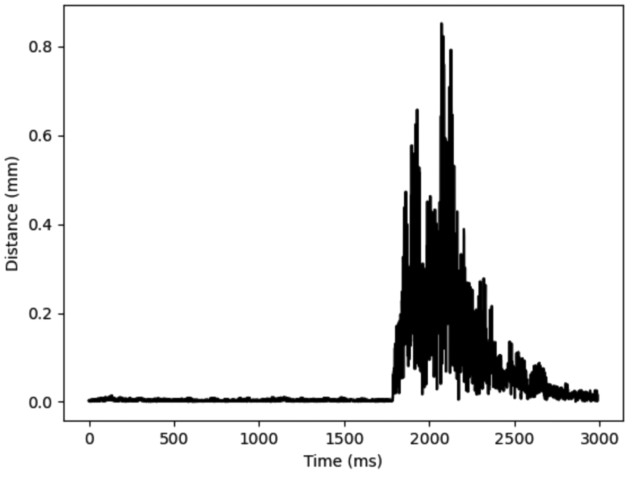

FIGURE 4 Feature C11: successive distances in 2D embedding space (see the text for details) for a randomly selected positive example from the training data.

RESULTS

Data exploration

Table 1 shows descriptive statistics for the 47 seismic events considered in the benchmark work to which we compare our models and also in this current work. For each variable tabulated, the mean is different from the median. This suggests that these variables are unlikely to be distributed according to a symmetric distribution, and in particular are unlikely to be normally distributed, which is also evident in the plots to follow.

Figure 5 (left) shows the frequency distribution of the number of stations used in event detection. There are 77 stations in the G-network. A large fraction of events were detected by less than 10 stations. A small fraction of events were detected by between 40 and 50 stations. Since most of the events are of small magnitude, it would be expected that they would predominantly be detected by a limited number of nearby stations. Figure 5 (right) shows the relationship between the number of stations used in event detection and the estimated magnitude of an event. The best fit line is upward

TABLE 1 Descriptive statistics related to event detections.

	Mean	Std.	Min.	25%	50%	75%	Max.
Number of stations used in detection	16.28	14.48	2	5	10	23.5	52
Magnitude	1.05	0.60	0.22	0.67	0.92	1.35	3.43
Distance to the nearest station (km)	1.61	1.31	0.13	0.99	1.37	1.97	8.94
Distance to the farthest station (km)	23.82	18.85	8.65	14.86	20.41	26.11	107.06

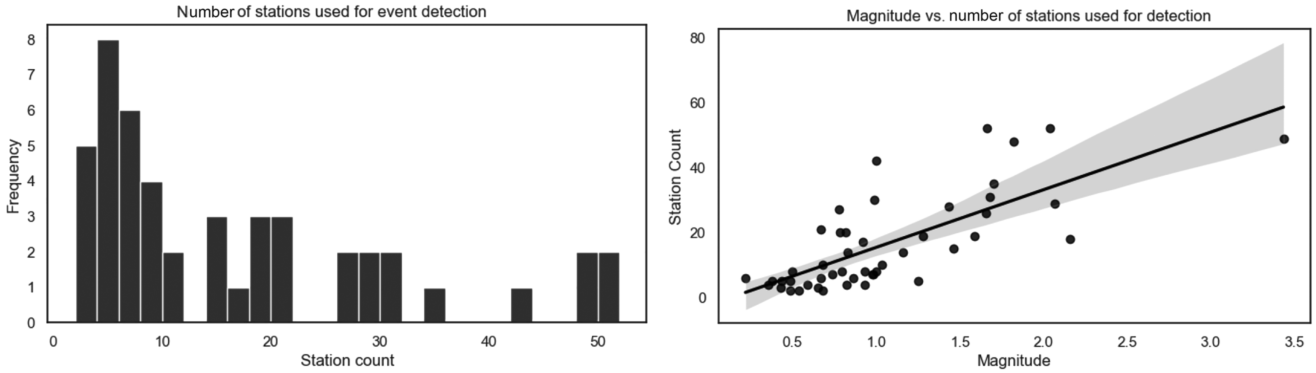

FIGURE 5 Distribution of station counts in event detection (left) and their relationship with event magnitude (right).

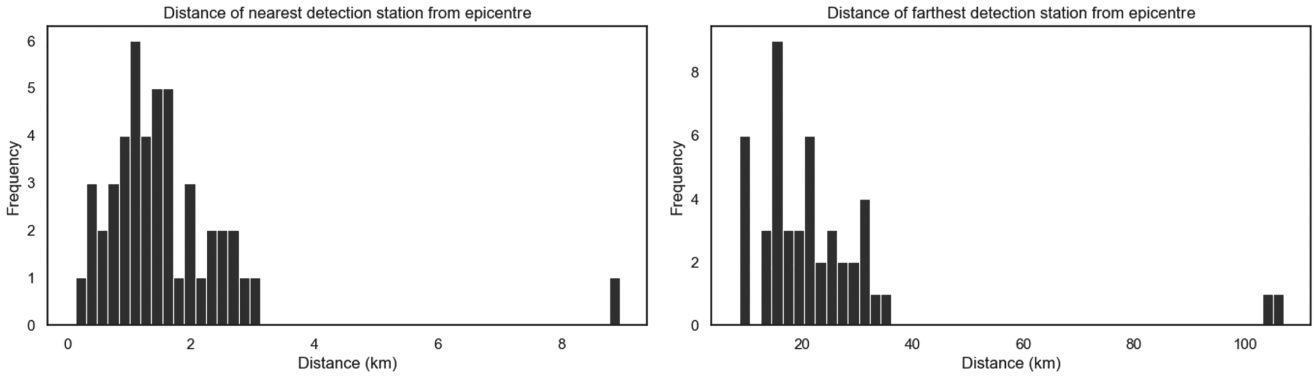

FIGURE 6 Distributions of distances from an event epicentre of the nearest (left) and furthest (right) stations used in detection.

sloping, as we would expect: higher magnitude events are detected by more stations. The minimum distance frequency distribution in Figure 6 (left) suggests that most events are detected by at least one station which is at most 3 km away from its epicentre, while the corresponding maximum distance plot in Figure 6 (right) suggests that the majority of events are detected by stations at most 40 km away from their epicentre.

Logistic regression experiments at baseline imbalance ratio

These first experiments, summarized in Table 2, aim to elucidate the value of the four selected catch22 features. The experiments are carried out at the same noise-to-signal ratio (IR) as our benchmark logistic regression (LR) model (Waheed et al., 2020), namely 1.73:1, in order to facilitate a direct comparison with this earlier work. However, our baseline model makes minor adjustments to this model, as noted in the section on data sourcing and pre-processing. We use a validation set for both LR models, to enable hyperparameter optimization and early stopping, in order to prevent overfitting. Our primary performance measure is the Matthews

TABLE 2 Logistic regression model test results at IR of 1.73:1.

Model description	Test accuracy (%)	Test MCC
Baseline (LR)	98.99	0.9786
Baseline + selected catch22 (LR+)	99.76	0.9948

Abbreviations: LR, logistic regression; MCC, Matthews correlation coefficient.

TABLE 3 LR and CNN test results at IR of 1.73:1.

Accuracy (%)			MCC		
LR	LR+	CNN	LR	LR+	CNN
98.99	99.76	99.61	0.9786	0.9948	0.9917

Abbreviations: LR, logistic regression; LR+, baseline (LR) + selected catch22; CCN, convolutional neural network.

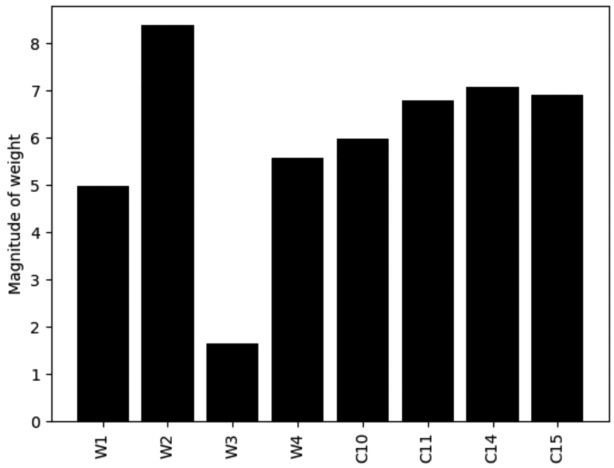

FIGURE 7 Feature importance plot from LR+ model.

correlation coefficient (MCC), for reasons explained in the performance measurement section. However, we also report test accuracy, due to its wide use in results reporting, and in our LR benchmark; again, as explained in this earlier section. Results are assessed for significance testing using a two-sided *t*-test at the 5% threshold.

It is evident that both the test accuracy and test MCC of the baseline model are very high. High-accuracy values were also obtained in both of our benchmark papers; they are in part due to the relative ease of detecting induced seismicity in a very densely configured network such as the G-network. It is noteworthy (as evidenced in the confusion matrices of Table 5) that neither the LR or LR+ models have any false negatives (undetected earthquakes). However, comparing the LR and LR+ models, there is even so a statistically significant benefit from the addition of the new features from catch22; this is associated with a decrease in the number of false positives (sections of seismic waveform data wrongly labelled as containing a seismic event). The proportional benefit of the new catch22 features in this respect will be seen to increase with increasing IR, as will be shown later in this results section.

The average weight magnitudes (evidencing the importance of the features to the resulting LR models) associated with all eight external inputs are shown in Figure 7, in which W1–W4 are the HCTSA feature weights and C10, C11, C14 and C15 the catch22 feature weights. It is evident that the four new catch22 features had a substantial impact on the augmented model's decision-making; all four are more highly weighted than all but one of the original HCTSA features. Nonetheless, three out of the four HCTSA features are clearly also important, and W2 highly so (in fact the most highly weighted feature overall), the only HCTSA feature of possibly limited importance being W3.

Comparison to convolutional neural network model at baseline imbalance ratio

Our feature-augmented LR model, LR+, described above, was then compared, on the same test dataset, to results from the model of Shaheen et al. (2021), our convolutional neural network (CNN) benchmark. Our objective was to discover whether the additional complexity of a CNN model was truly necessary for this task. The CNN model architecture had been made available to us by the authors of Shaheen et al. (2021), and, as explained earlier, was retrained on the training data. Table 3 shows the test performance of the LR+ model compared to that of the CNN model and the LR model. The CNN, as might be expected on the basis of the logistic regression results of Table 2, performed very well on this dataset. However, it was not the best-performing model; despite the CNN's very substantial additional complexity and opportunity to benefit from the use of the moveout pattern, the LR+ model in fact had a statistically significantly higher test MCC and accuracy. Moreover, as previously noted, neither of the LR models displayed any false negatives on the test dataset.

Effect of increasing IR on the performance of the models

As we noted in the Introduction, academic studies in earthquake detection usually train machine learning models on datasets with low imbalance ratios, with the hope, but with often limited evidence, that the trained models will perform equally well when tested on higher ratios (i.e., proportionally fewer earthquake examples) or in continuous tests. Table 4 shows, for the LR, LR+, and CNN models, results for expanded test datasets with higher IRs (constructed by oversampling noise data, as outlined in the section on data selection and preparation). Tables 5 and 6, in addition, show the confusion matrices, for all models considered, in the cases of the IRs 1.73:1 and 50:1, respectively. As would be expected, the test accuracies and MCCs drop, for all models, with increasing IR. This is seen more noticeably for the logistic regression models. However, the logistic regression models continue to have zero false negatives. It is notable, also, that the benefit of the additional catch22 features is progressively more evident at higher IRs, these four extra features, taking the total number of inputs to the LR models from

TABLE 4 LR and CNN test results at progressively higher IRs, where the columns headed 'LR+/LR' give the relative benefit, at each IR, of adding the four extra catch22 features to the logistic regression model.

	Accuracy (%)				MCC			
IR	**LR**	**LR+**	**LR+/LR**	**CNN**	**LR**	**LR+**	**LR+/LR**	**CNN**
1.73:1	98.99	99.76	1.008	99.61	0.9786	0.9948	1.017	0.9917
5:1	91.27	93.77	1.027	99.63	0.7663	0.8203	1.070	0.9866
10:1	71.33	81.70	1.146	99.69	0.4062	0.515	1.268	0.9814
25:1	62.59	75.27	1.203	99.69	0.2388	0.3164	1.325	0.9595
50:1	59.33	73.19	1.234	99.69	0.1642	0.2227	1.356	0.9265

Abbreviations: CCN, convolutional neural network; IR, imbalance ratio; LR, logistic regression; LR+, baseline (LR) + selected catch22.

TABLE 5 Confusion matrices at IR 1.73:1 for the LR, LR+, and CNN models.

	LR		LR+		CNN	
	Predicted negative	**Predicted positive**	**Predicted negative**	**Predicted positive**	**Predicted negative**	**Predicted positive**
True negative	1297	21	1313	5	1313	5
True positive	0	763	0	763	3	760

Abbreviations: CCN, convolutional neural network; LR, logistic regression; LR+, baseline + selected catch22.

TABLE 6 Confusion matrices at IR 50:1 for the LR, LR+, and CNN models.

	LR		LR+		CNN	
	Predicted negative	**Predicted positive**	**Predicted negative**	**Predicted positive**	**Predicted negative**	**Predicted positive**
True negative	22,295	15,805	27,681	1,0419	37,981	119
True positive	0	763	0	763	3	760

Abbreviations: CCN, convolutional neural network; IR, imbalance ratio; LR, logistic regression; LR+, baseline (LR) + selected catch22.

four to eight, leading to a 36% improvement in MCC for our LR+ model compared to the LR model, re-implemented from Waheed et al. (2020).

DISCUSSION AND CONCLUSIONS

In this study, we asked whether a logistic regression (LR) model with interpretable features can perform as well as a convolutional neural network (CNN) in detecting low-magnitude earthquakes in the Groningen Gas Field in the Netherlands. This question is important because there has been a recent move in academic seismology, as in many other fields, towards the use of complex deep learning models whose workings are impossible for a human analyst to inspect and understand in a simple way. However, in risk-sensitive settings such as earthquake detection, interpretability is a highly desirable model characteristic, making a benchmark study such as this valuable and timely.

In the first stage of our work, we replicated as closely as possible the procedures in our LR benchmark (Waheed et al., 2020), which used the same Groningen event dataset. We used an LR model trained with the same four features from the highly comparative time series analysis (HCTSA) package (Fulcher & Jones, 2014) that were used in our LR benchmark, though we adjusted this model in order to make the treatment of the data (e.g., length of time window) compatible with the model of Shaheen et al. (2021), our CNN benchmark, to which our LR results would later be compared. In the second stage, we improved on this initial LR model by the addition of four further interpretable features from the catch22 package (Lubba et al., 2019), selecting these features via a preliminary elastic net modelling phase. Finally, in the third stage, we benchmarked both of our LR models against the retrained CNN model, first at the imbalance ratio (IR) of 1.73:1 used in both of our benchmark papers, Waheed et al. (2020) and Shaheen et al. (2021), and then at progressively higher ratios, moving towards ones more typical of a natural setting (i.e.,

one in which, according to the labelling provided by (Royal Netherlands Meteorological Institute (KNMI), there are proportionally very few 30 s windows that contain a seismic event).

On the 1.73:1 test data, we discovered that our feature-augmented LR model (LR+) was surprisingly statistically significantly more effective, in relation to both accuracy and Matthews correlation coefficient (MCC), than the far more complex CNN model. The LR+ model (as did the four-input LR model) additionally had zero false negatives, that is, earthquake events incorrectly classified as noise, on this dataset. It can be strongly argued that false negatives are less tolerable than false positives in seismological applications not only because an undetected earthquake is liable to have more negative consequences than a false alarm but also because it is standard practice within seismological agencies such as KNMI to reduce false alarms by the manual review of all event detections (NORSAR, 2018). The number of false positives, therefore, needs only be rendered manageable, not necessarily reduced to zero.

When moving to data with larger proportions of noise (higher IRs), the performance of both LR models decreased with the amount of noise, though the LR models continued to have zero false negatives. The performance of the CNN did decrease substantially less rapidly with IR, compared to the LR models. However, the CNN had 283,700 free parameters, as opposed to the nine of the LR+ model. Furthermore, in relation to the LR models, we discovered that the proportional benefit of the additional catch22 features increased with IR; these four extra parameters were able to boost the MCC of the LR+ model, at an IR of 50:1, by 36% compared to the MCC of the baseline LR model.

That such a small number of extra features could have such a large proportional benefit, and moreover one that was observed to increase with IR, suggests that the use of further input features, either statistical or seismological in nature, might allow the creation of an interpretable LR model with a performance comparable to that of a CNN not only at a low IR, such as the initial 1.73:1 considered in this work and in both of our benchmark papers (where our LR+ model in fact outperformed the CNN), but at IRs more typical of a real-world scenario. We note here the distinction between the concepts of interpretability and explainability. The former requires only that the degree of influence of each input feature on the output of a model is readily apparent. The latter requires also that the means by which each feature affects the output is understandable in lay, or at least domain expert, terms. New features with a seismological origin would have an explainable influence. However, statistically motivated features derived from packages like HCTSA and catch22, in general, would not. It is for this reason we would term any linear model that included features of this latter type to be interpretable rather than explainable.

Considering the first additional statistically motivated features, we note that Waheed et al. (2020) selected their four HCTSA features using a substantially different method to the elastic net feature selection method used here; this may be the reason why their selected features did not, for example, include any of the four catch22 features we additionally used in the LR+ model, despite these being highly ranked in the feature importance plot of Figure 7. One approach that could be taken in further work might be to begin with all (around 7700) of the features computed by the HCTSA package and use a feature selection tool such as Minimum Redundancy Maximum Relevance (mRMR) (Ding & Peng, 2005) to choose the most relevant features. mRMR was developed initially for use in bioinformatics applications (Ding & Peng, 2005) but has since been used more widely, notably in Zhao et al. (2019), and including, within the geosciences, in an earthquake prediction model proposed by Asim et al. (2018).

Turning to domain-specific input features, Miranda et al. (2019) used three-dimensional measurements of the degree of polarization and vertical power radius against total power (RV2T) within an LR ensemble model which used data from four Colombian triaxial seismological stations, achieving 95% accuracy in the detection of seismic events. A variety of other domain-motivated features have been used within models for earthquake signal detection, for example, in Kaur et al. (2013), Vallejos and McKinnon (2013), Lindenbaum et al. (2016), and Reynen & Audet (2017), and there is clearly scope for the exploration of the use of such features alongside ones from statistical toolkits such as HCTSA. It would additionally of interest to compare the LR+ and CNN models on similar data from other regions, for example, to use the Oklahoma, USA, dataset from Reynen & Audet (2017).

We certainly do not claim there is no place in seismology for deep learning. Yet shallow models, such as LR, are of indisputable appeal due to their simplicity and transparency. We believe, on the basis of the results presented here for the Groningen Gas Field dataset, where our LR model was benchmarked against a far more complex CNN model for the same task, that it is worth further exploring the potential utility of LR models for this and for other seismic datasets, ones relating to both induced and tectonic source earthquakes.

ACKNOWLEDGEMENTS

The authors would like to thank Umair bin Waheed for helpful advice, and Ahmed Shaheen for the provision of the CNN model used after retraining as a deep learning benchmark.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available upon request from the authors.

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. & Farhan, L. (2021) Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 1–74.

Anikiev, D., Birnie, C., bin Waheed, U., Alkhalifah, T., Gu, C., Verschuur, D.J. & Eisner, L. (2023) Machine learning in microseismic monitoring. *Earth-Science Reviews*, 239, 104371.

Asim, K.M., Idris, A., Iqbal, T. & Martínez-Álvarez, F. (2018) Earthquake prediction model using support vector regressor and hybrid neural networks. *PLoS ONE*, 13(7), e0199004.

Bergen, K.J., Johnson, P.A., de Hoop, M.V. & Beroza, G. (2019) Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433), eaau0323.

Beroza, G.C., Segou, M. & Mousavi, S.M. (2021) Machine learning and earthquake forecasting-next steps. *Nature Communications*, 12(1), 1–13.

Chicco, D. & Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 1–13.

DeVries, P. M.R., Viégas, F., Wattenberg, M. & Meade, B.J. (2018) Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720), 632–634.

Ding, C. & Peng, H. (2005) Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2), 185–205.

Doshi-Velez, F. & Kim, B. (2017) Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv:1702.08608

Fulcher, B. & Jones, N. (2014) Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26, 3026–3037.

Gorse, D. & Goel, A. (2022) Deep vs. shallow learning: a benchmark study in low magnitude earthquake detection. In *83rd EAGE Annual Conference & Exhibition*. Houten, the Netherlands: European Association of Geoscientists & Engineers, pp. 1–5.

Kaur, K., Wadhwa, M. & Park, E. (2013) Detection and identification of seismic P-waves using artificial neural networks. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*. Piscataway, NJ: IEEE, pp. 1–6.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Lindenbaum, O., Rabin, N., Bregman, Y. & Averbuch, A. (2016) Multichannel fusion for seismic event detection and classification. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*. Piscataway, NJ: IEEE, pp. 1–5.

Lubba, C.H., Sethi, S., Knaute, P., Schultz, S., Fulcher, B. & Jones, N. (2019) catch22: CAnonical Time-series CHaracteristics selected through highly comparative time-series analysis. *Data Mining and Knowledge Discovery*, 33, 1821–1852.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442–451.

Mignan, A. & Broccardo, M. (2019) One neuron versus deep learning in aftershock prediction. *Nature*, 574(7776), E1–E3.

Mignan, A. & Broccardo, M. (2020) Neural network applications in earthquake prediction (1994-2019): meta-analytic and statistical insights on their limitations. *Seismological Research Letters*, 91(4), 2330–2342.

Miranda, J.D., Gamboa, C.A., 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)*. Piscataway, NJ: IEEE, pp. 1–5.

Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T. & et al. T.D,. (2022) Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, 127(1), e2021JB023499.

Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y. & Beroza, G.C. (2020) Earthquake transformer: an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1), 3952.

NORSAR (2018) Induced earthquake catalogue review. Technical report, Koninklijk Nederlands Meteorologisch Instituut. Accessed: 2 June 2021. https://kemprogramma.nl/file/download/18f4a605-4bbc-401a-b68e-8bd840b3d05b/1562832346kem11%20norsar_sodm_groningenreview_wp1.pdf

Paolucci, R., Mazzieri, I., Piunno, G., Smerzini, C., Vanini, M. & Özcebe, A.G. (2021) Earthquake ground motion modeling of induced seismicity in the Groningen gas field. *Earthquake Engineering & Structural Dynamics*, 50(1), 135–154.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. & Carvalhais, N. (2019) Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204.

Reuters (2023) Netherlands sticks to plan to close Groningen gas field by October. Accessed: 19 February 2023. https://www.reuters.com/markets/commodities/netherlands-sticks-plan-close-groningen-gas-field-by-october-ft-2023-01-22/

Reynen, A. & Audet, P. (2017) Supervised machine learning on a network scale: Application to seismic event classification and detection. *Geophysical Journal International*, 210(3), 1394–1409.

Royal Netherlands Meteorological Institute (KNMI) (1993) Netherlands Seismic and Acoustic Network. Accessed: 9 September 2022. https://doi.org/10.21944/e970fd34-23b9-3411-b366-e4f72877d2c5

Saad, O.M., Huang, G., Chen, Y., Savvaidis, A., Fomel, S., Pham, N. & Chen, Y. (2022) SCALODEEP: A highly generalized deep learning framework for real-time earthquake detection. *Journal of Geophysical Research: Solid Earth*, 126(4), e2020JB021473.

Shaheen, A., bin Waheed, U., Fehler, M., Sokol, L. & Hanafy, S. (2021) GroningenNet: Deep learning for low-magnitude earthquake detection on a multi-level sensor network. *Sensors*, 21, 8080.

Shinde, P.P. & Shah, S. (2018) A review of machine learning and deep learning applications. In *Proceedings of Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. Piscataway, NJ: IEEE, pp. 1–6.

Soto, H. & Schurr, B. (2021) DeepPhasePick: a method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks. *Geophysical Journal International*, 227(2), 1268–1294.

Trani, L., Pagani, G.A., Pereira Zanetti, J.P., Chapeland, C. & Evers, L.G. (2020) DeepQuake: An application of CNN for seismo-acoustic event classification in The Netherlands. *Earth and Space Science Open Archive*, 12.

Vallejos, J. & McKinnon, S. (2013) Logistic regression and neural network classification of seismic records. *International Journal of Rock Mechanics and Mining Sciences*, 62, 86–95.

van de Graaf, W.E., van Geuns, L. & Boersma, T. (2017) The termination of Groningen gas production: background and next steps. Accessed: 2 October 2022. https://energypolicy.columbia.edu/sites/default/files/pictures/CGEP_Groningen-Commentary_072518_0.pdf

van der Laat, L., Baldares, R.J., Chaves, E.J. & Meneses, E. (2021) OKSP: a novel deep learning automatic event detection pipeline for seismic monitoring in Costa Rica. In *Proceedings of IEEE 3rd International Conference on BioInspired Processing (BIP))*. Piscataway, NJ: IEEE, pp. 1–6.

Vlek, C. (2019) Rise and reduction of induced earthquakes in the Groningen gas field, 1991–2018: statistical trends, social impacts, and policy change. *Environmental Earth Sciences*, 78, 59. https://doi.org/10.1007/s12665-019-8051-4

Waheed, U.b., Shaheen, A., Fehler, M. & Fulcher, B. (2020) Winning with simple learning models: Detecting earthquakes in Groningen, the Netherlands. In *82nd EAGE Annual Conference & Exhibition*. Houten, the Netherlands: European Association of Geoscientists & Engineers, pp. 1–5.

Wang, X., Wirth, A. & Wang, L. (2007) Structure-based statistical features and multivariate time series clustering. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. Piscataway, NJ: IEEE, pp. 351–360.

Zhao, Z., Anand, R. & Wang, M. (2019) Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Piscataway, NJ: IEEE, pp. 442–452.

Zhu, W., Tai, K.S., Mousavi, S.M., Bailis, P. & Beroza, G.C. (2022) An end-to-end earthquake detection method for joint phase picking and association using deep learning. *Journal of Geophysical Research: Solid Earth*, 127(3), e2021JB023283.

How to cite this article: Goel, A., & Gorse, D. (2023) A comparison of deep and shallow models for the detection of induced seismicity. *Geophysical Prospecting*, 1–13. https://doi.org/10.1111/1365-2478.13386