processes

MDPI

Article

Interpretable Short-Horizon Prediction of High-Energy Microseismic Events for Rockburst Risk Screening Using TabPFN-SHAP

Feng Jin [1], Huilin Jia [2,*], Yubao Zhao [1,2], Yatian Wang [1], Qi Guo [1], Guorui Su [3] and Zhe Liu [2]

[1] China National Coal Group Corporation, Beijing 100120, China
[2] School of Safety Engineering, China University of Mining and Technology, Xuzhou 221116, China
[3] Information Institute of the Ministry of Emergency Management, Beijing 100029, China
* Correspondence: jhlcumtsafe@163.com

Abstract

Routine microseismic monitoring records event time, location and energy, yet converting these data into interpretable guidance for mine inspection remains difficult. We combine Tabular Prior-Data Fitted Networks (TabPFN) with permutation SHapley Additive exPlanations (SHAP) to estimate whether a high-energy microseismic event will occur in the same grid within 24 h, thereby supporting rockburst risk screening without directly predicting rockburst occurrence or damage. We aggregated 9116 events into 3298 grid-time samples described by 28 energy, temporal-activity and spatial-distribution variables. Samples were divided chronologically using 48 h buffers, and thresholds were selected only from validation data. On the held-out test period, TabPFN achieved the highest observed areas under the receiver operating characteristic and precision-recall curves (ROC-AUC 0.8064; PR-AUC 0.5330) and the lowest Brier score (0.1190) among six classifiers. Four controlled semi-synthetic domain-shift tests produced TabPFN ROC-AUC values of 0.6992–0.7752. SHAP identified preceding-window cumulative energy and event count as the leading contributors, with stable rankings under expanded sampling. The framework therefore links grid-level screening to monitoring quantities that engineers can inspect. These semi-synthetic results are stress tests rather than independent-mine validation, and no transferable warning threshold is claimed.

Keywords: rockburst risk; microseismic monitoring; TabPFN; explainable artificial intelligence; SHAP; spatiotemporal aggregation; high-energy event; tabular data

Check for updates

Academic Editor: Jae-Yoon Jung

Received: 1 September 2026
Revised: 15 October 2026
Accepted: 29 October 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As coal extraction extends to greater depths, high in-situ stress interacts with longwall advance, roadway excavation and geological structures. These processes can increase the likelihood of dynamic failure around mining openings. Rockburst is commonly associated with the sudden release of stored strain energy [1,2] and can damage the coal-rock mass, support systems and mining infrastructure. The interaction of stress redistribution, mining disturbance and material properties makes it difficult to assess this hazard with one indicator or empirical criterion [3].

Data-driven approaches have expanded the range of information used in rockburst assessment. Early work combined microseismic indicators through fuzzy comprehensive evaluation [4]. Supervised methods subsequently included support vector machines, neural networks and decision trees for case-based classification [5,6], alongside Bayesian inference

with incomplete observations [7]. More recent studies have investigated microseismic-based prediction of rockburst intensity [8], deep-learning analysis of multiple monitoring signals [9], and multi-task warning of coal mine hazards [10]. Reviews document these developments and their different data requirements [11,12].

The closest recent studies also illustrate the difference between the present task and established case-based prediction. Wang et al. used 88 cases with surface displacement, mining speed and face-advance variables to classify high-energy microseismic levels with a back-propagation neural network [13]. Kadkhodaei and Ghasemi used 93 labelled rockburst cases and six cumulative or rate-based microseismic indicators to train optimised random-forest models and interpret them with SHAP [14]. These studies show that monitored precursors can support short-term assessment, but their units of analysis are event/case records with additional engineering variables or labelled rockburst severity. They do not address repeated same-grid forecasts constructed from a continuous five-field event stream, nor do they test that setting with a chronological future period and a separately selected operating threshold. The practical question addressed here is therefore how a mine can use a basic event record to update local microseismic screening results and explain the basis of each alert.

However, many existing intelligent rockburst studies use curated engineering case tables, for which the sample size, available indicators and risk classes are constrained by the completeness of site records [5,7]. These datasets are useful for comparing risk between cases, but they do not directly represent how microseismic activity evolves within one mine as extraction progresses. For field warning, a model must update local risk estimates as new monitoring observations arrive. A risk class or probability is also difficult to use on site unless engineers can inspect whether the judgement arises mainly from energy accumulation, changing event activity or increasing spatial concentration [15].

Microseismic monitoring records the timing and location of local energy release and is an important source of information for rockburst-related assessment [4,16,17]. Its uses extend beyond Chinese coal mines. Seismic tomography has been used to update fracture-zone interpretations during deep block caving in Indonesia [18], while source parameters and legacy three-dimensional seismic data have been combined to examine rockburst-related structures at Kloof Gold Mine in South Africa [19]. In a broader seismic-prediction setting, Dzhangirov et al. compared linear inversion and neural methods for reservoir-property prediction in Kazakhstan [20]. These international examples address different commodities and prediction targets, but they show how monitoring interpretation depends on mining geometry, geology, network support and the intended decision.

Studies have also combined microseismic observations with acoustic emission and electromagnetic radiation for multi-level warning [21] and for recognition of interfering signals [22]. Automatic event-location methods can improve the preparation of monitoring data [23]. Related coal mine applications include reliability-constrained evidence fusion [24], explainable coal-gas risk identification [25], and knowledge-augmented recognition of outburst precursors [26]. Coal-and-gas outburst is not the same hazard as rockburst: the cited studies are included only because they show how heterogeneous monitoring signals can be represented, fused and explained under changing operating conditions. They are not treated as evidence that a rockburst model has been validated. Recent acoustic-emission research likewise showed that warning information may be expressed through temporal reorganisation rather than a simple change in raw amplitude [27].

Multi-source systems require compatible sensors and sufficiently complete records. The information consistently available in the present study consists of event time, location and energy. Point-cloud approaches have shown that such event streams contain useful spatiotemporal structure [28]. Nevertheless, an individual event is not a complete regional

state sample. We therefore describe the recent energy, activity and spatial state within a grid and ask whether a high-energy event follows in that grid.

Spatiotemporal aggregation provides cumulative energy, event rate, inter-event intervals and measures of spatial spread [29]. Event rate and cumulative energy are established monitoring quantities in mining seismology [30]. Other work has examined frequency–magnitude relationships and stress dependence [31,32], as well as time-varying acoustic-emission characteristics of coal with different burst proneness [33]. Field studies also support combining multiple microseismic descriptors [34]. Using the available event fields, we calculate descriptors over two adjacent historical windows and learn their association with future high-energy activity.

We use Tabular Prior-Data Fitted Networks (TabPFN) because a long event record does not necessarily yield a large labelled training set. After regional aggregation and chronological splitting, the baseline training set contains 2110 samples, including 341 future-high-energy positives. Training a large task-specific network is difficult to justify from this amount of site data alone. TabPFN instead uses a Transformer pre-trained on tabular prediction tasks and conditions its predictions on labelled examples from the current problem [35,36]. It is therefore a candidate for using nonlinear relationships in a modest-sized numerical dataset without training a deep network from scratch at each site.

Gradient-boosted trees remain strong alternatives for this type of data [37,38]. TabPFN is not assumed to solve class imbalance [39], temporal dependence or changes in mining conditions by itself. Those issues are addressed through sample construction and evaluation. The question tested here is whether pre-trained tabular inference is useful for daily screening of recently active grids, compared with established classifiers using the same historical descriptors. SHAP adds an inspectable explanation of the monitoring information used, rather than an improvement to the classifier's predictions.

A probability alone does not show which changes in microseismic activity produced an alert; explaining individual predictions has become a general problem in the application of black-box models [40,41]. To make the basis of model decisions inspectable, we use SHapley Additive exPlanations (SHAP) to decompose an individual probability into contributions from the input descriptors and to summarise the information repeatedly used across test samples [42]. Because TabPFN has neither directly readable linear coefficients nor a tree-split structure, we use model-agnostic permutation SHAP, following related work on the interpretation of TabPFN models [43]. The probability output can therefore be related to observable states such as energy accumulation, event activity and spatial concentration; interpretable machine-learning methods have likewise been applied to short-term rockburst risk prediction from monitored microseismicity, lending support to this line of reasoning [44].

We address four practical questions. Can basic microseismic event information alone be transformed into a repeatable short-horizon risk proxy? Does TabPFN improve probability ranking and calibration relative to competitive tree ensembles and simple baselines under chronological testing? Which derived descriptors drive the fitted model, and are the leading descriptors stable when the SHAP explanation sample changes? Do the main conclusions depend on a particular history window, grid size or high-energy threshold?

The main methodological contributions are as follows:

1. Discrete microseismic events are reconstructed as local grid-time state samples. Energy release, event activity and spatial distribution are aggregated within each grid-time unit, so the model represents the recent state of local microseismic activity rather than treating one event as a complete prediction case.
2. A TabPFN-SHAP method is proposed for short-horizon high-energy microseismic screening. Pre-trained tabular inference is evaluated under limited site-specific train-

ing data, while permutation SHAP relates the output to energy, activity and spatial descriptors that engineers can review.

3. Evaluation follows the sequence of fitting a model and screening a later monitoring period. Historical-only predictors, chronological splitting and boundary buffers reduce information sharing between subsets. Validation-selected thresholds and reporting of missed positives, false positives and alert workload connect the comparison to the practical demands of site review.

2. Materials and Methods

2.1. Microseismic Data and Prediction Target

Longwall advance and roadway excavation in deep coal mines continually alter the stress and fracture activity of the surrounding coal-rock mass. Field microseismic systems provide a basic record of this evolution by continuously logging event time, location and energy. Owing to the monitoring configuration and completeness of the available records, these were the fields that could be obtained consistently throughout the study period. A single event represents only one local energy release at a particular time and location and cannot constitute a complete risk case. We therefore organised the continuous event observations into spatial-temporal samples that describe the recent state of a local region.

The data were obtained from the 615 working face of Longjiabao Coal Mine in Jilin Province, China. The face lies between the −1010 and −1050 m mining levels, at a reported mining depth of 1139.8–1233.5 m. It extracts the combined No. 2 and No. 3 seams, with a working thickness of 9–13 m (12.5 m on average). The panel is approximately 463 m along strike and 178 m along dip; seam dip ranges from 10° to 20°, with an average of 15°. Mining uses a retreating strike-longwall, fully mechanised top-coal-caving method, and the roof is managed by total caving. The engineering information in this subsection was supplied by the mine data owner and is not publicly available because of site-confidentiality requirements.

The basic roof comprises mainly sandstone, tuff and conglomerate, whereas the immediate roof is dominated by mudstone and siltstone. Three normal faults with reported throws of 1, 5 and 6 m intersect or influence the face, and small undulating folds are also present. The combined coal seam and its roof were assessed as having weak burst proneness, the floor as having no burst proneness, and the face as having medium rockburst hazard. For regional dynamic-load monitoring, two geophones were installed in each of the transport and return gates, at nominal along-roadway intervals of 300–500 m; the sensors were repositioned outward as the face advanced so that the array continued to enclose the active face. The microseismic system returned event occurrence time, source location and released energy, which are the fields analysed here. The technical material available to this study does not state the equipment manufacturer and model, acquisition sampling rate, location algorithm or uncertainty, or verified system-uptime record; these quantities are therefore not inferred.

After basic quality control (parsing occurrence times, coercing coordinates and energy to numeric values, removing rows with missing core fields or duplicate records, and retaining only positive-energy events), the catalogue contains 9116 records. Each event contains five basic fields: occurrence time, x, y, z and recorded energy. The monitoring interval extends from 21 August 2022 to 5 January 2025. The coordinate ranges are 69,271–70,336 m in x, 72,814–73,949 m in y, and −1179 to −448 m in z; recorded energy ranges from 39 to 2.247×10^6 J.

The model-input table does not include grid-specific stress measurements, geological annotations, excavation progress, support response, damage observations or rockburst grades. The analysis is therefore based on the event stream and the descriptors derived from it; the site description above supplies engineering context but is not used as a predictor.

We defined the prediction target as whether the same spatial grid contains a high-energy microseismic event during the following 24 h. Given the recent microseismic state of a grid, the model estimates the probability of that event-based outcome during the next day. The resulting probability provides a proxy indicator for short-horizon rockburst risk screening [45]. It does not represent engineering damage, injury, rockburst occurrence or a rockburst grade that can be transferred directly to another site.

Throughout this paper, prediction refers only to the future high-energy microseismic event defined above. The model output, $\hat{p}_{g,\tau}$, is therefore called the predicted high-energy-event probability. Screening means using this probability to prioritise eligible grids, and a screening flag is generated when the probability exceeds the validation-selected threshold. Rockburst risk screening describes the intended engineering use of the probability and flag after review with other site information; it is not the model's directly observed outcome. A high-energy event is not equivalent to a damaging rockburst, and the absence of the proxy event is not evidence that a location is safe. Without matched rockburst observations, the study cannot estimate sensitivity or false-alarm rates for actual rockbursts.

Figure 1 summarises the microseismic data in three connected parts: temporal energy evolution, three-dimensional event location, and a statistical overview comprising the energy distribution and monthly event count. The event density along the time axis is uneven: some periods are densely recorded, whereas others are sparse or contain long gaps. Such gaps may reflect monitoring interruptions, changes in the excavation stage or changes in recording conditions, and should not automatically be treated as periods with no microseismic activity. If sample windows were allowed to cross unobserved intervals, the model could treat missing monitoring time as normal low activity, which would affect both feature calculation and validation. Therefore, timestamp-defined monitoring blocks are identified before sample construction, and purge intervals are placed around train-validation-test boundaries to reduce window overlap and misuse of future information.

The analysis shows how far the routinely recorded event stream itself can be used. Time was used to order events, detect monitoring continuity and define historical and forecast windows. The three coordinates were used to assign events to spatial cells and to describe local spatial distribution. Energy was used to characterise release magnitude and to define future high-energy microseismic events. Microseismic-state descriptors were built directly from the event stream, without pre-existing damage grades or site-specific geological annotations. Energy is expressed in joules (J).

The baseline high-energy threshold was 3644 J, the P90 energy value among the first 60% of chronologically ordered events (5469 records). This event-based calibration period ended on 8 June 2023. By comparison, the purged training period ended on 20 October 2023, validation began on 24 October 2023 and the independent test period began on 9 April 2024. The P90 calculation therefore used only an early subset of the eventual training-period event stream and no validation- or test-period events. The 60% event cutoff is not the same index as the later 60% split of unique grid-time anchors: one is counted over individual events, whereas the other is counted over prediction times after sample construction.

P90 identifies the upper part of the observed energy distribution while retaining enough positive grid-time samples for model comparison. The value of 3644 J is specific to this dataset and is not a universal rockburst criterion. The P85 and P95 analyses in the Results section examine whether model ranking depends on this choice.

2.2. Event-to-Sample Transformation and Feature Engineering

After quality control, the event sequence was divided wherever a monitoring gap exceeded 24 h. This cutoff matches one full history or forecast interval: allowing a 24 h sample window to cross a longer event-free interval would calculate activity descriptors

as though that interval had verified monitoring coverage. There were 99 such gaps, with durations from 24.02 to 674.14 h (median 36.80 h), producing 100 timestamp-defined blocks before sample eligibility was applied. Only 24 blocks contained grid-time units that met the history/current-activity/forecast construction requirements. Each retained block was discretised into 100 m × 100 m × 100 m grid cells with a 24 h time step. Grid indices were referenced to the minimum coordinates of the full catalogue, so an index such as $(3, 2, 2)$ refers to this data-specific origin rather than an existing site grid.

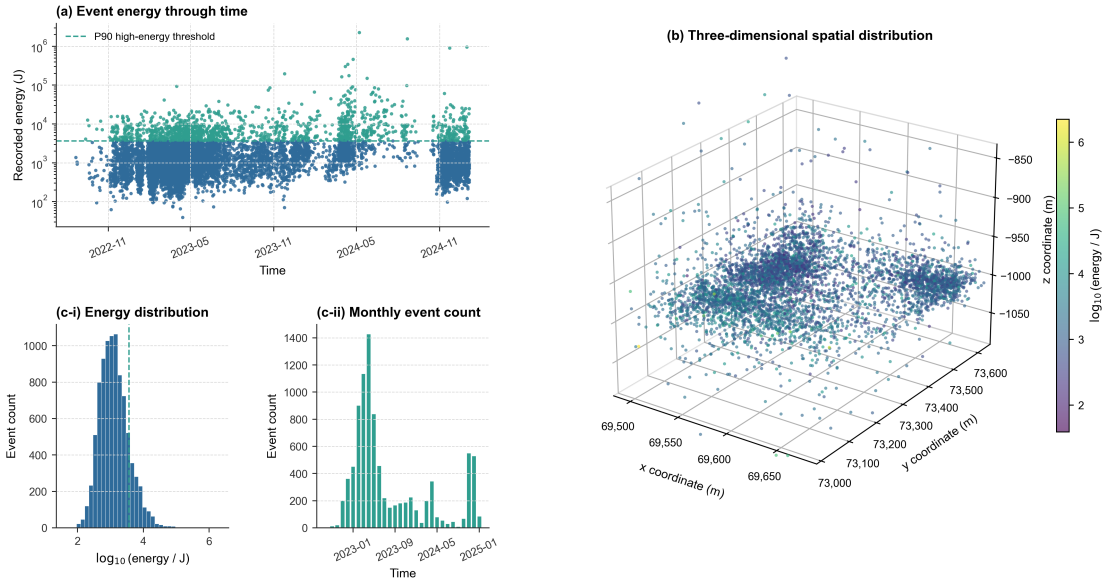


Figure 1. Temporal, spatial and statistical characteristics of the microseismic data: (**a**) recorded energy over time, with the baseline P90 high-energy threshold; (**b**) three-dimensional event locations coloured by energy; (**c-i**) log-energy distribution and P90 threshold; and (**c-ii**) monthly event count.

These blocks are defined from event timestamps, not independently verified system-uptime logs. A long inter-event gap may reflect either low activity or a recording interruption. The 24 h rule avoids spanning a complete analysis window with no event support, but it cannot establish monitoring completeness within every retained window and has not been calibrated against an external uptime record.

For a grid cell and anchor time t, the current window $(t - 24\,\text{h}, t]$ supplied the current-state predictors, the preceding window $(t - 48\,\text{h}, t - 24\,\text{h}]$ supplied the comparison state, and the forecast window $(t, t + 24\,\text{h}]$ supplied the binary label. Thus, each historical window has duration $\Delta_h = 24$ h, while the two adjacent windows provide a total predictor lookback of $2\Delta_h = 48$ h. These half-open intervals assign an event on a shared boundary to exactly one adjacent window. Grid-time units without events in the current window were excluded because the analysis concerns subsequent risk when local microseismic activity has already been observed.

Consequently, all reported performance measures concern recently active grids, not every location in the mine. A grid with no recorded event in the current window receives no prediction under this sampling rule; it must not be assigned a low-risk class by default. This selection can exclude a location that is quiet before a high-energy event, so whole-mine coverage and the frequency of such missed transitions have not been established. In practice, event absence must also be distinguished from a monitoring interruption using system-status information.

The 24 h per-window baseline gives a daily screening unit and a 24 h forecast horizon over which site personnel can review newly recorded activity. The two adjacent historical windows permit comparison of recent states over a total 48 h lookback. Shorter per-window

durations may contain too few events for stable summaries, whereas longer durations combine activity over a broader stage of mining. The 100 m cubic grid is an analysis scale for regional aggregation, not a measured fracture dimension or a stated event-location accuracy. It balances the desire to retain local variation against sparse observations in smaller cells. Site-specific location-error estimates and excavation geometry were not available to establish a uniquely appropriate grid size. The alternative scales in Section 4.3 therefore test sensitivity to these choices rather than demonstrate a universal optimum.

Sample construction follows temporal order: features use only events before the anchor time, and labels use only events after it. Samples are generated separately within each timestamp-defined monitoring block so that a long event-free interval is not automatically treated as a period of verified zero activity. The mathematical definitions are given in Section 2.3.

Each sample is described by derived features of recent local activity rather than by individual event records. The selected quantities include cumulative energy, event rate and spatial spread. Energy-state features describe cumulative and extreme release and changes from the preceding window; activity features describe event count and occurrence rate [46]; spatial features describe three-dimensional dispersion and energy-centroid movement [47]. Some descriptors are strongly correlated, including cumulative energy and energy release rate. We retain them and account for their dependence when interpreting the SHAP results.

We handled sparse histories with fixed rules before model fitting. Count and energy variables were allowed to be zero when the corresponding event type was absent from an otherwise valid historical window. Log-energy summaries were calculated only from observed positive-energy events. Interval-based descriptors were assigned conservative defaults when fewer than two events were available. These rules retain low-activity windows without deleting samples solely because a statistic is undefined. Samples are constructed only within timestamp-defined monitoring blocks to reduce the effect of long event gaps on the identification of low activity. No transformation used information from the forecast window.

The baseline setting produced 3298 grid-time samples, including 509 positive high-energy-event labels (15.43%). The 28 predictors were grouped as energy state, temporal activity and spatial distribution (Table 1), and their complete definitions are provided in Appendix A. SHAP therefore explains descriptors calculated from the event catalogue, rather than the raw coordinates or energy of an individual event.

Table 1. Derived feature groups used to identify future high-energy-event risk.

Feature Group	Number	Derived Variables
Energy state	11	Current-window cumulative energy, maximum energy, P90 energy, mean and standard deviation of log-energy, energy concentration, high-energy count and ratio, energy release rate, preceding-window cumulative energy, and energy change.
Temporal activity	7	Current-window event count and event rate, preceding-window event count, event-count change, mean and minimum inter-event interval, and inter-event coefficient of variation.
Spatial distribution	10	Standard deviations in x, y and z, RMS and maximum spatial spread, energy-weighted spread, event and energy density, centroid offset, and centroid migration distance.

2.3. Proposed TabPFN-SHAP Prediction and Explanation Framework

Figure 2 summarises the complete route from microseismic event records to grid-level engineering decision support. The event stream is first reconstructed as grid–time samples containing 28 descriptors of energy state, temporal activity and spatial distribution. Prediction features are calculated only from the preceding and current windows; events in the next 24 h are used to define the training label and are not included in the model input. TabPFN then estimates the probability of a high-energy event in the same grid. A validation-selected threshold converts this probability into a screening flag, and permutation SHAP explains the feature basis of the probability at both global and individual-sample levels. In Figure 2, the displayed "Risk Alert" is this threshold-based screening output, not a prediction that a damaging rockburst will occur. Engineers may use the event probability, alert and explanation to support rockburst risk screening, but the model does not output a rockburst probability.

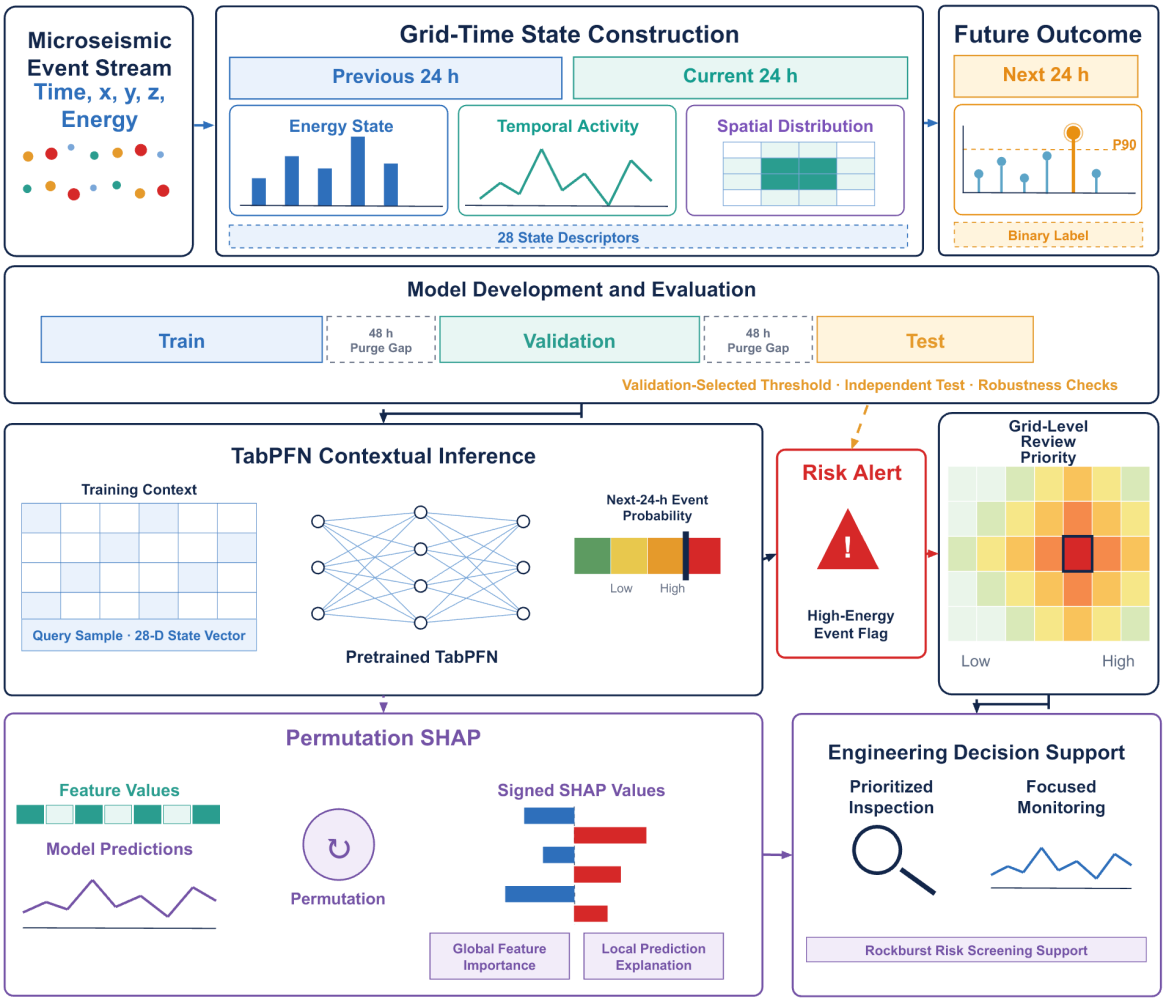


Figure 2. Overall framework of the proposed TabPFN-SHAP method, from historical microseismic-state construction and future high-energy-event labels to probability estimation, threshold-based risk alerts, SHAP interpretation and engineering review. The risk alert denotes a high-energy-event screening flag rather than a predicted rockburst occurrence.

The sample construction and model explanation are defined as follows. A raw microseismic event is written as

$$e_i = (t_i, \mathbf{u}_i, E_i), \qquad \mathbf{u}_i = (x_i, y_i, z_i), \tag{1}$$

where t_i is the occurrence time, $\mathbf{u}_i$ is the three-dimensional location and E_i is the released energy. For a spatial grid g and an anchor time τ, the current window, preceding window and forecasting event set are

$$\begin{aligned} \mathcal{H}_{g,\tau} &= \{e_i : t_i \in (\tau - \Delta_h, \tau],\ \mathbf{u}_i \in g\}, \\ \mathcal{H}^-_{g,\tau} &= \{e_i : t_i \in (\tau - 2\Delta_h, \tau - \Delta_h],\ \mathbf{u}_i \in g\}, \\ \mathcal{F}_{g,\tau} &= \{e_i : t_i \in (\tau, \tau + \Delta_f],\ \mathbf{u}_i \in g\}, \end{aligned} \tag{2}$$

with $\Delta_h = \Delta_f = 24$ h in the baseline experiment. The binary high-energy-event label is defined as

$$y_{g,\tau} = \mathbb{I}\left(\sum_{e_i \in \mathcal{F}_{g,\tau}} \mathbb{I}(E_i \geq E_q) \geq 1 \right), \tag{3}$$

where E_q is the energy threshold used to construct the proxy label. The sum is zero when the forecast set is empty, provided that the forecast window is treated as observed under the monitoring-block rule. This is equivalent to testing the maximum energy for a non-empty set but also defines the empty-set case. Only samples with $|\mathcal{H}_{g,\tau}| \geq 1$ are included. The model input is a derived state vector

$$\mathbf{s}_{g,\tau} = \Phi(\mathcal{H}_{g,\tau}, \mathcal{H}^-_{g,\tau}) = [s_1, s_2, \ldots, s_{28}], \tag{4}$$

where $\Phi(\cdot,\cdot)$ denotes the feature-extraction operator that summarises event count, cumulative energy, energy density, spatial dispersion and centroid migration from $\mathcal{H}_{g,\tau}$, and computes change-type descriptors by comparing $\mathcal{H}_{g,\tau}$ with $\mathcal{H}^-_{g,\tau}$; the idea of contrasting long-term and short-term microseismic features for rockburst risk prediction has recently received attention [48]. For compactness, define $f(\mathbf{s}) = f_{\mathrm{TabPFN}}(\mathbf{s} \mid \mathcal{D}_{\mathrm{train}})$ after the training context has been fixed. The TabPFN probability and the SHAP explanation can then be written as

$$\hat{p}_{g,\tau} = f\left(\mathbf{s}_{g,\tau}\right), \qquad \hat{p}_{g,\tau} - \mathbb{E}_{\mathbf{s}\sim B}[f(\mathbf{s})] = \sum_{j=1}^{28} \varphi_j(\mathbf{s}_{g,\tau}), \tag{5}$$

where B is the SHAP background distribution and φ_j is the attribution assigned to the jth derived feature. Global importance is reported as $\mathbb{E}(|\varphi_j|)$ over the explained test samples.

In the prediction stage, TabPFN uses the labelled training samples as context and estimates the high-energy-event probability of a new grid-time sample from $\mathbf{s}_{g,\tau}$ [36].

In the second stage, we apply permutation SHAP to the fitted TabPFN classifier. SHAP keeps the predicted probability unchanged and assigns that probability change to the engineered input features under a background reference distribution. The global SHAP analysis identifies the main descriptors that repeatedly influence model output across independent-test samples, whereas the SHAP waterfall plot shows the feature contributions for an individual high-probability prediction. Because the input features are aggregated from events within a grid and historical window, the SHAP results can be related to monitored states such as recent energy accumulation, persistent activity and spatial concentration; this practice of aggregating local attributions into global understanding is consistent with explainable artificial intelligence [49].

3. Experimental Design

3.1. Chronological Validation and Screening Threshold

We ordered the samples by anchor time and divided them into training, validation and independent test periods [50,51]. At each train–validation or validation–test boundary, samples with anchor times falling within the preceding 48 h or the following 48 h were excluded. Thus, "48 h buffer" refers to the one-sided buffer length; the total anchor-time

exclusion zone centred on each boundary was 96 h. After this purge, the training, validation and test sets contained 2110 samples (341 positives), 478 samples (50 positives) and 671 samples (109 positives), respectively. The independent test period began on 9 April 2024. This temporal separation does not make neighbouring grids or consecutive test anchors statistically independent; it prevents direct cross-subset window overlap, while residual dependence is addressed separately in the uncertainty analysis below.

We selected the alert threshold for each classifier on the validation set only, by maximising balanced accuracy. When several thresholds tied, the lower threshold was retained to favour sensitivity in a safety-screening context. The selected threshold was then fixed before the independent test set was scored. This separates model fitting, operating-threshold selection and final evaluation, and avoids the optimistic practice of choosing a threshold after seeing the test labels.

For a candidate threshold c, balanced accuracy is

$$\mathrm{BA}(c) = \frac{1}{2}\left[\frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FN}(c)} + \frac{\mathrm{TN}(c)}{\mathrm{TN}(c) + \mathrm{FP}(c)}\right]. \quad (6)$$

This gives equal weight to the detection rate in each class and avoids selecting a threshold simply because negative samples are more common. It does not assume equal engineering consequences for one missed event and one false alert. No site-specific costs were available, so balanced accuracy provides a reproducible reference operating point rather than a cost-optimal field policy.

The reported TabPFN threshold of 0.1184 means that an eligible grid is flagged when its estimated next-24-h high-energy-event probability reaches this value. It does not mean an 11.84% probability of a damaging rockburst or define a universal safety limit. Its numerical value is specific to the model output scale and validation period. A threshold below 0.5 is compatible with screening an uncommon event when sensitivity matters, but the choice still entails false alerts. For deployment, the mine would need to specify acceptable missed-event and review workloads, select the operating point on separate calibration/validation data, and then evaluate it prospectively. The reported test results apply only to the threshold fixed for this study, not to thresholds selected after inspecting the test outcomes.

The buffer length was determined from the widest predictor lookback: the current and preceding 24 h historical windows together extend 48 h before an anchor time, while the label uses the following 24 h. A sample immediately adjacent to a chronological boundary could otherwise share historical information or overlap in its future label period with a sample assigned to another subset. Applying the same 48 h buffer on both sides is deliberately conservative relative to the 24 h forecast horizon: it reduces the number of available observations but makes the validation and test periods more defensible as future-like evaluations. The test period is therefore a temporally separated sequence of grid-time decisions whose labels were not consulted during threshold selection.

Here, an independent test period means a held-out period, not a set of mutually independent observations. Adjacent test anchors may share historical events, and neighbouring grids may respond to the same mining disturbance. The boundary buffer limits overlap between subsets but does not remove these within-period relationships. This distinction affects uncertainty estimates and the interpretation of model comparisons.

3.2. Compared Models and Implementation

The comparison included logistic regression, random forest, XGBoost, LightGBM and CatBoost [52–55]. The three gradient-boosted trees are competitive methods for tabular classification [56,57], while logistic regression and random forest provide linear and bagging references; earlier studies have also used microseismic features with decision-tree classifiers to extract rockburst risk discrimination knowledge [6].

The reported experiment used fixed configurations and random seed 20260802 rather than repeated adjustment against test results. The baselines use model-specific class weights to address class imbalance [39] (Table 2). No additional weighting or resampling was applied to TabPFN. The comparison therefore does not impose identical imbalance treatment: the weighted baselines alter their fitted objective, whereas TabPFN retains its default inference and all models receive a validation-selected decision threshold. This distinction may affect both probability scale and classification results and is part of the definition of the compared implementations, not evidence that one imbalance strategy is generally preferable.

TabPFN was used without extensive task-specific hyperparameter search. All six models received the same 28 features, chronological subsets and validation-based threshold procedure. The tree configurations combine a low learning rate with restricted depth or leaf size to limit model complexity on the available training samples. Logistic regression supplies a linear reference and random forest a bagging reference. These choices define a fixed-configuration comparison, not an equal-budget search for the best attainable performance of each algorithm. Accordingly, the results compare the listed implementations and should not be read as a general ranking of model families.

Table 2. Models and configurations reported for the comparison.

Model	Role	Configuration
TabPFN v2.6	Classifier in the proposed TabPFN-SHAP framework	Official v2.6 classifier checkpoint; random seed 20260802; no class weighting and no task-specific hyperparameter search.
CatBoost	Gradient-boosting baseline	500 iterations, depth 6, learning rate 0.03, L2 leaf regularisation 3, Bayesian bootstrap, class weights.
XGBoost	Competitive gradient-boosting baseline	500 trees, depth 4, learning rate 0.03, minimum child weight 3, 0.8 subsampling and column subsampling, regularisation and class weight.
LightGBM	Competitive gradient-boosting baseline	500 estimators, learning rate 0.03, 15 leaves, 0.8 subsampling and column subsampling, class weight.
Random forest	Bagging baseline	500 trees, minimum leaf size 3, balanced subsampling.
Logistic regression	Linear baseline	Standardisation, balanced class weight, maximum 5000 iterations.

The reported software environment was Python 3.9.23, NumPy 2.0.2, pandas 2.3.2, SciPy 1.13.1, scikit-learn 1.6.1, PyTorch 2.7.1+cu128, TabPFN 8.0.3 with the official v2.6 classifier checkpoint, SHAP 0.49.1, XGBoost 2.1.4, LightGBM 4.6.0 and CatBoost 1.2.10. The three original SHAP resamples used seeds 20260802–20260804; the enlarged SHAP calculations used seed 1729 with deterministic offsets for the three repetitions. These values, together with the archived prediction tables, specify the computational setting more fully than the model parameters alone.

3.3. Performance Evaluation and Interpretability Protocol

We report ROC-AUC, PR-AUC, Brier score, F1 score, balanced accuracy, precision, recall, alert rate and alert counts. ROC-AUC describes ranking discrimination, while PR-AUC emphasises performance on the uncommon positive class [58,59]. The Brier score measures mean squared probability error [60,61]. Lower Brier error is desirable, but does not by

itself establish calibration: probability error reflects both calibration and discrimination. Calibration concerns whether groups assigned a given probability exhibit a similar observed event frequency [62,63].

Pairwise ROC-AUC differences were first assessed with the DeLong test [64]. This paired comparison accounts for different models predicting the same cases, but the ordinary implementation does not account for dependence between different grid-time cases. Its confidence intervals and p-values are therefore retained as nominal, exploratory comparisons.

To represent serial dependence more directly, we additionally used a paired temporal-block bootstrap within each timestamp-defined monitoring segment. Non-overlapping, non-circular 14-day blocks were sampled with replacement within each segment; the same sampled blocks were applied to all six models, while model predictions and validation-selected thresholds were held fixed. We requested 2000 replicates and excluded replicates in which a required metric was undefined because the resample contained only one outcome class. This left 1980 valid and 20 invalid replicates. Percentile 95% intervals were calculated for all reported metrics, and paired intervals were calculated for the TabPFN-minus-baseline ROC-AUC differences. The results are reported in Appendix B.

Probability reliability was evaluated using the Brier score and a common set of ten equal-width bins. For bin b, let n_b be its sample count and let $\bar{p}_b$ and $\bar{y}_b$ denote mean predicted probability and observed outcome frequency. Expected calibration error was calculated as $\mathrm{ECE} = \sum_b (n_b/n)|\bar{y}_b - \bar{p}_b|$, and maximum calibration error as $\mathrm{MCE} = \max_b |\bar{y}_b - \bar{p}_b|$. Calibration intercept α and slope β were estimated from $\mathrm{logit}\{P(Y=1)\} = \alpha + \beta\,\mathrm{logit}(p)$ after clipping probabilities away from zero and one. These summaries describe the held-out period; no post-hoc recalibration was fitted.

The baseline permutation-SHAP analysis used five stratified training samples as background and explained 30 stratified samples from the held-out test period. This low-cost setting is small relative to 28 predictors and was treated as a preliminary explanation rather than sufficient evidence on its own. Stratification retained both outcome classes rather than restricting explanation to high-scoring cases. With 28 predictors, the forward-and-reverse evaluation budget was 57 per explained sample. We therefore repeated the analysis with backgrounds of 20 and 50 cases, explanation sets of 100 and 300 cases, and maximum evaluation budgets of 285 and 570. Each configuration used three repetitions. Rank agreement, Top-10 overlap and the maximum absolute reconstruction residual between the model probability and the SHAP decomposition were recorded (Appendix B). These checks test whether the main ranking survives materially larger computation; they do not turn SHAP values into causal effects or demonstrate complete coverage of every operating condition.

Permutation SHAP does not require linear coefficients, a tree structure or a particular differentiable form, making it suitable for TabPFN's pre-trained Transformer inference. The baseline sample was limited for computational reasons, whereas the enlarged settings were added specifically to assess whether that limitation changed the leading-feature conclusion.

SHAP was applied after the model had produced the test probabilities and did not alter those predictions. Background samples provided the reference distribution, and the explained samples came from the independent test period. Mean absolute SHAP values were used for global ranking; the integrated global explanation combines global importance, resampling variation and the direction of contributions across test samples, while the local waterfall explanation shows how individual features raise or lower the probability for one representative sample.

SHAP values describe the fitted model relative to the selected background, not isolated physical effects [65–67]. Substituting feature values against a background can also produce combinations that are uncommon among strongly dependent descriptors. Attributions may therefore be redistributed among related inputs when the background or feature

representation changes. A high attribution for preceding-window energy indicates model use of that descriptor; it does not establish that changing energy alone would cause a future event.

Some dependencies follow directly from the feature definitions. At fixed per-window duration Δ_h and grid volume V, energy release rate C/Δ_h and energy density C/V are scaled versions of cumulative energy C. Likewise, event rate and event density are scaled versions of event count. Their separate SHAP values do not demonstrate independent temporal or spatial mechanisms. In particular, C/V does not measure within-grid spatial concentration when V is constant. Measures of spatial spread and centroid displacement describe spatial configuration more directly. Interpretation therefore emphasises related groups of monitoring information and the stability of the leading descriptors.

3.4. Controlled Semi-Synthetic Domain-Shift Analysis

Because a compatible independent-mine catalogue was unavailable, we added a controlled stress test to examine how the fixed workflow behaves when the event distribution changes. Four deterministic semi-synthetic catalogues were constructed from the temporal blocks and empirical distributions of the source catalogue. Scenario A introduced mild coordinate transformation, localisation noise and energy perturbation. Scenario B additionally shifted spatial hotspots, changed spatial scales and strengthened the high-energy tail. Scenario C combined stronger temporal clustering, hotspot displacement and energy-distribution change. Scenario D represented monitoring degradation through low-energy-event loss, random event dropout, localisation error and ten simulated outage periods. Each catalogue retained only event time, three-dimensional location and energy. The generated records are not observations from other mines. The seeds and exact transformation parameters are reported in Appendix C.

The same 24 h-per-window/100 m sample-construction procedure and 28 descriptors were applied to each catalogue. A scenario-specific early-period P90 energy cutoff retained the relative definition of a high-energy event under its shifted energy distribution. The six model implementations were fitted on the original-mine training subset. For each semi-synthetic scenario, the first 20% of anchor times formed a local calibration segment and the remaining times formed an evaluation segment, with a 48 h buffer on both sides of the boundary. Direct transfer used the screening threshold selected on the original-mine validation set. A secondary analysis kept the fitted models fixed but reselected the operating threshold on the early target segment. This second analysis tests limited site-specific threshold adjustment rather than model retraining. The scenarios are single predefined realisations, so the analysis describes sensitivity to these controlled shifts and does not provide a sampling interval for cross-mine performance.

4. Results

4.1. Predictive Performance in the Held-Out Period

On the independent test set, TabPFN obtained ROC-AUC 0.8064, PR-AUC 0.5330, F1 0.5612, balanced accuracy 0.7768 and recall 0.7156, the highest observed values among the six classifiers; its Brier score of 0.1190 was the lowest observed value (Figure 3, Table 3). Among the three boosting models, CatBoost was closest to TabPFN, with ROC-AUC 0.6524, PR-AUC 0.3256 and a Brier score of 0.1525.

Relative to CatBoost, TabPFN increased ROC-AUC by 0.1540 and PR-AUC by 0.2074 and reduced the Brier score by 0.0335. Relative to logistic regression, the corresponding differences were 0.0599, 0.1518 and 0.0571. Under the ordinary paired DeLong analysis, the comparisons with the four tree-based models had nominal p-values below 0.001 ($z \geq 3.86$). The comparison with logistic regression did not reach statistical significance ($z = 1.66$,

$p = 0.096$; Figure 3). These results are conditional on the reported configurations and test period. Their uncertainty may be understated because temporal and spatial dependence between samples is not represented in that test.

The 14-day block bootstrap gave TabPFN 95% intervals of 0.7356–0.8660 for ROC-AUC, 0.4330–0.6321 for PR-AUC and 0.0942–0.1491 for Brier score. The paired ROC-AUC difference relative to logistic regression was 0.0599 (95% interval −0.0210 to 0.1280), again showing that the two-model difference is uncertain in this monitoring period. The intervals for the differences from Random Forest, CatBoost, XGBoost and LightGBM remained above zero (Appendix B). Because the block length is an analysis choice and spatial correlation may remain, these intervals are dependence-aware sensitivity results rather than universal uncertainty bounds.

All six models were evaluated on the same 671 held-out grid-time samples. Their probability outputs allow different screening operating points to be considered on validation data. The counts reported below concern future-high-energy labels, not observed rockburst damage.

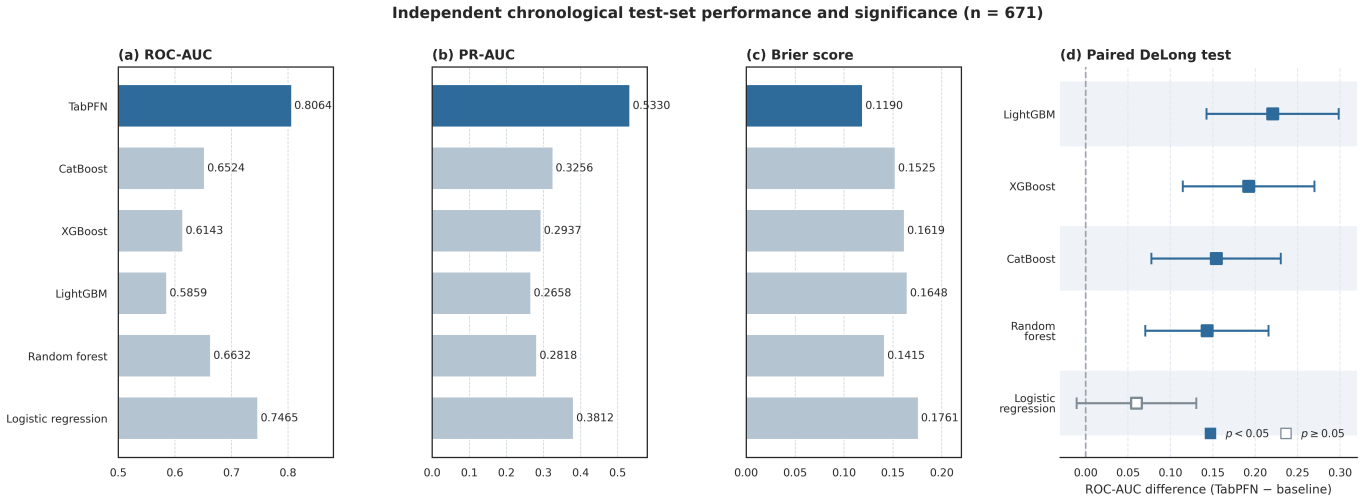


Figure 3. Held-out model comparison: (**a**) ROC-AUC, (**b**) PR-AUC and (**c**) Brier score; (**d**) TabPFN-minus-baseline AUC differences with nominal DeLong 95% intervals. Filled squares denote nominal $p < 0.05$. These intervals do not adjust for between-sample temporal or spatial dependence.

Table 3. Independent chronological test-set performance. The alert threshold for each model was selected on validation data only.

Model	Threshold	ROC-AUC	PR-AUC	Brier	F1	Balanced Accuracy	Precision	Recall
TabPFN	0.1184	0.8064	0.5330	0.1190	0.5612	0.7768	0.4615	0.7156
CatBoost	0.3391	0.6524	0.3256	0.1525	0.3345	0.6045	0.2697	0.4404
XGBoost	0.2246	0.6143	0.2937	0.1619	0.2967	0.5668	0.2057	0.5321
LightGBM	0.3030	0.5859	0.2658	0.1648	0.2968	0.5731	0.2289	0.4220
Random forest	0.2632	0.6632	0.2818	0.1415	0.3106	0.5843	0.2347	0.4587
Logistic regression	0.3754	0.7465	0.3812	0.1761	0.4260	0.6906	0.3144	0.6606

Figure 4 presents discrimination and a visual summary of probability reliability. Its calibration panel is retained as a descriptive plot, whereas Table 4 provides a direct comparison under a common ten-bin, equal-width rule. TabPFN had ECE 0.061 and MCE 0.191. Random Forest had a slightly lower ECE (0.054), but its MCE was higher (0.199), its Brier score was higher and its calibration slope was above one. The other four comparators had ECE values of 0.079–0.112. TabPFN combined the lowest Brier score and MCE with a calibration slope of 0.97, the value closest to the ideal slope of one among the evaluated models. Its

intercept of −0.26 still indicates an overall probability offset, so the outputs should support ranking and review rather than be interpreted as perfectly calibrated event frequencies.

Table 4. Probability calibration on the held-out test period using ten common equal-width bins.

Model	Brier	ECE	MCE	Intercept	Slope
TabPFN	0.1190	0.061	0.191	−0.26	0.97
Logistic regression	0.1761	0.112	0.290	−0.52	0.92
Random forest	0.1415	0.054	0.199	−0.12	1.18
CatBoost	0.1525	0.079	0.240	−0.31	0.92
XGBoost	0.1619	0.091	0.260	−0.43	0.71
LightGBM	0.1648	0.105	0.310	−0.47	0.66

As a descriptive reference, assigning every test sample the test-period prevalence, $\pi = 109/671$, gives a Brier score of $\pi(1-\pi) = 0.1361$. The reported TabPFN value of 0.1190 is approximately 12.5% lower. This prevalence-only reference is calculated retrospectively from the reported test counts, not fitted as a deployable baseline. It provides context for overall probability error but does not separate calibration from discrimination or establish reliability around the alert threshold.

The threshold 0.1184 is used as a screening operating point, not as proof that all cases near that output have the same true event frequency. Although the common-bin analysis provides an overall calibration check, reliability specifically near the operating point would require enough observations in that probability range. Accuracy was not used as the main metric: with only 16.24% positive test samples, always predicting the negative class would yield 83.76% accuracy while detecting no positives [39,59].

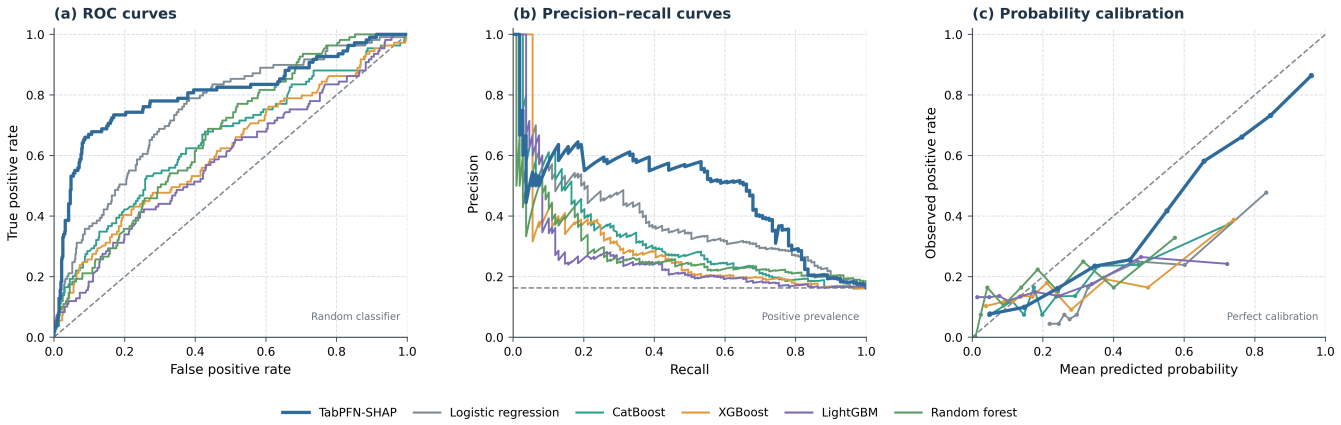


Figure 4. Held-out (**a**) ROC, (**b**) precision–recall and (**c**) calibration curves. Dashed lines indicate random ranking, positive prevalence and ideal calibration, respectively. The heavier navy line identifies TabPFN within the TabPFN-SHAP framework. Common-bin numerical calibration results are reported in Table 4.

Figure 5 brings the dependence-adjusted comparison and the numerical calibration diagnostics together. The paired 14-day block-bootstrap interval was above zero for each tree-based comparator, while the interval for logistic regression crossed zero. Across the calibration summaries, TabPFN had the lowest Brier score and MCE and the slope closest to one. Random Forest retained the lowest ECE and the intercept closest to zero. Thus, the updated probability analysis supports TabPFN's overall reliability advantage without implying that every calibration statistic is uniformly best.

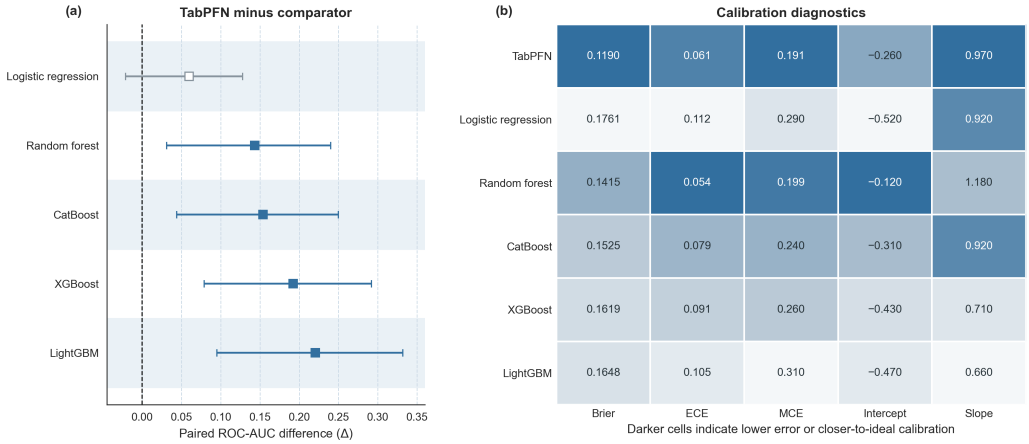


Figure 5. Dependence-adjusted model comparison and probability calibration on the held-out period. (**a**) Paired TabPFN-minus-comparator ROC-AUC differences with 95% confidence intervals from the 14-day temporal block bootstrap; the open marker denotes the comparison whose interval crosses zero. (**b**) Calibration diagnostic matrix. Cells report the original Brier score, ECE, MCE, intercept and slope; darker blue indicates lower error or a value closer to the corresponding ideal (intercept 0; slope 1).

To examine the spatial distribution of model output, Figure 6 aggregates the test-period samples by grid and compares the mean predicted high-energy-event probability with the observed proportion of future high-energy events. The two depth layers with the largest sample counts are shown; they contain 412 and 227 grid-time samples and together cover 95.2% of the test set. Each coloured cell is an observed 100 m test grid rather than an interpolated surface, and the printed count gives its number of contributing grid-time samples. Cells with fewer than five samples are hatched and excluded from the colour scale; cells with no test sample remain white. This makes the uneven spatial support visible when the predicted and observed patterns are compared.

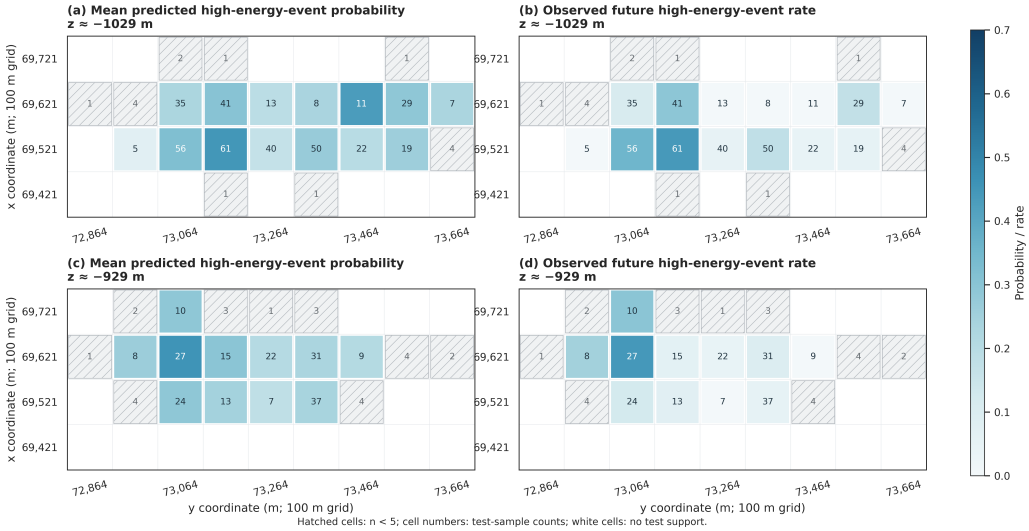


Figure 6. Support-aware test-period spatial summaries at the two most populated depth layers: (**a**,**c**) mean TabPFN high-energy-event probability and (**b**,**d**) observed future-high-energy-label frequency. Colours show the actual aggregate for each populated 100 m grid and cell labels give its test-sample count. Hatched cells have fewer than five samples and are masked from the colour scale; white cells have no test support. No between-grid interpolation is used.

The principal elevated areas show partial spatial agreement, although discrepancies remain at some grids. Because the panel averages over the complete test period, it is a

retrospective support-aware summary rather than a prospective alarm map for a particular day. The masking cutoff is used only for display and does not remove samples from model evaluation. A field display would need to show the current anchor time, data coverage and unscreened grids explicitly.

Table 5 compares the reported operating points. TabPFN flagged 25.2% of the grid-time samples and detected 78 positives, whereas logistic regression flagged 34.1% and detected 72. Thus, the reported comparison has six additional positive samples, 60 fewer flagged samples and 66 fewer false positives for TabPFN. These are grid-time counts; repeated flags may refer to one region or a continuing episode and are not necessarily distinct site incidents or inspection visits. A false positive means that no qualifying high-energy event followed within the specified grid and horizon, not that a site inspection would have been unnecessary. No economic savings or reduction in actual rockburst losses was measured.

The block-bootstrap intervals for TabPFN recall and precision were 0.6164–0.8049 and 0.3811–0.5510, respectively (Appendix B). Reporting both quantities makes the screening trade-off visible: detecting more future-high-energy samples does not mean that every flag is followed by a qualifying event. The alert-rate interval was 0.1975–0.3115 under the same resampling scheme.

Table 5. Operational contrast between TabPFN and the strongest linear baseline at validation-selected alert thresholds.

Model	Alerts	Alert Rate	True Positives	Missed Positives	False Positives	True Negatives
TabPFN	169	25.2%	78	31	91	471
Logistic regression	229	34.1%	72	37	157	405

We also compared the full framework with a single-variable threshold rule under the same split and validation-only threshold procedure.

The best single-variable rule on the validation set was preceding-window cumulative energy, which is also the highest-ranked SHAP feature. On the independent test set, this rule reached ROC-AUC 0.7334, PR-AUC 0.3703, balanced accuracy 0.6880 and recall 0.6055 with 195 alerts (Table 6). Preceding-window event count was the second-best single feature, with ROC-AUC 0.7150. Both descriptors are calculated from the same preceding events and are strongly correlated ($\rho = 0.9599$), so they mainly represent the local microseismic activity intensity over the same period.

Relative to the cumulative-energy rule, the reported multivariable model improved ROC-AUC by 0.0730 and PR-AUC by 0.1627. It produced 26 fewer alerts and 38 fewer false positives while detecting 12 more positive samples (78 versus 66). The single-variable rule was selected from 28 candidates on validation data. The comparison supports evaluating a multivariable state description rather than cumulative energy alone, but does not isolate the incremental benefit of the spatial feature group. That would require a matched feature-group ablation.

Table 6. Comparison between the best single-variable rule and the TabPFN probability model on the independent chronological test set. The single-variable rule uses preceding-window cumulative energy with a threshold selected on validation data only.

Model	ROC-AUC	PR-AUC	Balanced Accuracy	Recall	Alerts	TP	FP
Single feature (preceding-window energy)	0.7334	0.3703	0.6880	0.6055	195	66	129
TabPFN	0.8064	0.5330	0.7768	0.7156	169	78	91

4.2. SHAP Explanation of Energy, Activity and Spatial Features

Figure 7a combines global SHAP importance with resampling stability. In the baseline explanation, preceding-window cumulative energy had a mean absolute SHAP value of 0.0256, followed by preceding-window event count at 0.0219. The next contributors were energy change (0.0134), current-window cumulative energy (0.0109), energy release rate (0.0106), event-count change (0.0098) and energy density (0.0092). The panel compares the baseline values with the mean and standard deviation over three explanation resamples, while Figure 7b shows how these features raise or lower the predicted high-energy-event probability across the explained test samples.

These baseline values come from one explanation run. Across three different stratified explanation samples, preceding-window cumulative energy had an average mean absolute SHAP value of 0.0348. The difference from the baseline value reflects the explained samples; the fitted TabPFN model did not change.

Preceding-window cumulative energy and event count were strongly correlated ($\rho = 0.9599$). They measure total energy and event number in the same preceding window and mainly represent local microseismic activity intensity over that period. They should not be treated as independent physical effects. Taken together, recent activity intensity was the main basis of the TabPFN predictions, with additional information from energy change, release rate, event-count change and energy density.

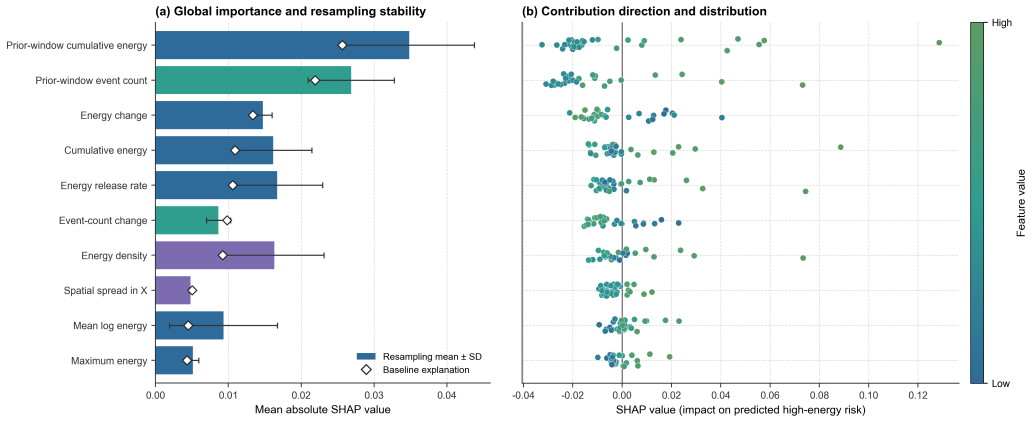


Figure 7. Integrated global permutation-SHAP explanation for TabPFN: (**a**) mean absolute SHAP values and standard deviations for the ten leading features over three explanation-sample resamples, with diamonds indicating the baseline explanation values; (**b**) direction and distribution of feature contributions in the baseline explained samples, with colour representing relative feature value.

Figure 8 illustrates a positive grid-time sample (anchor time 2024-04-13 23:04:11; grid index 3, 2, 2). The probability reported for this explanation was 0.8827, above the reported screening threshold of 0.1184. Preceding-window cumulative energy, current-window cumulative energy, energy release rate, energy density and preceding-window event count made the largest positive contributions in the explanation. The figure shows the model basis of a high-energy-event flag; it does not document a damaging rockburst.

The global ranking summarises contributions across the explained samples, whereas Figure 8 concerns one case. The several energy-related contributions should be interpreted together because they include rescaled versions of the same underlying quantity.

Preceding-window cumulative energy and event count describe earlier activity, while energy change contrasts adjacent windows. At the fixed baseline window and grid size, energy release rate and energy density do not add independent information to cumulative energy. Their attributions cannot separately establish an acceleration of release or an increase in within-grid spatial concentration. The ranking is evidence about the fitted model's use

of correlated descriptors, and small differences between adjacent ranks should not be interpreted as differences in physical effect.

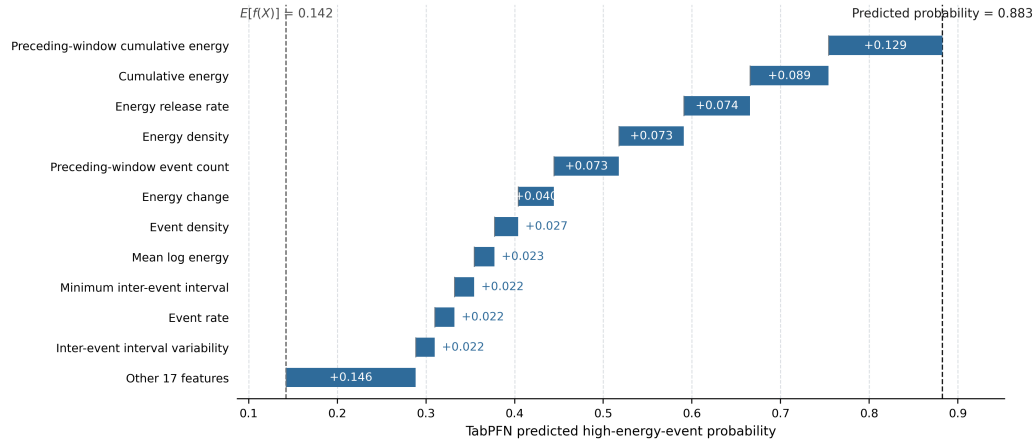


Figure 8. Local SHAP waterfall explanation for a representative positive high-energy-event sample.

4.3. Robustness to Spatiotemporal Scale, Label Definition and SHAP Resampling

The first analysis examined the duration of each of the two adjacent historical windows and the grid size. Five spatiotemporal configurations were compared while retaining the 24 h forecast window, data split, buffer rule and threshold-selection procedure. TabPFN retained the highest observed ROC-AUC among the compared implementations in these settings, so the model ordering was comparatively stable, but its absolute ROC-AUC ranged from 0.7481 to 0.8064 (Figure 9). The baseline setting of 24 h per historical window and a 100 m grid reached 0.8064; the alternatives ranged from 0.7481 to 0.7645, approximately 0.04–0.06 lower. The baseline parameters had been specified before the sensitivity analysis and were not reselected from these results. This noticeable reduction shows that discrimination is sensitive to temporal and spatial aggregation even when the relative model ranking is maintained. With the grid fixed at 100 m, the 12 h- and 48 h-per-window settings differed in ROC-AUC by only 0.0042. Because sample size and positive prevalence vary between configurations, PR-AUC and Brier scores should be interpreted within the corresponding data distribution.

The label analysis used the P85, P90 and P95 calibration-period energy quantiles. TabPFN ROC-AUC was 0.7658, 0.8064 and 0.7784, respectively (Figure 9); within each label definition, it had the highest observed ROC-AUC and PR-AUC and the lowest observed Brier score among the evaluated models. The P95 label produces fewer positives, so its PR-AUC and Brier score should not be compared directly with P85. Similar ranking results under all three definitions indicate that the observed performance is not confined to the P90 label.

Finally, the fitted TabPFN model was held fixed while the SHAP sampling and evaluation budget were varied. In the original three resamples, preceding-window cumulative energy and event count remained first and second, pairwise Spearman correlations of the full rankings were 0.8451–0.9119, and the Top-10 overlap was 7–9 features. In the expanded checks, a 20-case background with 100 explained cases produced rank agreement of 0.901 and a Top-10 overlap of 9; using 50 background cases increased rank agreement to 0.918 with the same overlap. Doubling the evaluation budget gave complete Top-10 overlap, and increasing the explained set to 300 retained rank agreement of 0.909. The largest SHAP reconstruction residual across these configurations was 4×10^{-6} (Appendix B). Within the examined settings, the principal ranking is therefore not an artefact of the original five-case background or 30-case explanation set. The results remain statistical descriptions of model behaviour and do not establish a causal rockburst mechanism.

Figure 10 shows the accuracy–cost trade-off in these checks. Increasing the background, explanation set or evaluation budget retained rank agreement above 0.90 and Top-10 overlap of at least nine, while reconstruction residuals remained at the order of 10^{-6}. Runtime increased from 8 s in the baseline calculation to 372 s for 300 explained samples. The larger calculations therefore reinforced the leading-feature conclusion, but the increase in computation did not produce a materially different ranking.

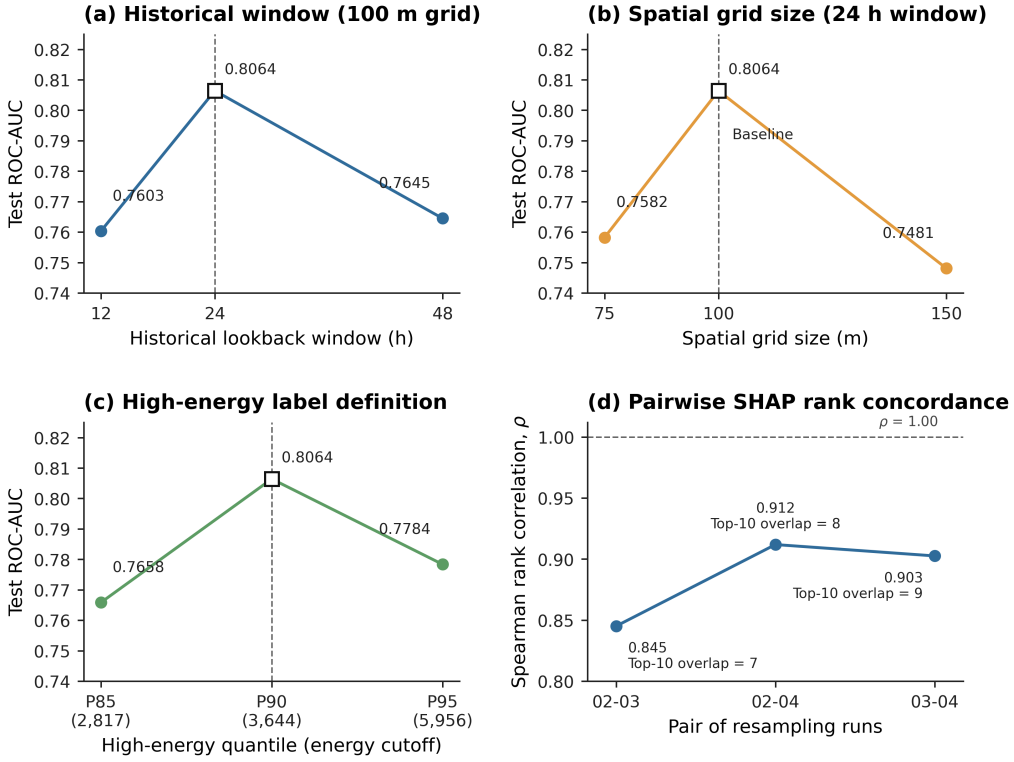


Figure 9. Sensitivity of test ROC-AUC to (**a**) the duration of each of the two adjacent historical windows, (**b**) grid size and (**c**) label threshold; (**d**) SHAP rank correlations across three resamples, annotated with Top-10 overlap. Squares and dashed lines mark the P90/24 h-per-window/100 m baseline in (**a**–**c**).

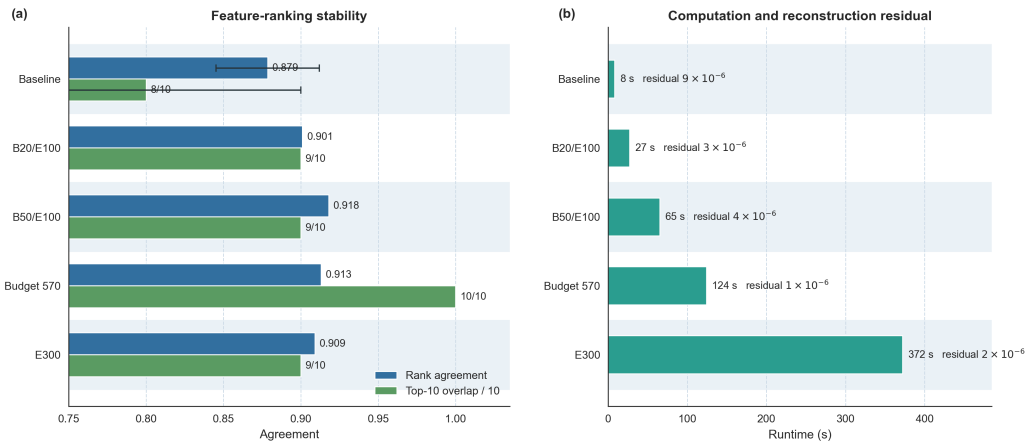


Figure 10. Permutation-SHAP stability under enlarged background, explanation-set and evaluation-budget settings. (**a**) Spearman feature-ranking agreement and Top-10 overlap expressed on a common 0–1 scale; error bars for the baseline denote its observed ranges over three resamples. (**b**) Computation time, with the maximum absolute probability-space reconstruction residual annotated for each setting. B and E denote background and explained-sample counts, respectively.

These checks strengthen the stability assessment but do not establish complete coverage of every grid, operating stage or rare state. A high rank correlation also does not guarantee that an individual case's SHAP magnitude or sign is unchanged. The anonymised high- and low-output decompositions in Appendix B illustrate how the same grouped descriptors can support contrasting predictions without implying a physical intervention effect.

The spatiotemporal and label analyses examine how sample construction affects model performance, while SHAP resampling examines the sensitivity of feature ranking to the explained samples. All results still come from one catalogue, and validation with data from other mines remains necessary (Section 5.2).

Figure 11 reports the Spearman correlation matrix for six high-impact derived descriptors across all 3298 grid-time samples. Preceding-window cumulative energy and preceding-window event count were strongly positively associated ($\rho = 0.9599$), whereas energy change was inversely associated with both descriptors ($\rho = -0.8666$ and -0.8384). Mean log energy had weak associations with the two preceding-state descriptors ($|\rho| \leq 0.0818$), and spatial spread in X showed moderate positive associations with them ($\rho = 0.4602$ and 0.4916).

The matrix provides context for interpreting SHAP. Because attribution is distributed among correlated descriptors, individual SHAP values should not be read as independent causal effects.

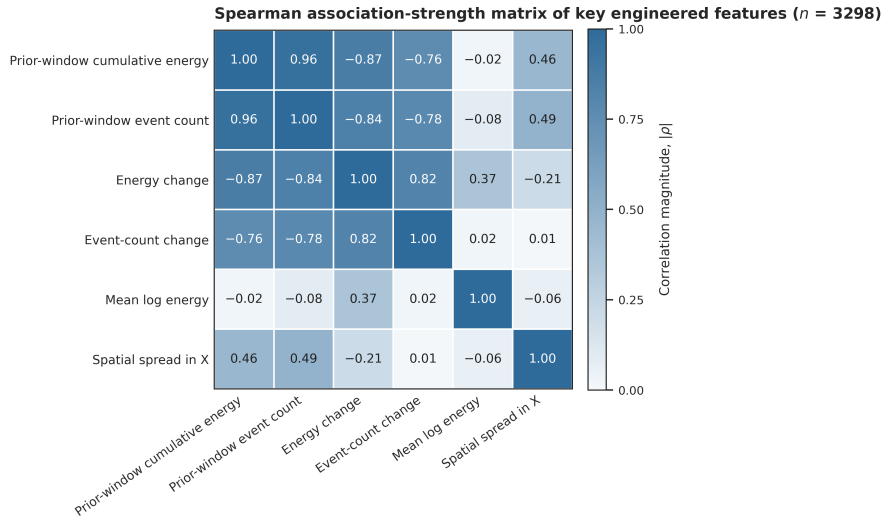


Figure 11. Spearman correlation matrix of six high-impact engineered microseismic features over all 3298 grid-time samples. Cell colour encodes correlation magnitude ($|\rho|$, pale to dark blue), while each cell reports the signed correlation coefficient. The matrix provides dependence context for the TabPFN-SHAP attribution results.

4.4. Performance Under Controlled Semi-Synthetic Distribution Shifts

The later evaluation segments contained 2504, 3433, 4512 and 2057 grid–time samples in Scenarios A–D, with positive rates of 9.98%, 9.82%, 10.22% and 7.88%, respectively. Under direct transfer of the original-mine operating threshold, TabPFN achieved ROC-AUC values of 0.7752, 0.7527, 0.6992 and 0.7703 (Table 7). The corresponding PR-AUC values were 0.3737, 0.2780, 0.2350 and 0.2468. TabPFN ranked first among the six implementations for ROC-AUC, PR-AUC and Brier score in every scenario; it ranked first for balanced accuracy in A–C and second in D (Figure 12).

Table 7. Direct-transfer performance of TabPFN in the four controlled semi-synthetic scenarios.

Scenario	n	Positive Rate	ROC-AUC	PR-AUC	Brier	Balanced Accuracy	Recall
A: mild site shift	2504	0.0998	0.7752	0.3737	0.0788	0.7020	0.6600
B: spatial–energy shift	3433	0.0982	0.7527	0.2780	0.0848	0.6954	0.6825
C: strong distribution shift	4512	0.1022	0.6992	0.2350	0.0935	0.6487	0.5640
D: monitoring degradation	2057	0.0788	0.7703	0.2468	0.0718	0.6846	0.6173

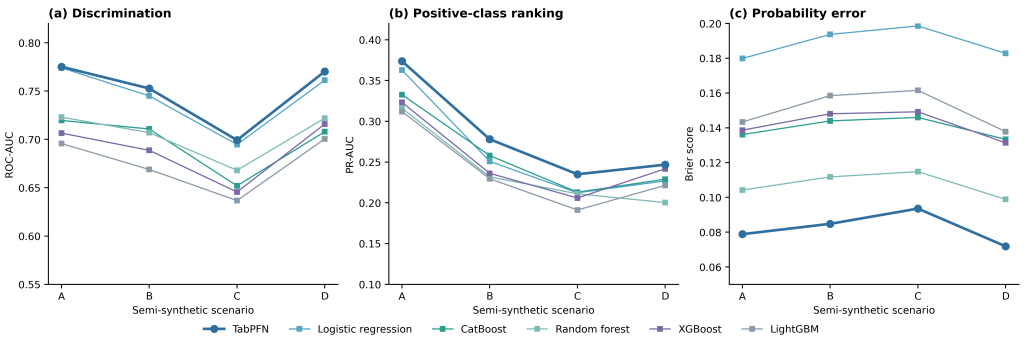


Figure 12. Direct-transfer performance across four controlled semi-synthetic domain-shift scenarios. (**a**) ROC-AUC, (**b**) PR-AUC and (**c**) Brier score for the six fixed model implementations. Lower values are preferable only for the Brier score. A–D denote mild site shift, spatial–energy shift, strong distribution shift and monitoring degradation, respectively.

For descriptive comparison, the ROC-AUC values in A, B and D remained within the 0.7356–0.8660 block-bootstrap interval obtained from the original held-out period, whereas C fell below its lower limit. That interval quantifies uncertainty in the original period and is not a prediction interval for a new domain; the comparison is used only to show the magnitude of degradation. Raw PR-AUC, precision and F1 were lower partly because the semi-synthetic positive rates were 6–8 percentage points below the original test prevalence. Relative to the positive-class prevalence, the PR-AUC lift was 3.74, 2.83, 2.30 and 3.13 in A–D, compared with 3.28 in the original held-out period. The strong-shift scenario therefore produced the clearest reduction in ranking performance.

Local threshold selection changed the TabPFN threshold from 0.1184 to 0.0885, 0.1368, 0.1241 and 0.0985 in A–D. It increased recall from 0.6600 to 0.7480 in A and from 0.6173 to 0.7099 in D, but also increased their alert rates by 0.1018 and 0.0510. Changes in balanced accuracy were −0.0077, +0.0031, +0.0015 and +0.0225. Threshold adjustment did not change ROC-AUC, PR-AUC or Brier score because the predicted probabilities and their ranking were held fixed. These results support local selection of an operating point according to review capacity and missed-event consequences, rather than transferring 0.1184 as a universal cutoff.

5. Discussion

5.1. Interpreting the Main Findings

Rockburst prediction studies commonly organise engineering cases as tabular samples and divide them randomly into training and test sets [5]. The present study instead uses a time-stamped microseismic record from one mine. Neighbouring samples may share historical events and overlapping forecast periods, so random splitting is unsuitable. To follow the field sequence of building a model and then processing later data, we divided the samples chronologically, excluded anchor times within 48 h on either side of each split boundary, and fixed the alert thresholds before testing. Tables 3–6 therefore describe performance on a later monitoring period.

Across the same 671 held-out samples, TabPFN had the highest point estimates for ROC-AUC, PR-AUC, F1 score, balanced accuracy and recall, and the lowest Brier error. Relative to logistic regression, the differences were 0.0599 in ROC-AUC, 0.1518 in PR-AUC and a reduction of 0.0571 in Brier error. Its operating point detected six additional positives with 60 fewer flags and 66 fewer false positives. The paired 14-day block bootstrap supported positive AUC differences from the four tree models, whereas the interval relative to logistic regression crossed zero. The results therefore support TabPFN's practical advantage over the evaluated tree implementations in this period, while showing that its discrimination advantage over the linear baseline is less certain. None of these comparisons demonstrates reduced rockburst losses or superiority across mines.

SHAP makes the monitoring information used by the fitted classifier easier to inspect. Preceding-window energy and event count were the most stable leading descriptors, consistent with the importance of activity measures in mining seismology [30]. Their strong correlation means that the two ranks should be read together. Other energy and activity descriptors also contribute (Figure 7), but scaled versions of cumulative energy do not establish separate physical effects. The single-feature comparison supports multivariable screening; it does not identify which feature group causes the improvement. This distinction preserves the engineering value of an explanation without interpreting it as a causal rockburst mechanism [34,66].

5.2. Engineering Scope and Further Validation

The relation between microseismic activity and potential rockburst scale provides an engineering motivation for screening [2,45], but it does not make the future-high-energy label equivalent to a rockburst observation. Acoustic-emission evidence has likewise described the pre-failure stage as a reorganisation towards more persistent and temporally clustered high-intensity episodes rather than a uniform change in one indicator [27]. Persistent or clustered seismic activity may nevertheless be predictable without a corresponding damaging event. The present data cannot separate that persistence from a precursor specifically associated with rockburst damage.

A proposed field workflow would start with checking sensor operation, event quality, energy units and the coordinate system. At each daily anchor, historical features would be calculated for recently active grids using a fixed site grid. The model would output a probability and the main SHAP contributions for each eligible grid. Flags would enter a review queue in which engineers check excavation advance, stress information, geological structures and support observations before deciding whether additional monitoring or inspection is warranted. Quiet or unobserved grids would remain explicitly unscreened, not marked safe. This is a proposed use of the framework, not a deployment trial completed in this study.

Transfer to another mine requires more than applying the same numerical threshold. Lithology, depth, excavation layout, sensor coverage, detection limits and energy estimation can change both the descriptors and the event frequency [17,47]. The controlled semi-synthetic analysis gives an intermediate stress test: TabPFN retained the strongest discrimination and probability-error results among the compared implementations in all four scenarios, but its ROC-AUC declined from 0.8064 in the original held-out period to 0.6992 under the strongest combined shift. Threshold adjustment improved recall in some scenarios only by increasing the review workload. These results make the likely direction of transfer sensitivity more explicit, but the catalogues inherit statistical structure from the source data and contain prescribed rather than observed site differences. They are therefore not evidence of independent-mine field generalisation.

The high-energy quantile, spatial scale and probability operating point would still require local review, with a later period reserved for evaluation. Independent-mine field validation has not been performed, and the semi-synthetic stress test does not replace it. Adding geological and stress observations may also change the prediction task and would require a new evaluation.

Remaining uncertainties include only 50 validation positives, the choice of a 14-day resampling block and possible spatial dependence not removed by temporal resampling. Consecutive anchors and neighbouring grids in the held-out period should therefore be viewed as correlated monitoring decisions, not independent incidents. Although the calibration slope was close to one, the intercept still indicates an overall offset, and reliability around the selected alert threshold remains uncertain. The enlarged SHAP analysis supports the principal ranking under the examined computation budgets, but it cannot show that all rare operating conditions are represented.

A prospective evaluation should preserve the predictions at each anchor, record which grids were eligible and observed, and link the resulting review decisions to later microseismic and damage observations. This would allow probability reliability, screening coverage and actual engineering outcomes to be assessed separately. Independent-mine validation is also required before transferring the numerical threshold or reported performance to another site.

6. Conclusions

This study connects basic microseismic event records to interpretable screening of future high-energy activity. Using event time, three-dimensional location and energy, 9116 records were reconstructed as 3298 grid-time samples with 28 descriptors. The methodological contribution is the combination of historical state construction, pre-trained tabular inference and inspection of feature contributions in a chronological evaluation. The output concerns recently active grids and is intended to support rockburst-related site review, not to predict damage grades directly.

The test comparison gives TabPFN ROC-AUC 0.8064, PR-AUC 0.5330 and Brier score 0.1190, respectively the highest, highest and lowest point estimates among the six listed implementations. Its operating point detected six additional positives with 60 fewer flags and 66 fewer false positives than logistic regression. Paired 14-day block-bootstrap results support its AUC advantage over the four evaluated tree models, while the interval against logistic regression crosses zero. Common-bin calibration and enlarged SHAP checks further show that the model's probability quality and leading descriptor ranking are not conclusions drawn from Brier error or one small explanation sample alone.

The engineering value of the framework is to accompany each screening output with monitoring quantities that can be checked before action. Preceding-window energy and event count led the reported SHAP rankings, indicating model reliance on recent activity rather than independent causal effects. The relative model ordering remained comparatively stable under the tested history-window and grid-size changes, but TabPFN ROC-AUC fell from 0.8064 at the baseline scale to 0.7481–0.7645 in the alternative settings; the aggregation scale therefore matters to absolute performance. In the controlled semi-synthetic analysis, TabPFN retained the highest ROC-AUC and PR-AUC among the six implementations across all four scenarios, while the decline to ROC-AUC 0.6992 under the strongest combined shift further shows why site-specific checking is necessary. A practical application would combine these explanations with data-quality checks and site observations, leave unsupported grids unscreened, and choose thresholds according to local review capacity and missed-event consequences. These semi-synthetic results are robustness evidence, not independent-mine validation; prospective use and cross-mine transfer remain to be tested with field data.

Author Contributions: Conceptualization, F.J. and H.J.; methodology, F.J., H.J. and Y.Z.; software, F.J.; validation, F.J., Y.W. and Q.G.; formal analysis, F.J. and H.J.; investigation, F.J. and Y.Z.; resources, F.J. and Z.L.; data curation, F.J. and G.S.; writing—original draft preparation, F.J.; writing—review and editing, H.J., Y.Z., Y.W., Q.G., G.S. and Z.L.; visualization, F.J.; supervision, H.J.; project administration, F.J. and H.J.; funding acquisition, F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China National Coal Group Key Technology Project, grant number 20221CY001.

Data Availability Statement: The microseismic event data and sample-level model outputs are not publicly available due to confidentiality requirements and data ownership restrictions of the mine. Aggregated performance, calibration and SHAP-stability results required to assess the conclusions are reported in the manuscript and appendices. Requests concerning additional available material may be directed to the corresponding author and are subject to approval by the data owner.

Conflicts of Interest: Authors Feng Jin, Yubao Zhao, Yatian Wang, and Qi Guo were employed by China National Coal Group Corporation. The research was funded by the China National Coal Group Key Technology Project (grant number 20221CY001), and the mine data owner provided access to the monitoring data and relevant engineering information. The funder had no additional role in the analysis or interpretation of the data, the writing of the manuscript, or the decision to publish the results beyond the contributions of the named authors. The remaining authors declare no conflicts of interest.

Appendix A. Derived Feature Definitions

For a grid-time sample, let $\mathcal{H}$ denote the events in the current 24 h history window and $\mathcal{H}^-$ the events in the preceding 24 h window; let E_i be event energy, $\mathbf{u}_i = (x_i, y_i, z_i)$ the event coordinates, $\bar{\mathbf{u}}$ the arithmetic centroid, $\bar{\mathbf{u}}_E$ the energy-weighted centroid, $V = (100\ \mathrm{m})^3$ the cell volume, $\Delta_h = 24$ h and E_q the high-energy threshold. Table A1 defines all 28 features.

Table A1. Definitions of the 28 derived features used by the high-energy-event prediction model.

Feature	Symbol	Definition
Current-window event count	N	$\lvert\mathcal{H}\rvert$, number of events in the current window.
Event rate (current)	r	N/Δ_h.
Cumulative energy	C	$\sum_{i\in\mathcal{H}} E_i$.
Maximum energy	$E_{\max}$	$\max_{i\in\mathcal{H}} E_i$.
P90 energy	E_{90}	90th percentile of $\{E_i : i \in \mathcal{H}\}$.
Mean log energy	$\overline{\ln E}$	$\frac{1}{N}\sum_{i\in\mathcal{H}} \ln(1+E_i)$.
Log-energy variability	$s_{\ln E}$	standard deviation of $\{\ln(1+E_i) : i \in \mathcal{H}\}$.
Energy concentration	κ	$E_{\max}/C$.
High-energy count	n_q	$\lvert\{i \in \mathcal{H} : E_i \geq E_q\}\rvert$.
High-energy ratio	p_q	n_q/N.
Energy release rate	r_E	C/Δ_h.
Preceding-window event count	N^-	$\lvert\mathcal{H}^-\rvert$.
Preceding-window cumulative energy	C^-	$\sum_{i\in\mathcal{H}^-} E_i$.
Event-count change	ΔN	$(N - N^-)/(N^- + 1)$.
Energy change	ΔC	$(C - C^-)/(C^- + 1)$.
Mean inter-event interval	$\overline{\tau}$	mean of consecutive event time gaps in $\mathcal{H}$; defaults to Δ_h when fewer than two events.
Minimum inter-event interval	$\tau_{\min}$	minimum of those gaps; defaults to Δ_h.

Table A1. *Cont.*

Feature	Symbol	Definition
Inter-event interval variability	c_v	ratio of the standard deviation to the mean of the consecutive gaps in $\mathcal{H}$; 0 if undefined.
Spatial spread in x	s_x	standard deviation of $\{x_i : i \in \mathcal{H}\}$.
Spatial spread in y	s_y	standard deviation of $\{y_i : i \in \mathcal{H}\}$.
Spatial spread in z	s_z	standard deviation of $\{z_i : i \in \mathcal{H}\}$.
RMS spatial spread	d_{rms}	$\sqrt{\frac{1}{N}\sum_i \lVert\mathbf{u}_i - \bar{\mathbf{u}}\rVert^2}$.
Maximum spatial spread	$d_{\max}$	$\max_i \lVert\mathbf{u}_i - \bar{\mathbf{u}}\rVert$.
Energy-weighted spread	d_E	$\sqrt{\frac{\sum_i E_i \lVert\mathbf{u}_i - \bar{\mathbf{u}}_E\rVert^2}{\sum_i E_i}}$.
Event density	ρ_n	N/V.
Energy density	ρ_E	C/V.
Energy-centroid offset	δ_E	$\lVert\bar{\mathbf{u}}_E - \bar{\mathbf{u}}\rVert$.
Centroid migration	δ_{mig}	$\lVert\bar{\mathbf{u}}_{\mathcal{H}} - \bar{\mathbf{u}}_{\mathcal{H}^-}\rVert$; displacement of arithmetic event centroids; 0 if $\mathcal{H}^-$ is empty.

Appendix B. Additional Dependence, Calibration and Explanation Checks

Table A2 reports 14-day temporal-block-bootstrap intervals for TabPFN. The 1980 valid paired resamples used the same block indices for every model; 20 requested resamples were invalid because a required metric could not be calculated. These intervals retain temporal ordering within the sampled blocks but do not remove all possible spatial dependence.

Table A2. TabPFN test estimates and 95% percentile intervals from the 14-day paired block bootstrap.

Metric	Estimate	Lower Limit	Upper Limit
ROC-AUC	0.8064	0.7356	0.8660
PR-AUC	0.5330	0.4330	0.6321
Brier score	0.1190	0.0942	0.1491
F1 score	0.5612	0.4714	0.6410
Balanced accuracy	0.7768	0.7127	0.8362
Precision	0.4615	0.3811	0.5510
Recall	0.7156	0.6164	0.8049
Alert rate	0.2519	0.1975	0.3115

Table A3. Paired TabPFN-minus-baseline ROC-AUC differences from the 14-day block bootstrap.

Comparator	Difference	Lower limit	Upper Limit
Logistic regression	0.0599	−0.0210	0.1280
Random forest	0.1432	0.0310	0.2400
CatBoost	0.1540	0.0440	0.2500
XGBoost	0.1921	0.0790	0.2920
LightGBM	0.2205	0.0950	0.3320

Table A4. Expanded permutation-SHAP stability checks. Each configuration used three repetitions.

Configuration	Background	Explained	Max Evaluations	Rank Agreement	Top-10 Overlap
Baseline	5	30	57	0.8451–0.9119	7–9
Expanded background	20	100	285	0.901	9
Larger background	50	100	285	0.918	9
Larger evaluation budget	50	100	570	0.913	10
Larger explanation set	50	300	570	0.909	9

The maximum absolute reconstruction residual was 9×10^{-6} for the baseline setting and no more than 4×10^{-6} for the expanded settings. Table A5 gives two anonymised probability-space decompositions from the expanded explanation output. They supplement, rather than replace, the representative case in Figure 8.

Table A5. Grouped signed SHAP contributions for anonymised high- and low-output examples.

Descriptor or Quantity	High-Output Case	Low-Output Case
Preceding-window cumulative energy	0.250	−0.040
Preceding-window event count	0.180	−0.030
Energy change	0.100	−0.010
Current-window cumulative energy	0.080	−0.010
Energy release rate	0.060	−0.008
Event-count change	0.040	−0.007
Other 22 descriptors, combined	−0.050	−0.015
Background probability	0.180	0.180
Sum of SHAP contributions	0.660	−0.120
Reconstructed probability	0.840	0.060

Appendix C. Semi-Synthetic Scenario Parameters

Tables A6–A8 report the deterministic settings used to generate the four domain-shift catalogues. Moving blocks were sampled from the source event sequence before the listed temporal, spatial, energy and monitoring transformations were applied. These settings define controlled perturbations of the source distribution; they do not describe four observed mines.

Table A6. Temporal-resampling settings for the semi-synthetic scenarios.

Scenario	Seed	Block Length (d)	Event-Count Factor	Activity Bias	Time Jitter (min)	Time Trend
A: mild site shift	20261011	30	0.96	0.00	8	0.08
B: spatial–energy shift	20261012	10	1.14	0.45	15	0.22
C: strong distribution shift	20261013	5	1.28	1.00	28	0.55
D: monitoring degradation	20261014	14	1.10	0.00	12	−0.10

Table A7. Spatial transformations for the semi-synthetic scenarios. Values in parentheses are ordered as stated in the column heading.

Scenario	Rotation (°)	Scale (x, y, z)	Translation (x, y, z) (m)	Cluster Shift (m)	Location Noise (m)
A	7	(1.05, 0.96, 1.03)	(180, −140, −25)	10	6
B	18	(1.18, 0.86, 1.10)	(−260, 220, 35)	28	12
C	31	(1.30, 0.78, 1.18)	(420, −360, 70)	58	20
D	9	(1.04, 1.02, 0.98)	(−120, −180, −15)	12	26

Table A8. Energy and monitoring transformations for the semi-synthetic scenarios.

Scenario	Energy Scale	Log-Energy Spread	Log-Energy Noise	Tail Amplification	Detection Floor	Random Dropout	Outages
A	0.90	0.92	0.08	0.00	1.00	0.00	0
B	1.15	1.10	0.20	0.18	1.00	0.00	0
C	1.32	1.19	0.32	0.34	1.00	0.00	0
D	0.94	1.02	0.18	0.02	0.36	0.08	10 (30–84 h)

References

1. Ma, T.-H.; Tang, C.-A.; Tang, S.-B.; Kuang, L.; Yu, Q.; Kong, D.-Q.; Zhu, X. Rockburst mechanism and prediction based on microseismic monitoring. *Int. J. Rock Mech. Min. Sci.* **2018**, *110*, 177–188. [CrossRef]
2. Askaripour, M.; Saeidi, A.; Rouleau, A.; Mercier-Langevin, P. Rockburst in underground excavations: A review of mechanism, classification, and prediction methods. *Undergr. Space* **2022**, *7*, 577–607. [CrossRef]
3. Zhou, J.; Li, X.; Mitri, H.S. Evaluation method of rockburst: State-of-the-art literature review. *Tunn. Undergr. Space Technol.* **2018**, *81*, 632–659. [CrossRef]
4. Cai, W.; Dou, L.; Zhang, M.; Cao, W.; Shi, J.; Feng, L. A fuzzy comprehensive evaluation methodology for rock burst forecasting using microseismic monitoring. *Tunn. Undergr. Space Technol.* **2018**, *80*, 232–245. [CrossRef]
5. Zhou, J.; Li, X.; Mitri, H.S. Classification of rockburst in underground projects: Comparison of ten supervised learning methods. *J. Comput. Civ. Eng.* **2016**, *30*, 04016003. [CrossRef]
6. Zhao, H.; Chen, B.; Zhu, C. Decision tree model for rockburst prediction based on microseismic monitoring. *Adv. Civ. Eng.* **2021**, *2021*, 8818052. [CrossRef]
7. Li, N.; Feng, X.; Jimenez, R. Predicting rock burst hazard with incomplete data using Bayesian networks. *Tunn. Undergr. Space Technol.* **2017**, *61*, 61–70. [CrossRef]
8. Yin, X.; Liu, Q.; Huang, X.; Pan, Y. Real-time prediction of rockburst intensity using an integrated CNN-Adam-BO algorithm based on microseismic data and its engineering application. *Tunn. Undergr. Space Technol.* **2021**, *117*, 104133. [CrossRef]
9. Di, Y.; Wang, E.; Li, Z.; Liu, X.; Huang, T.; Yao, J. Comprehensive early warning method of microseismic, acoustic emission, and electromagnetic radiation signals of rock burst based on deep learning. *Int. J. Rock Mech. Min. Sci.* **2023**, *170*, 105519. [CrossRef]
10. Liu, C.; Wang, E.; Li, Z.; Zang, Z.; Li, B.; Yin, S.; Zhang, C.; Liu, Y.; Wang, J. Research on multi-factor adaptive integrated early warning method for coal mine disaster risks based on multi-task learning. *Reliab. Eng. Syst. Saf.* **2025**, *260*, 111002. [CrossRef]
11. Zhang, Y.; Fang, K.; He, M.; Liu, D.; Wang, J.; Guo, Z. Rockburst prediction using artificial intelligence techniques: A review. *Rock Mech. Bull.* **2024**, *3*, 100129. [CrossRef]
12. Basnet, P.M.S.; Mahtab, S.; Jin, A. A comprehensive review of intelligent machine learning based predicting methods in long-term and short-term rock burst prediction. *Tunn. Undergr. Space Technol.* **2023**, *142*, 105434. [CrossRef]
13. Wang, H.; Xu, L.; Yu, H.; Zhang, J. Research on prediction of high energy microseismic events in rock burst mines based on BP neural network. *Sci. Rep.* **2024**, *14*, 29934. [CrossRef]
14. Kadkhodaei, M.H.; Ghasemi, E. Interpretable real-time monitoring of short-term rockbursts in underground spaces based on microseismic activities. *Sci. Rep.* **2025**, *15*, 911. [CrossRef]
15. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
16. Feng, G.L.; Feng, X.T.; Chen, B.R.; Xiao, Y.X.; Yu, Y. A microseismic method for dynamic warning of rockburst development processes in tunnels. *Rock Mech. Rock Eng.* **2015**, *48*, 2061–2076. [CrossRef]
17. Ma, T.-H.; Tang, C.-A.; Liu, F.; Zhang, S.-C.; Feng, Z.-Q. Microseismic monitoring, analysis and early warning of rockburst. *Geomat. Nat. Hazards Risk* **2021**, *12*, 2956–2983. [CrossRef]
18. Melati, S.; Wattimena, R.K.; Sahara, D.P.; Simangunsong, G.M.; Wibowo, A.; Hidayat, W.; Riyanto, E. Modeling cave propagation in deep block caving by incorporating the fracture zone inferred from seismic tomography. *Min. Miner. Depos.* **2025**, *19*, 87–97. [CrossRef]
19. Masethe, R.T.; Manzi, M.S.D.; Durrheim, R.J. Using legacy 3D seismic data and source parameters of mining-induced earthquakes to mitigate the risk of rockbursting in Kloof Gold Mine, South Africa. *Geophys. Prospect.* **2023**, *71*, 1281–1311. [CrossRef]
20. Dzhangirov, A.N.; Bekbaltina, G.K.; Umirova, G.K.; Temirkhanova, R.G.; Abdoldina, F.N. Comparative analyzing the technology of predicting reservoir properties according to seismic data based on linear and nonlinear prediction algorithms. *Eng. J. Satbayev Univ.* **2024**, *146*, 54–60. [CrossRef]
21. Zhao, S.; Wang, E.; Wang, J.; Jia, H.; Zhang, Q.; Li, Z. Multi-level fusion early warning of rock burst utilizing acoustic emission and electromagnetic radiation chaotic prediction and image feature mining: Case studies. *Rock Mech. Rock Eng.* **2025**, *58*, 14021–14044. [CrossRef]
22. Di, Y.; Wang, E.; Huang, T. Identification method for microseismic, acoustic emission and electromagnetic radiation interference signals of rock burst based on deep neural networks. *Int. J. Rock Mech. Min. Sci.* **2023**, *170*, 105541. [CrossRef]
23. Zhan, K.; Wen, X.; Xu, R.; Wang, X.; Wang, C.; Song, P. LGF SeismoLocator: A deep learning model for precision microseismic event localization in coal mines. *Rock Mech. Rock Eng.* **2024**, *57*, 10717–10730. [CrossRef]
24. Liu, C.; Wang, E.; Wang, Y.; Li, Z.; Xu, J. Online early warning of coal and gas outbursts through reliability-constrained evidence fusion of acoustic emission, electromagnetic radiation, and gas concentration. *Reliab. Eng. Syst. Saf.* **2026**, *275*, 112772. [CrossRef]
25. Liu, C.; Wang, E.; Li, Z.; Xu, J.; Yin, S. Coal-gas risk fusion identification and warning based on deep explainable expert learning networks: A case study in mining information systems. *Inf. Process. Manag.* **2026**, *63*, 104665. [CrossRef]

26. Liu, C.; Wang, E.; Li, B.; Li, Z.; Wang, Y. A knowledge-augmented generative-recognition-warning framework for precursor risk identification of coal and gas outbursts. *Adv. Eng. Inform.* **2026**, *74*, 104617. [CrossRef]
27. Liu, Z.; Wang, E.; Tan, T.; Li, B. Entropy-driven characterization of acoustic emission precursors for rockbursts: A systematic comparison of four entropy metrics and early warning verification. *J. Appl. Geophys.* **2026**, *255*, 106521. [CrossRef]
28. Tang, S.; Wang, J.; Tang, L.; Ding, S. Automatic early warning of rockbursts from microseismic events by learning the feature-augmented point cloud representation. *Tunn. Undergr. Space Technol.* **2024**, *147*, 105692. [CrossRef]
29. Li, X.; Mao, H.; Li, B.; Xu, N. Dynamic early warning of rockburst using microseismic multi-parameters based on Bayesian network. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 715–727. [CrossRef]
30. Mendecki, A.J. *Seismic Monitoring in Mines*; Chapman & Hall: London, UK, 1997.
31. Gutenberg, B.; Richter, C.F. Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.* **1944**, *34*, 185–188. [CrossRef]
32. Schorlemmer, D.; Wiemer, S.; Wyss, M. Variations in earthquake-size distribution across different stress regimes. *Nature* **2005**, *437*, 539–542. [CrossRef]
33. Yin, S.; Wang, E.; Li, Z.; Zang, Z.; Liu, X.; Zhang, C.; Ding, X.; Aihemaiti, A. Multifractal and b-value nonlinear time-varying characteristics of acoustic emission for coal with different impact tendency. *Measurement* **2025**, *248*, 116896. [CrossRef]
34. Yu, Q.; Zhao, D.; Xia, Y.; Jin, S.; Zheng, J.; Meng, Q.; Mu, C.; Zhao, J. Multivariate early warning method for rockburst monitoring based on microseismic activity characteristics. *Front. Earth Sci.* **2022**, *10*, 837333. [CrossRef]
35. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A Transformer that solves small tabular classification problems in a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 1–5 May 2023.
36. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef]
37. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 18932–18943.
38. Borisov, V.; Leemann, T.; Sessler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519. [CrossRef]
39. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
40. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 1135–1144.. [CrossRef]
41. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 93:1–93:42. [CrossRef]
42. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774.
43. Rundel, D.; Kobialka, J.; von Crailsheim, C.; Feurer, M.; Nagler, T.; Rügamer, D. Interpretable machine learning for TabPFN. *arXiv* **2024**, arXiv:2403.10923. [CrossRef]
44. Basnet, P.M.S.; Jin, A.; Mahtab, S. Developing an explainable rockburst risk prediction method using monitored microseismicity based on interpretable machine learning approach. *Acta Geophys.* **2024**, *72*, 2597–2618. [CrossRef]
45. Liu, G.-F.; Jiang, Q.; Feng, G.-L.; Chen, D.-F.; Chen, B.-R.; Zhao, Z.-N. Microseismicity-based method for the dynamic estimation of the potential rockburst scale during tunnel excavation. *Bull. Eng. Geol. Environ.* **2021**, *80*, 3605–3628. [CrossRef]
46. Li, Z.; He, S.; Song, D.; He, X.; Dou, L.; Chen, J.; Liu, X.; Feng, P. Microseismic temporal-spatial precursory characteristics and early warning method of rockburst in steeply inclined and extremely thick coal seam. *Energies* **2021**, *14*, 1186. [CrossRef]
47. Lai, X.; Jia, C.; Cui, F.; Chen, J.; Zhou, Y.; Feng, G.; Gao, Y. Microseismic energy distribution and impact risk analysis of complex heterogeneous spatial evolution of extra-thick layered strata. *Sci. Rep.* **2022**, *12*, 10832. [CrossRef]
48. Zhang, H.; Li, H.; Zhang, X.; Xue, S.; Ying, H.; Li, A.; Zheng, J.; Shi, X. Supervised and unsupervised general framework for rockburst risk prediction based on feature contrast of long-term and short-term microseismic data. *Geomech. Geophys. Geo-Energy Geo-Resour.* **2025**, *11*, 139. [CrossRef]
49. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
50. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
51. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
52. 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*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]

53. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 3146–3154.
54. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 6638–6648.
55. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
56. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 507–520. [CrossRef]
57. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
58. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*; ACM: New York, NY, USA, 2006; pp. 233–240. [CrossRef]
59. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
60. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
61. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
62. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
63. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [CrossRef]
64. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef]
65. Aas, K.; Jullum, M.; Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* **2021**, *298*, 103502. [CrossRef]
66. Janzing, D.; Minorics, L.; Blöbaum, P. Feature relevance quantification in explainable AI: A causal problem. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*; PMLR: Cambridge, MA, USA, 2020; Volume 108, pp. 2907–2916.
67. Kumar, I.E.; Venkatasubramanian, S.; Scheidegger, C.; Friedler, S. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2020; Volume 119, pp. 5491–5500.

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