This is a peer-reviewed, accepted author manuscript of the following conference paper: Jiang, J, Stankovic, V, Stankovic, L, Murray, D & Pytharouli, S 2024, Explainable AI for transparent seismic signal classification. in 2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE International Symposium on Geoscience and Remote Sensing (IGARSS), IEEE, Piscataway, NJ.

EXPLAINABLE AI FOR TRANSPARENT SEISMIC SIGNAL CLASSIFICATION

Jiaxin Jiang, Vladimir Stankovic, Lina Stankovic, David Murray, Stella Pytharouli

University of Strathclyde
Glasgow, United Kingdom

ABSTRACT

Deep learning has found extensive applications in classifying seismic signals in recent years. However, as a black box algorithm, deep learning is still rarely exploited in real-world applications, such as landslide monitoring. This is particularly a concern for geoscientists who prefer to classify seismic signals based on their physical properties, through feature engineering. To build trust in deep learning model outputs, we propose a CNN multi-classifier architecture to classify seismic signals into four classes (earthquake, micro-quake, rockfall and noise), and explain its outputs based on Layer-wise Relevance Propagation. We demonstrate that the provided explanations can lead to a more interpretable model by relating network outputs to geophysical phenomena and showing that distinguishing features extracted by the network are aligned with those identified by geoscientists as pertinent to classes of interest.

***Index Terms*—** Microseismic event classification, Trustworthy AI, Explainable AI

1. INTRODUCTION

Detecting endogenous seismicity due to deformation of slow-moving clay-rich landslides has become an important research topic to mitigate threats to humans, especially around roads, dams and train tracks. Seismometers provide accurate recordings of seismic signals; however, due to their high sensitivity, distinguishing between seismic signals originating from seismic activities and any other signals contained in the recordings is not an easy task. Deep learning-based approaches dominate recent literature on detection and classification of seismic activities, and a detailed review can be found in [1]. The deep learning models achieve state-of-the-art performance in detecting and classifying seismic signals avoiding cumbersome manual feature generation, selection and extraction process, with their ability to automatically learn most discriminative features from raw recordings. However, this also means that these models are limited by the used training set, and may learn specifically spurious correlations with the prediction target [2], [3]. Furthermore, the fact that the feature engineering task is taken away from the designer, makes deep learning models opaque, and hence often referred to as "black box", which limits their adoption.

Explainable artificial intelligence (XAI) [4], [5] is a research direction that provides human-interpretable explanations. XAI tools have been used in computer vision (e.g., [6]) and time-series signal analysis problems (e.g., [7]); however, XAI work on micro-seismic signal analysis is in its infancy. In [8], a heatmap-based visualisation tool was presented to explain model outputs via the outputs of activation functions of each filter in the convolutional layers and then overlapping the result with the raw input signal. However, it is not clear how explanations are formed by fusing outputs of the activation functions from different layers, only earthquake and other high SNR events are considered, the approach does not exploit advanced XAI methods, and it is not used to explain any false predictions. In [9], the authors proposed a Dual-Channel CNN together with an explanation module, EUG-CAM (elaborate upsampling-based gradient-weighted class activation mapping) that builds upon the principles of GradCAM (gradient weighted class activation mapping), harnessing the influence of feature map values and gradients to elucidate the importance of diverse features in the last convolutional layer. Recognising the discrepancy between feature map sizes and input data dimensions, EUG-CAM uses a strategic amalgamation of transposed convolution, unpooling, and interpolation, to generate feature mappings from a coarse localisation map. This results in an explanation feature map that effectively encapsulates class activation, learning insights, and network architecture considerations. However, the model's limitation is in classifying only two classes (rock fracturing vs. noises) and its confinement to binary classification. Furthermore, the reliance on a 1-D CNN model facilitates explanations primarily within the time domain, possibly neglecting the benefits of frequency-domain insights. Additionally, the visualisation maps cannot show the adverse input signal influence (negative contribution) on classification results hampering a comprehensive and well-rounded comprehension of the model's decision-making

This work was supported by EPSRC Prosperity Partnership research and innovation programme EP/S005560/1 and in part by EPSRC New Horizons research programme EP/X01777X/1. The contextual data interpretation and labelling work by experts on the SZ dataset was supported by RSE Saltire International Collaboration Awards.

process. In [1], visualization of feature maps is used to understand the CNN network's internal workings. The authors examine feature maps at various convolutional layers and the second fully connected (FC) layer, gaining insights into feature extraction. The main observation is that early layers locate event positions and extract basic features, while deeper layers refine these features into abstract representations for classification. The second FC layer's feature distributions vary across seismic events, indicating the network's capability to distinguish event types from noise based on learned features.

In this paper we identify the learnt features from a deep neural multi-classifier and demonstrate that these features are in agreement with the physical properties of seismic signals and hand-crafted features used in literature [10]. We leverage on state-of-the-art XAI tools to explain deep learning models for detection and classification of micro-seismic signals and show how these explanations can explain correct and wrong predictions. Specifically, we adopt Layer-wise Relevance Propagation (LRP) [11] to explain the decision making process, and analyse the basis of the model for event classification and discuss the reasons for misclassification. The outputs, both true and false positives, are therefore explainable in terms of geophysical features, thus building transparency into the operation of deep learning approaches.

2. METHODOLOGY

2.1. Data preparation and pre-processing

The dataset used in all our experiments is obtained from the open access Résif Seismological Data Portal, acquired by the French Landslide Observatory OMIV (Observatoire Multi-disciplinaire des Instabilités de Versants). The waveform data is acquired by Super-Sauze C (SZC) stations in MT network which are installed at the east and west sides of the Super-Sauze landslide in Southeast France (Latitude: 44.34787, Longitude: 6.67805). The seismometers consist of one 3-component sensor and 3 vertical one-component sensors (organised as equilateral triangle) at 250Hz sampling rate. The seismometers recorded three periods: from 11 Oct. to 19 Nov. 2013; from 10 to 30 Nov. 2014; and from 9 June to 15 Aug. 2015. We use a catalogue of manually labelled events that occurred during these periods for classification into four classes [1], [12]–earthquake, quake, rockfall and natural/ anthropogenic (N/A) noise. Rockfalls mainly occur at the main scarp of the landslide, where the rigid block falls from the steep slope (height $>$ 100m). The quake is likely to be triggered by material damage, surface cracks and openings. The earthquakes represent regional seismic events in this area and the teleseisms. N/A noise events include all anthropogenic and environmental noise, due to, e.g., transportation, pedestrian walking, heavy rain, animals, etc. We manually selected relatively higher-SNR (Signal-to-Noise Ratio) events from the catalogue for model training. Labelled events from 26th to 28th Nov. 2014, which are not included in the training set, are used for testing. We use a band-pass filter with 5-60Hz bandwidth to remove low frequency noise and retain events of interest, as identified from the cataloguing process [12]. We have 15-second sliding windows of continuous filtered recordings as input window. In particular, we normalize time series input by subtracting mean and dividing by the maximum of the absolute value of each input window. We use Short-time Fourier Transform (STFT) maps as model inputs that was shown in [1] to provide in average better results compared to feeding directly time-series signals. For the STFT map input, in order to get good time and frequency resolution, 'Boxcar' window with length of 128 samples with 70% overlap is used. The input shape for STFT-based model is $65 \times 95 \times 3$ samples.

2.2. Seismic signal classification

An STFT-based CNN model inspired by VGGNet [13] and adapted from [1] is used. The architecture of the the model is shown in Figure 1. Convolutional layers perform feature representation and extraction, followed by max-pooling layers that downsample the extracted feature into a feature map with smaller size. Compared to [1], to effectively process longer-duration seismic events within continuous data streams, we increase the input window of the CNN model to 15 seconds (from 10 seconds), simultaneously optimising associated parameters. Moreover, recognising the prevalence of seismic waveforms captured by 3-component sensors in the field, the input to the network is 3-channel input data, in contrast to 6-channel used in [1], which significantly expands the model's applicability to field studies.

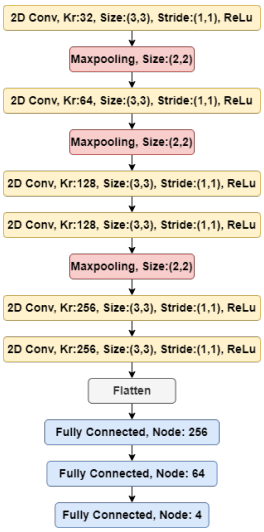


Fig. 1: STFT-based CNN for seismic classification.

2.3. Sliding window-based detection and post-processing

A sliding window method is used to segment the continuous stream into smaller windows [14], [15]. The window size and overlap are selected based on the temporal resolution required for the signal processing task. In particular, a window size of

3750 samples, which is the length of the input signal for the CNN model, is used. The overlap between consecutive windows is set to 93% of the window size, i.e., 3500 samples, which allows the CNN model to capture the temporal dynamics of the signal. For each window, the CNN model is used to predict the class of the signal using the trained weights. While the sliding window technique enables continuous detection, it can introduce certain challenges. One of the main issues is that it may break the continuity of the event waveform, leading to potential inconsistencies or artifacts in the classification results. This occurs because the sliding window segments are treated independently, without considering the temporal context or smooth transitions between adjacent windows. To address this problem, the proposed post-processing system is based on threshold filtering, median filtering, and Gaussian kernel filtering of the softmax output of the CNN. In addition, a peak selection method is applied to resolve cases where two classes of events have very similar detection results. (1) The softmax output of the CNN is filtered with a threshold value (0.5), and all values below this threshold are set to zero. (2) After the threshold filtering step, the probability distribution may contain isolated spikes. To remove these isolated spikes, we apply a median filter to each class separately. In addition to removing isolated spikes, the median filter can also merge spikes that are very close together, resulting in smoother and more continuous probability distributions. We set the size of the median filter to 5. (3) Gaussian kernel filtering, defined with a sum of 1 and a length of 15, is applied to the median filtered output to smooth the probability distribution. (4) We select the highest peak as the final output. This peak selection method allows us to choose the class of the event with the longest duration, as it indicates a higher confidence level in the classification result.

2.4. Explainability tools

LRP [11] is a state-of-the-art XAI method, that shows the contribution of each sample in the input data to the classification results and can be implemented in the pre-trained model. The LRP method starts from the output of the model, sets the output value before activation function as relevance, and gradually back propagates the relevance, iteratively, layer by layer, to the input nodes. In the backpropagation, the relevance follows the conservation law, that is, a neuron's relevance equals to the sum of the relevance it flows out toward all other neurons. Various propagation rules have been proposed, such as LRP-γ, LRP-ϵ, LRP-0 rule [16] [16]. In this paper, we used LRP-ϵ rule which is suitable for convolutional layers and max pooling layers [17], and is define as:

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k, \tag{1}$$

where R_j represents the relevance score assigned to neuron j, a_j denotes an input activation, w_{jk} is the weight connecting neuron j to neuron k in the layer above, $\sum_{0,j}$ denotes that we sum over all neurons j in the lower layer plus a bias term w_{0k} with $a_0 = 1$. ϵ is a regularisation term, i.e., a small value that prevents the denominator from being 0.

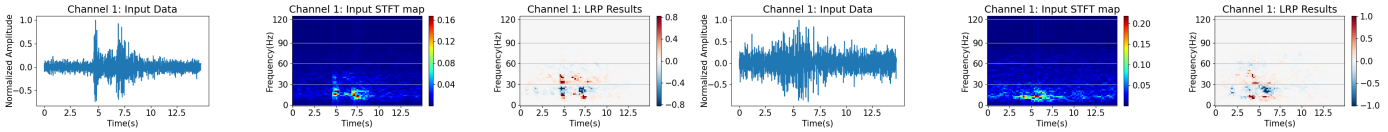
3. RESULTS

Our models are implemented in Keras framework. Since the activation function of the output layer is softmax, we use categorical cross entropy as loss function. The used optimiser is AdaDelta. The classification results are shown in Table 1. It can be seen that generally the model leads to a high recall (sensitivity) for the earthquake (88.2%) and rockfall (97.3%) classes. The worse results are obtained for N/A noise (79.2%) and quake (68.4%) signals, due to heterogeneity of the N/A noise signal and very low signal amplitude of quake signals. The results are aligned with those from [1] and [12].

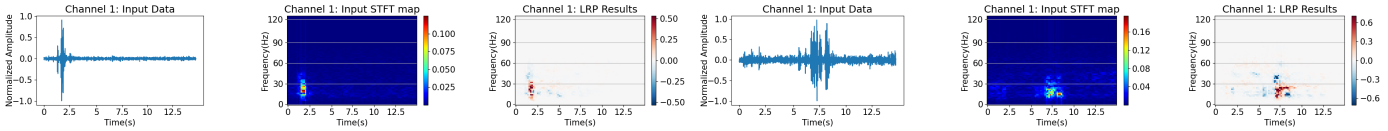
Table 1: The confusion matrix for STFT-based CNN model.

	Quake	Earthquake	Rockfall	Noise
Quake	**26**	2	8	2
Earthquake	0	**15**	1	1
Rockfall	2	0	**73**	0
Noise	95	11	37	**546**

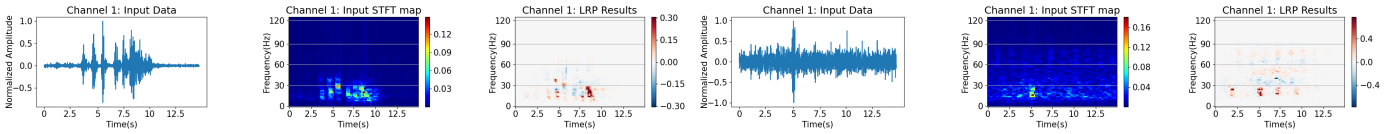
The used package for embedding LRP into our models is iNNvestigate [18]. Figure 2 shows an example of 3 correctly classified events. Positive and negative values of the LRP relevance indicate that the corresponding STFT values have positive and negative contributions to the classification results, respectively. The distribution of relevance for earthquake is focused on the high frequencies (about 40 to 50Hz) when the P-wave is picked as well as the low frequencies (around 15 to 20Hz) of the P-wave and (around 5sec) the low frequencies of the S-wave with intermediate noise shown in light blue correctly identified as not contributing (negative contribution). This example shows that the model learnt, and uses as basis for its predictions, that the P-waves of earthquake tend to have both high and low frequencies and that high energy content of S-Waves follows in time. Quake events are of shorter duration than earthquakes, have lower amplitudes, with energy focused in low frequencies. The relevance is concentrated in the single peak (positive and negative) of the event waveform, suggesting that the normalised maximum amplitude is the key distinguishing feature. In the frequency domain, the LRP map clearly shows the importance of the peak that has energy mainly focused below 30Hz while there is also a small positive contribution between 30 to 40Hz. While the relevance of quake events is concentrated on a single peak, the relevance of rockfall events is concentrated on multiple peaks, which also shows an important property of rockfall events $--$ multiple significant peaks. Looking at the LRP map, the relevance has multiple focused points corresponding to multiple short waves a characteristic of rockfalls. In addition, although both the rockfall and the quake events have a frequency band between 10 to 30Hz, the relevance are mostly concentrated at frequencies greater than 20Hz for rockfalls and below 20Hz for quakes.



(a) Correctly classified earthquake



(b) Correctly classified quake



(c) Correctly classified rockfall

Fig. 2: Correctly classified examples: The first column shows the time-series signal, middle column the STFT, and the right LRP relevance heatmap.

(a) noise mis-classified as earthquake

(b) rockfall mis-classified as quake

(c) quake mis-classified as rockfall

Fig. 3: Three mis-classified examples.

3.1. Explaining origin of mis-classification

The confusion matrix presented in Table 1 shows that the quake signals are sometimes misclassified as rockfalls. Interestingly, however, rockfall signals are rarely misclassified as quakes (only 2 mis-classified events). To investigate this further, Figure 3(c) shows an example of a quake event misclassified as rockfall, where the relevance distribution on the LRP map is very scattered. That is, the LRP relevance is not focused on the quake event's peak, but instead picked up several consecutive peaks, where the positive relevance is correctly concentrated at 5 sec. This indicates that the model correctly recognised the quake event's peak appearing around 5 sec, but there was a high energy signal in nearby frequency bands, influencing the final prediction. On the other hand, there are many positive relevances at different times that correspond to frequencies between 20Hz to 30Hz, which is akin to the learnt rockfall 'behavior'. Thus, the main reason of misclassification between quake and rockfall is that the SNR of the quake was very low, with a noise signal appearing immediately after mimicking multiple peaks of rockfall events, as shown in LRP map of Figure 2 (b).

In Fig. 3(a), the misclassification of noise as an earthquake is shown. The noise exhibits prominent peaks around 4 sec and 5.5 sec. Examination of the LRP map reveals the model's recognition of low-frequency and high-frequency components around the 4-second mark, along with low-frequency signals at 5.5 sec. This aligns with the characteristic features of P-waves and S-waves in earthquake signals, as shown in the LRP map of Figure 2(a), resulting in the model's mis-classification of this event as an earthquake.

In Fig. 3(b), we show an instance in which a rockfall is mis-classified as a quake. The rockfall displays multiple peaks, but aside from the principal one, all are of negligible magnitude. Analysis of the LRP illustrates a concentration of positive effects (red) at the primary peak of the event. Conversely, numerous negative contributions (blue) are observed at the secondary peaks, suggesting that the presence of these peaks is not taken into account due to their small magnitudes.

4. CONCLUSION

The paper proposed an STFT-based CNN model that achieves good performance on earthquake and rockfall events. Further using LRP to explain outputs of the proposed CNN, we identified properties of the signals extracted by the network when making decisions. Based on this, we concluded, for example, that the main reason why quake events are often misclassified as rockfall is due to appearance of a noise signal at multiple higher frequencies that mimics rockfalls. By observing the insights gained through XAI, we can discern specific features of input events that are prone to mis-classification. This knowledge can be instrumental in enhancing the robustness and generalisability of our model. This can be achieved by adding more events in the training set that closely resemble the challenging input patterns identified through XAI. The availability of LRP maps as visual aids can also offer a valuable tool to support cataloguing by geoscientists. This collaborative synergy between automated classification and manual classification can further enhance the accuracy of microseismic catalogues, contributing to a better understanding of geological phenomena.

5. REFERENCES

[1] Jiaxin Jiang, Vladimir Stankovic, Lina Stankovic, Emmanouil Parastatidis, and Stella Pytharouli, "Microseismic event classification with time-, frequency-, and

wavelet-domain convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[2] Charlotte Soneson, Sarah Gerster, and Mauro Delorenzi, "Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation," *PloS one*, vol. 9, no. 6, pp. e100335, 2014.

[3] Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba, "Understanding the role of individual units in a deep neural network," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30071–30078, 2020.

[5] Andreas Holzinger, "From machine learning to explainable ai," in *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 2018, pp. 55–66.

[6] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2912–2920.

[7] David Murray, Lina Stankovic, and Vladimir Stankovic, "Transparent ai: explainability of deep learning based load disaggregation," in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2021, pp. 268–271.

[8] Luca Trani, Giuliano Andrea Pagani, João Paulo Pereira Zanetti, Camille Chapeland, and Läslo Evers, "Deepquake—an application of cnn for seismo-acoustic event classification in the netherlands," *Computers & Geosciences*, vol. 159, pp. 104980, 2022.

[9] Xin Bi, Chao Zhang, Yao He, Xiangguo Zhao, Yongjiao Sun, and Yuliang Ma, "Explainable time–frequency convolutional neural network for microseismic waveform classification," *Information Sciences*, vol. 546, pp. 883–896, 2021.

[10] Jiangfeng Li, Lina Stankovic, Stella Pytharouli, and Vladimir Stankovic, "Automated platform for microseismic signal analysis: Denoising, detection, and classification in slope stability studies," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7996–8006, 2020.

[11] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, pp. e0130140, 2015.

[12] Floriane Provost, Clément Hibert, and J-P Malet, "Automatic classification of endogenous landslide seismicity using the random forest supervised classifier," *Geophy. Research Let.*, vol. 44, no. 1, pp. 113–120, 2017.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] Omar M Saad and Yangkang Chen, "Earthquake detection and p-wave arrival time picking using capsule neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6234–6243, 2020.

[15] Omar M Saad and Yangkang Chen, "Capsphase: Capsule neural network for seismic phase classification and picking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[16] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[17] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.

[18] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans, "innvestigate neural networks!," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.