

False-Alarm-Controlled Evaluation of Gas Indices and Deep Learning for Early Detection of Coal Spontaneous Combustion

ULAS CINAR

ulas.cinar@comu.edu.tr

Canakkale Onsekiz Mart University

Article

Keywords: Coal Spontaneous Combustion, Index Gases, Early Warning, False-Alarm Rate, Underground Coal Mining, Mine Safety

Posted Date: August 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10667830/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

False-Alarm-Controlled Evaluation of Gas Indices and Deep Learning for Early Detection of Coal Spontaneous Combustion

ULAS CINAR^{1*}

¹Department of Occupational Health and Safety, Canakkale Onsekiz Mart University, Canakkale, Turkey.

* Correspondence: ulas.cinar@comu.edu.tr; Tel.: +90-535-332-6891, ORCID: 0000-0003-3924-0768

Abstract

Gas ratios such as Graham's index remain the principal field diagnostics for coal spontaneous combustion, yet their reliability is rarely assessed against monitoring periods in which no heating develops. Reported detection rates and lead times are therefore uninterpretable, because any alarm rule can be made to look prescient by alarming continuously. We propose a false-alarm-controlled protocol in which every method, index threshold or neural network, is reduced to a continuous alarm score and compared at a matched false-alarm rate. It is developed on a stochastic simulation of 210 episodes in which coal temperature is the latent state generating the gas observations, then applied unchanged to 18 field episodes from three coal mines: three confirmed fires, fifteen event-free records. In both settings Graham's index detects no fire at any threshold whose false-alarm rate is at or below 0.20. In the field its ordering inverts, exceeding 24 in seven of thirteen event-free ventilated episodes while staying below 0.50 in two sealed-area fires, because the oxygen deficiency in its denominator is near zero in ventilated roadways and above 18 behind seals. At a matched false-alarm rate a carbon-monoxide threshold detected all three fires with 345 h median lead time; attention-LSTM models detected two.

Keywords: Coal Spontaneous Combustion; Index Gases; Early Warning; False-Alarm Rate; Underground Coal Mining, Mine Safety.

1. Introduction

1.1 Coal spontaneous combustion and its gas-phase diagnosis

Spontaneous combustion is among the oldest and most persistent hazards in underground coal mining and remains a leading cause of unplanned panel sealing, production loss and mine closure. Uncontrolled low-temperature oxidation of coal is simultaneously a safety hazard and a substantial environmental burden: coal-seam fires burn for years or decades across China, India, the United States, Indonesia and South Africa, releasing heat, particulate matter and greenhouse gases^{1,2}, with individual fire zones emitting thousands of tonnes of CO₂, SO₂ and CO over their active lifetime³ and a poorly captured share of anthropogenic greenhouse gas attributed to underground coal-mine fires⁴⁻⁶.

Because self-heating develops slowly and without visible signs, detection depends almost entirely on the composition of the ventilation air. For more than a century the primary field diagnostic has been the interpretation of that composition through empirical index gases and ratios, above all Graham's index, defined as the ratio of carbon monoxide to oxygen deficiency, together with the CO/CO₂ and Jones-Trickett ratios. Singh et al.⁷ give the standard modern account, emphasising that no single ratio should be interpreted in isolation; Liang et al.⁸ review the wider family of index gases, and Liang et al.⁹ survey combustion mechanisms and control once a fire is established. A parallel literature characterises the intrinsic liability of a coal to self-heat from rank, petrography and mineral content¹⁰⁻¹⁵ and relates low-temperature oxidation to standard oxygen-consumption measurements¹⁶. These indices are embedded in mine safety regulation and routine practice because they are physically motivated, trivial to compute and require no infrastructure beyond standard gas monitoring.

An often-overlooked property of these indices is their numerical conditioning. Graham's index and the Jones-Trickett ratio both divide by the oxygen deficiency, a quantity that is by construction close to zero whenever the atmosphere approaches fresh air. In the baseline regime, precisely the regime in which false alarms are generated, the ratios are therefore ill-conditioned, and small sensor errors in O₂ or N₂ produce large excursions. This property, rather than any deficiency of the underlying oxidation chemistry, turns out to explain the behaviour reported below.

1.2 Data-driven early-warning models

Statistical learning entered the field through liability and susceptibility prediction: Zhang et al.¹⁷ predicted self-ignition tendency from coal properties, Shukla et al.¹⁸ compared learners for seam-level susceptibility, Qi et al.¹⁹ predicted borehole spontaneous combustion, and Zhuo et al.²⁰ applied particle-swarm-optimised gradient boosting to combustion temperature. Sequence modelling followed. Wang et al.²¹ used a sparrow-search-optimised convolutional network for temperature prediction; Yang et al.²² combined particle-swarm optimisation with an LSTM for gas-concentration forecasting; Ding et al.²³ reported a Bayesian-optimised LSTM; Shao et al.²⁴ coupled a convolutional network with a bidirectional GRU and a convolutional attention block, reporting a coefficient of determination of 0.974; and Xing et al.²⁵ fused multiple sensor channels in a CNN-BiLSTM with attention. Su et al.²⁶ constrained an LSTM with the convection-diffusion equation governing gas transport, Danish and Onder²⁷ mapped gas and temperature variables to fire intensity through a Mamdani fuzzy inference system, and Zhao et al.²⁸ added post-hoc interpretability through SHAP attribution of a stacked ensemble. Reviews by Onifade and Genc¹² and Deng et al.²⁹ survey this trajectory. Reported accuracy is consistently strong, with coefficients of determination above 0.95 now commonplace.

1.3 The evaluation gap

This progress is real but it is measured in a way that does not answer the operational question. A mine does not need a model that reconstructs a temperature curve accurately; it needs an alarm that fires early enough to permit intervention, whether by ventilation adjustment, inertisation, sealing or withdrawal, and rarely enough that the responsible engineer does not learn to discount it. These are different quantities, and the second is almost never reported. Across the studies summarised in Supplementary Table S1, evaluation is by regression error, by classification accuracy at an unreported operating point, or by lead time measured on confirmed events only. None quantifies how frequently the same method would raise an alarm during a monitoring period in which no heating develops.

The omission matters for a reason that is normative as well as statistical. Any alarm rule can be made to appear arbitrarily prescient by lowering its threshold, at the cost of alarming continuously; a method that alarms on nine out of ten quiet periods will, on the events that do occur, appear to have detected them very early. Without the false-alarm rate a reported lead time therefore carries almost no information. Process-safety practice makes the same point in definitional terms: ANSI/ISA-18.2 and the companion standard IEC 62682 define an alarm as an indication to the operator of a condition that requires a response^{30,31}, so a signal that activates when no response is warranted is not a poorly performing alarm but formally not an alarm at all. EEMUA Publication 191, issued after the Piper Alpha explosion and now the reference for alarm-system design in the United Kingdom, Europe and Australia³², treats the elimination of nuisance alarms as a safety objective in its own right, because an indicator known to be usually spurious is discounted on the occasion when it is not. A gas-monitoring installation feeding a mine control room is an alarm system in this sense, and a proposed early-warning method is a candidate alarm rule.

1.4 Contributions

Rather than proposing another architecture, this article proposes and demonstrates an evaluation protocol, and uses it to re-examine the comparison between classical indices and deep sequence models. The contributions are: a false-alarm-controlled benchmarking protocol in which every candidate method, whether a fixed index threshold or a neural classifier, is reduced to a continuous alarm score, swept across its full operating range and compared at a matched false-alarm rate with bootstrap confidence intervals; a physically grounded stochastic simulation in which coal temperature is the latent state and gas concentrations are its noisy observations, so that ground-truth risk labels are constructed independently of the gas-derived indices supplied to the models; evidence, in both simulation and field data, that Graham's index detects no fire at any threshold achieving a false-alarm rate at or below 0.20, its field ordering between fire and non-fire episodes being inverted; and an ablation and permutation analysis showing that the classical derived indices contribute nothing measurable to a sequence model that already receives their constituent gases, an attention-augmented LSTM being outperformed in the field by a threshold on a single carbon-monoxide channel.

The protocol is developed on simulated data, where ground truth is known by construction, and then applied unchanged to field monitoring records from three coal mines. Both components have limitations, stated here rather than deferred: the simulator cannot reproduce the full complexity of a real mine atmosphere, and the field sample contains only three confirmed fires. The claim advanced is not that any particular model is best, but that the evaluation practice currently standard in this literature is capable of inverting the apparent ranking of methods, and that the correction is inexpensive.

Representative studies are summarised against a single question: what would this method do during a monitoring period in which no fire develops? In every case the answer is unreported (Supplementary Table S1).

2. Results

Figure 1 summarises the study design. A physics-based simulator generates monitoring episodes in which coal temperature is the latent state; a sensor model converts the true gas concentrations into noisy, drifting observations; features are computed from the observations alone; candidate methods produce a continuous alarm score; and all methods are compared on a common operating characteristic. Risk labels are derived from the latent temperature and are never visible to any method.

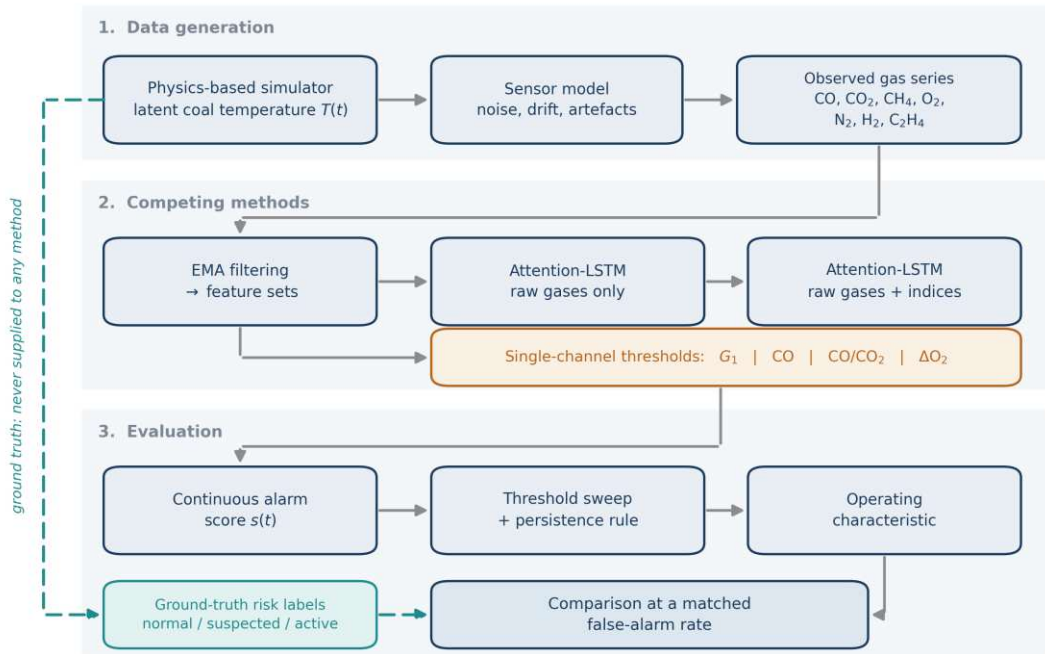


Figure 1

2.1. Characteristics of the simulated data

In a representative simulated test episode the latent temperature remains near ambient for roughly the first three weeks, crosses the suspected-heating threshold and then accelerates to the active-fire threshold. Carbon monoxide rises over three orders of magnitude across the episode, whereas methane fluctuates on a diurnal and random-walk pattern unrelated to the heating, and oxygen falls by approximately six percentage points as carbon dioxide rises (Supplementary Fig. S1).

Over the same episode Graham's index rises by more than two orders of magnitude but is visibly noisy in the early, near-baseline period, where the oxygen deficiency in its denominator is small. The Jones-Trickett ratio remains within the band conventionally associated with coal as the fuel for most of the episode, which is the

correct behaviour for a fuel-identification index and confirms that the simulated chemistry is internally consistent; it is not a severity signal and does not behave as one (Supplementary Fig. S2).

Across the test set the measured gases are correlated through their common temperature driver but far from collinear, and methane is essentially uncorrelated with the oxidation channels, as intended. The derived indices are strongly related to their constituent gases, which anticipates the ablation result of Section 2.3 (Supplementary Fig. S3).

2.2. Classification performance

At window level both variants reach 0.985 overall accuracy on the held-out test episodes. This figure is, however, almost entirely determined by the normal class, which comprises 82 % of test windows: both models classify normal and active-fire windows nearly perfectly, and both struggle on the rare transitional class, where precision falls to 0.53 and 0.54, respectively. Macro-F1, which weights the three classes equally, is 0.884 for the physics-informed variant and 0.876 for the raw-gases ablation, a difference of 0.008, well within the variation observed between random seeds (Supplementary Table S4).

These figures also illustrate the problem this article addresses. Reported in isolation, an accuracy of 0.985 would appear strong and comparable with figures throughout the literature summarised in Supplementary Table S1, yet Section 2.5 shows the same models detect only half of the fires once their false-alarm behaviour is constrained.

2.3. Contribution of the derived indices

Permutation importance for the physics-informed model, measured as the decrease in test macro-F1 when a single channel is randomly permuted across samples, ranks carbon monoxide as the most important channel, followed by the oxygen deficiency. Graham's index and the Jones-Trickett ratio, by contrast, produce changes of 0.000 and -0.002 respectively; that is, the model's performance is unaffected by destroying them entirely (Supplementary Table S5 and Supplementary Fig. S4).

This is not evidence that Graham's index is uninformative in absolute terms, but that it is redundant here: G_1 is a deterministic function of CO and ΔO_2 , both already supplied, and a recurrent network can represent that function internally. Supplying hand-derived ratios to a model that already observes their inputs should therefore not be expected to help, and the ablation in Supplementary Table S4 confirms it does not.

2.4. Learned temporal attention

The mean attention weight, taken as a function of position within the input window and averaged separately over windows of each true class, is close to uniform for normal windows, indicating that the model integrates the whole window when nothing is happening. For windows labelled as suspected heating or active fire the profile tilts towards the most recent samples, consistent with a model that weights recent evidence more heavily once a trend is established. The effect is modest, and the attention mechanism should be understood as providing a diagnostic view of the model rather than a physical explanation of the process (Supplementary Fig. S5).

2.5. False-alarm-controlled benchmark

Figure 2 presents the central result. Panel (a) shows detection rate against false-alarm rate for all six methods across the full threshold sweep; panel (b) shows median lead time against the same abscissa. The shaded region marks false-alarm rates at or below 0.10.

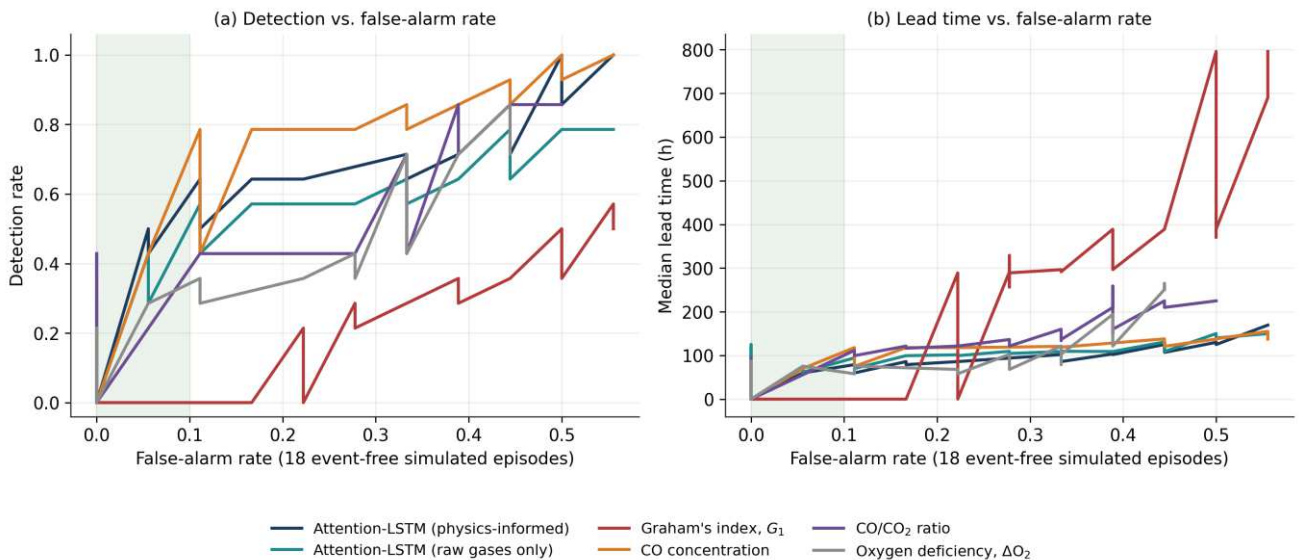


Figure 2

Graham's index behaves qualitatively differently from every other method. Its curve does not enter the acceptable operating region at all: across the entire threshold sweep, no value of τ yields a false-alarm rate at or below 0.20 while detecting any fire. When the false-alarm constraint is removed, the same index appears extremely effective; at $G_1 > 1.0$ it detects 86 % of fires with a median lead time of 664 h, and at $G_1 > 2.0$ it detects 86 % with 624 h, but those operating points carry false-alarm rates of 0.78 and 0.61 respectively. An alarm that fires during more than three-quarters of quiet monitoring periods is not an early-warning system.

Table 2 compares all methods at a matched false-alarm rate of 0.06. The physics-informed Attention-LSTM detects half of the fires with a median lead time of 60 h; the raw-gases ablation, the CO threshold and the CO/CO₂ threshold each detect 43 %, with median lead times of 70, 70 and 94 h. The confidence intervals of these four methods overlap almost completely, and no pairwise difference approaches significance with 14 fire episodes. The only method that is clearly separated from the others is Graham's index, and it is separated downwards.

Two conclusions follow from Fig. 2 and Table 2. First, the apparent superiority of the classical index in lead time is an artefact of an uncontrolled operating point, and reverses completely once that point is matched. Second, among the remaining methods, a sophisticated sequence model with attention performs no better than a threshold on a single carbon monoxide sensor, to within the resolution of this experiment.

2.6. Field dataset

The protocol developed above was applied unchanged to a field dataset comprising 18 continuously monitored episodes from three underground coal mines. Each episode covers approximately 30 days of tube-bundle and sensor monitoring at a four-hourly sampling interval, giving 3,240 measurement points in total. Three episodes ended in an independently confirmed fire and fifteen did not. The event-free episodes are the component that makes the false-alarm rate measurable, and they were retained specifically for that purpose. Episodes span goaf, return-airway and sealed-area monitoring locations, and three coal ranks, so the dataset also exposes each method to conditions well outside any single operating regime. Table 1 summarises the episodes.

Table 1. Field episodes used for validation.

Location type	Episodes	Confirmed fires	Event-free	Coal rank
Goaf	8	1	7	High-vol / Med-vol bituminous, Sub-bituminous A
Return airway	6	0	6	High-vol / Med-vol bituminous, Sub-bituminous A
Sealed area	4	2	2	Med-vol bituminous, Sub-bituminous A
Total	18	3	15	n/a

Ground truth was taken from the operating record rather than from any gas-derived quantity: the confirmation time of each fire is the moment at which it was established by thermal camera, in-situ multipoint temperature sensors or PT100 arrays behind a seal, together with the associated operational response. Lead time is measured backwards from that documented moment. Because the confirmation evidence is independent of CO, CO₂, O₂ and N₂, the circularity that would arise from defining a fire by the same index being evaluated is avoided in the field data exactly as it was in the simulation.

Field records were resampled onto the hourly grid used during training, and the identical feature pipeline, persistence rule and threshold sweep were applied. The neural models were used exactly as trained on simulated data, with no retraining or fine-tuning, so their performance here is a zero-shot transfer result.

2.7. Graham's index inverts in the field

The most consequential field result concerns Graham's index. Across the full threshold sweep no value achieves a false-alarm rate at or below 0.20 while detecting any of the three fires; its detection rate is 0.00 [95 % CI 0.00, 0.00] at every admissible operating point, reproducing on field data the failure identified in simulation.

Figure 3 shows why, and the reason is more severe than the simulation suggested. In seven of the thirteen event-free ventilated episodes the index exceeds 24, reaching 44 in one goaf record, one to two orders of magnitude above the conventional alarm level of unity; in eleven of those thirteen it exceeds unity at some point. In two of the three confirmed fires, both of them behind seals, it never exceeds 0.50. The ordering between fire and non-fire episodes is therefore not merely weak, it is inverted: the highest values of Graham's index in this dataset occur in episodes where nothing happened, and the lowest occur during confirmed fires. No threshold on this index can separate the two groups.

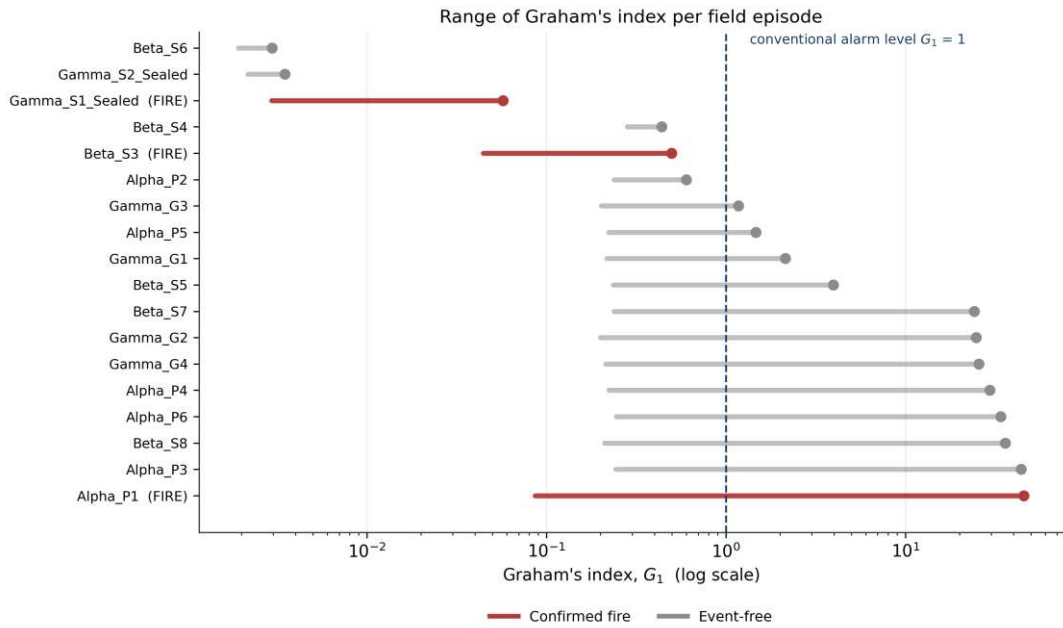


Figure 3

Two opposite mechanisms produce this. In ventilated goaf and return-airway locations the oxygen deficiency is close to zero, so ordinary sensor noise in O_2 and N_2 generates large excursions of the ratio, and the denominator occasionally becomes negative. In sealed areas the deficiency instead rises to between 1 and 21 across the four episodes, with a median of 18 in three of them against medians of 0.05 to 0.47 in the ventilated episodes, and the index is correspondingly suppressed: in the sealed-area fire, coal temperature reached 118.5 °C and carbon monoxide rose from 5 to 128 ppm, yet Graham's index never exceeded 0.06. Figure 4 contrasts the two cases.

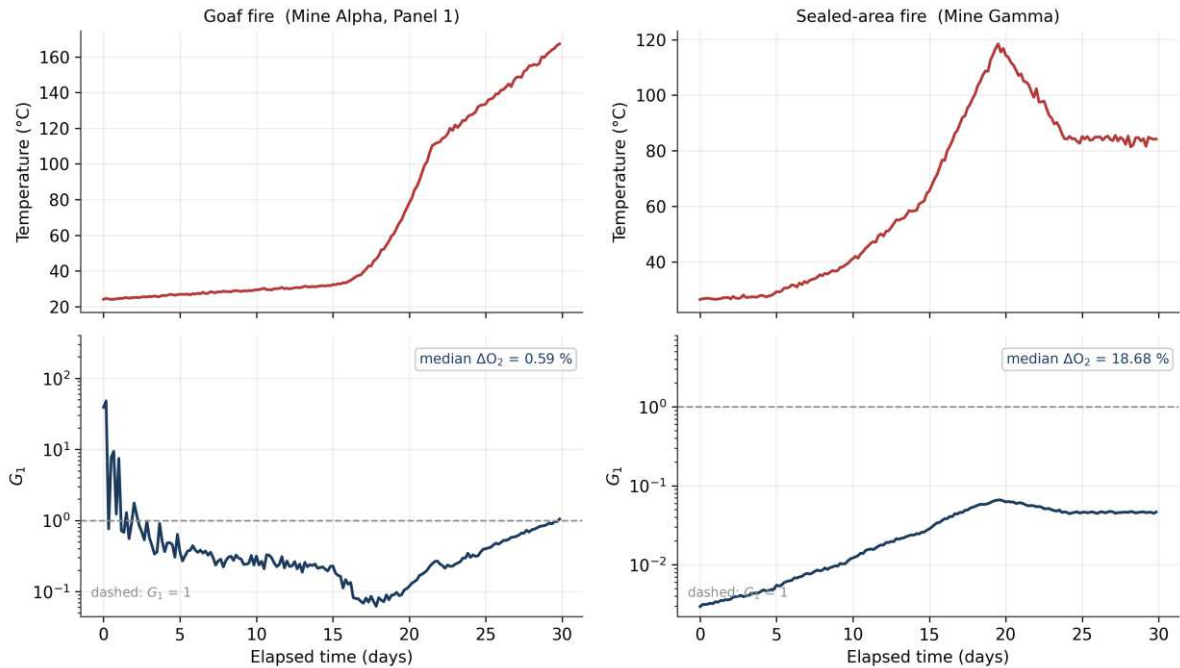


Figure 4

A single numerical threshold on Graham's index therefore cannot be transferred between monitoring locations within the same mine, let alone between mines: one calibrated in a goaf is crossed continuously in ordinary operation, and the same value applied behind a seal does not respond to an established fire.

2.8. Field method comparison at a matched false-alarm rate

Figure 5 gives the operating characteristics of all six methods on the field data, and Table 2 reports the comparison at a matched false-alarm rate of 0.10, the achieved rate being 0.07 for every method at its selected threshold.

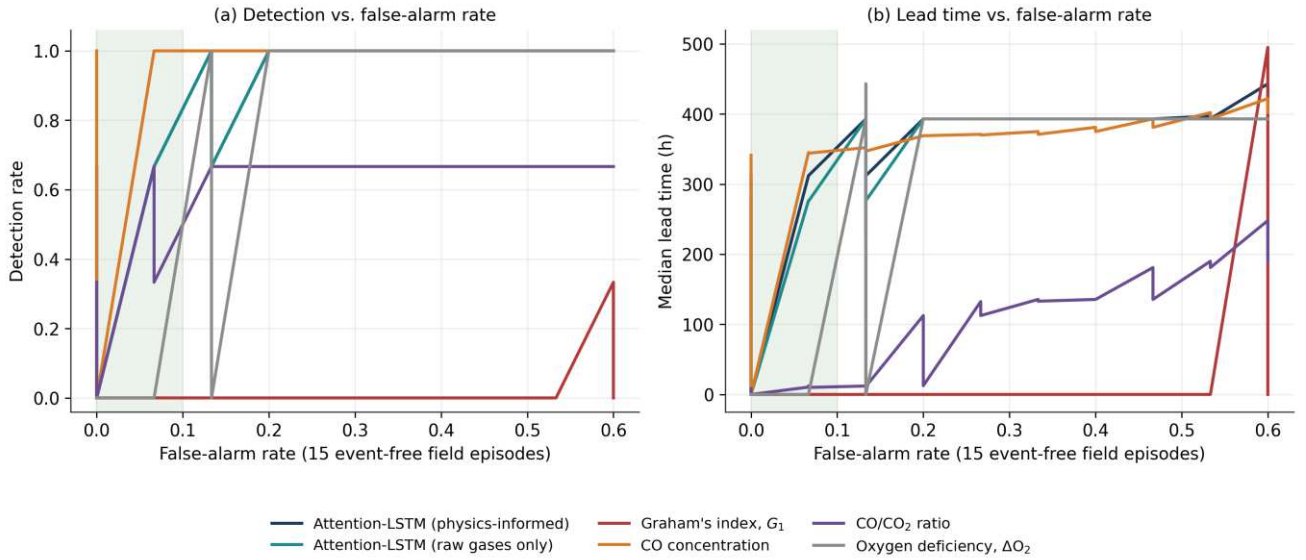


Figure 5

Table 2. Method comparison at a matched false-alarm rate. Simulation: ceiling 0.10, achieved 0.06 ($n = 14$ fire and 18 non-fire episodes). Field: ceiling 0.10, achieved 0.07 ($n = 3$ fire and 15 non-fire episodes). Confidence intervals are 95 % bootstrap percentile intervals over episodes.

Method	Simulation ($n = 14$ fires)		Field ($n = 3$ fires)		
	Detection [95 % CI]	Lead time (h)	Detection [95 % CI]	Lead time (h)	Fires
CO concentration	0.43 [0.14, 0.71]	70	1.00 [1.00, 1.00]	345	3 / 3
Attention-LSTM (physics-informed)	0.50 [0.21, 0.71]	60	0.67 [0.00, 1.00]	312	2 / 3
Attention-LSTM (raw gases only)	0.43 [0.14, 0.71]	70	0.67 [0.00, 1.00]	276	2 / 3
CO/CO ₂ ratio	0.43 [0.21, 0.71]	94	0.67 [0.00, 1.00]	10	2 / 3
Oxygen deficiency, ΔO_2	0.29 [0.07, 0.50]	76	0.00 [0.00, 0.00]	n/a	0 / 3
Graham's index, G_1	0.00 [0.00, 0.00]	n/a	0.00 [0.00, 0.00]	n/a	0 / 3

A threshold on a single carbon-monoxide channel detects all three fires with a median lead time of 345 h, while the two attention-LSTM models detect two, with lead times of 312 and 276 h. The simulation result is thus reproduced and if anything strengthened: the additional capacity of an attention-based sequence model does not convert into better alarm performance than a single well-chosen channel.

2.9. Where the learned models fail

Both neural models missed the same fire, and which one is instructive: the goaf fires were detected, the sealed-area fire was not. The models were trained only on simulated episodes with oxygen near 20.9 % and methane below about 1 %, whereas the sealed-area record has oxygen near 4 % and methane above 11 %, so the training standardisation maps those inputs to a region of feature space the network never encountered. The carbon-monoxide threshold, scale-free in the other species, was unaffected.

This is a limitation of the trained model rather than of the protocol, and the lesson transfers: a detector trained on one atmospheric regime should not be assumed to generalise to another, the sealed-area regime differing from the ventilated one categorically rather than marginally.

3. Discussion

The mechanism is numerical rather than chemical. In a ventilated roadway close to fresh air the oxygen deficiency is by construction near zero: across the simulated test set its fifth percentile is 0.001 %, against a mean of 0.98 %. The index is therefore evaluated as a ratio with a vanishing denominator precisely in the baseline regime, so ordinary sensor noise in the oxygen and nitrogen channels produces excursions that are large in absolute terms, the maximum over the test set being 144 against a nominal alarm threshold of 1.0. Any threshold low enough to catch a developing heating is therefore also low enough to be crossed repeatedly by noise during quiet periods. A further contributor is that CO is not exclusively a combustion product: coal seams release a measurable primal CO of coalification origin even without oxidation, so part of the baseline fluctuation in the numerator reflects a source unrelated to heating³⁵.

This is consistent with, and gives a quantitative account of, the long-standing practitioner caution that no single ratio should be interpreted in isolation⁷ and that index gases must be read against a site-specific baseline⁸. It also suggests a practical remedy that does not require machine learning at all: conditioning the index on a minimum oxygen deficiency, or replacing it with a numerically better-conditioned statistic such as the CO concentration itself, both of which perform substantially better in Table 2.

At a matched false-alarm rate the attention-augmented LSTM is statistically indistinguishable from a threshold on a single CO channel, and in the field it detects one fewer fire. This does not show that sequence models are useless for the problem; it shows that here the additional capacity is not converted into detection performance. Two explanations are consistent with the data. The signal may be close to single-channel in character, since a heating event raises CO monotonically over orders of magnitude, leaving limited additional structure for a temporal model to exploit; and the training set of 147 episodes may be too small for a recurrent model to generalise across the parameter variation deliberately built into the simulator.

The finding nonetheless has a direct bearing on how results in this field are reported. The studies in Supplementary Table S1 that report R2 above 0.95 for temperature or concentration forecasting are not in conflict with the present result, because they measure a different quantity. A model can reconstruct a concentration trajectory almost perfectly and still provide no usable alarm, if the reconstruction is accurate mainly during periods when nothing is happening. Reporting an operating characteristic alongside the regression metric would resolve the ambiguity at essentially no cost.

The behaviour documented here is a nuisance-alarm problem of the kind alarm-management practice exists to prevent. An indicator that activates during 78 % of monitoring periods in which no fire develops does not satisfy the ISA-18.2 requirement that an alarm correspond to a condition requiring a response, and would not survive the rationalisation stage of the alarm lifecycle^{32,31}. The hazard is not merely wasted effort: an indicator known to be usually spurious is discounted, degrading response to the genuine event it was installed to catch. In an underground coal mine, where that response may involve ventilation changes, inertisation or withdrawal of personnel, the degradation is consequential.

For mine practice, the results therefore support a conservative reading: the classical indices remain valuable as diagnostic instruments interpreted by an experienced engineer against a known site baseline, which is how they have always been used and how the source literature recommends they be used^{7,8}, but their transfer into automatic fixed-threshold alarm rules is difficult to justify without site-specific characterisation of the nuisance rate. Where such a rule is nonetheless required, the conditioning remedy discussed above should be considered, and the resulting alarm should be rationalised and monitored in the same manner as any other safety-related alarm on the installation.

For research practice the recommendation is concrete. Studies proposing early-warning methods should retain and report monitoring periods in which no fire developed, report a false-alarm rate alongside any lead time or detection claim, compare methods at a matched operating point rather than at each method's own convenient threshold, and report uncertainty, since episode counts in this field are small enough that confidence intervals span most of the achievable range. None of this requires new instrumentation or new data.

- Simulated and field data: The methodological development and the ablation analysis rest on synthetic episodes, and the field component comprises 18 episodes from three mines. The simulator cannot reproduce the full complexity of a real mine atmosphere, and the field sample, while it reproduces the central findings, is small. The two components are complementary rather than interchangeable: the simulation supports the mechanistic argument, the field data establish that the failure of Graham's index is not an artefact of the simulator.
- Small event counts: The simulated test partition contains 14 fire and 18 non-fire episodes, and the field dataset 3 fire and 15 non-fire episodes. Bootstrap intervals on field detection rate accordingly span [0.00, 1.00] for methods detecting two of three fires, and no pairwise difference between the learned models and the CO threshold is statistically established. The zero detection rate of Graham's index is the exception: it holds at every admissible threshold in both settings and is not a marginal result.
- Location-type dependence: The field data show that the behaviour of Graham's index depends strongly on monitoring location, with the oxygen deficiency differing by one to two orders of magnitude between ventilated roadways and sealed areas. Only four sealed-area episodes were available, so the sealed-area result rests on a small sample and warrants confirmation on a larger set.
- Limited model exploration and transfer: A single architecture at a single scale was evaluated, without extensive hyperparameter search, and the models were applied to field data zero-shot, without retraining. Their failure on the sealed-area fire is therefore a statement about transfer from the simulated regime, not about the architecture in general. No comparison was made with the gradient-boosting or hybrid physics-informed models cited in Supplementary Table S1.
- Labelling: In the simulation, risk states are defined by latent temperature thresholds. In the field data, fire confirmation is taken from the operating record (thermal camera, in-situ temperature arrays, PT100 sensors behind seals), each of which carries its own detection delay; lead times are therefore measured from the

moment of documented confirmation rather than from the physical onset of heating, and are correspondingly conservative.

The protocol requires only that event-free monitoring periods be retained, so it could be applied retrospectively to existing archives from any mine logging continuous gas monitoring, and a larger multi-site study would allow the coal-rank dependence documented by Onifade and Genc¹² to be examined. The numerical conditioning of the classical indices also deserves direct attention: a regularised or baseline-referenced formulation of Graham's index may recover much of its diagnostic value while eliminating the failure mode identified here. Finally, since the transitional state is both the rarest and the operationally most valuable, methods targeted at that regime, such as change-point detection or anomaly scoring against a learned site baseline, may suit it better than three-class classification.

4. Conclusions

This article set out to test whether the comparison between classical gas indices and deep sequence models for coal-fire early warning is conducted on a sound footing. It concludes that it is not, and that the deficiency is inexpensive to remedy.

Under a false-alarm-controlled protocol applied to 210 simulated episodes and then, unchanged, to 18 field episodes from three coal mines, Graham's index achieved a detection rate of zero at every threshold whose false-alarm rate was at or below 0.20. In the field the failure is more severe than the simulation predicted: the index exceeded 24 in seven of the thirteen event-free ventilated episodes, reaching 44, while remaining below 0.50 in the two confirmed fires behind seals, an inversion of the ordering it is supposed to provide. The mechanism is the oxygen deficiency in its denominator, near zero in ventilated roadways and one to two orders of magnitude larger behind seals, where the index stayed below 0.06 during a fire reaching 118.5 °C. At a matched false-alarm rate a single carbon-monoxide threshold detected all three field fires with 345 h median lead time, while attention-augmented LSTM models detected two.

The field sample is small and the confidence intervals on the learned models span most of the achievable range, so no pairwise difference between the neural models and the carbon-monoxide threshold should be regarded as established. The zero detection rate of Graham's index is the exception, holding at every admissible threshold in both settings. The structural conclusion is more robust than any individual number: a lead time reported without a matched false-alarm rate does not constrain a method's operational value, and comparisons at unmatched operating points can invert the true ranking. Retaining event-free monitoring periods and reporting a single operating characteristic would allow recent progress in this field to be assessed against the criterion any alarm on an underground installation must already meet: that it indicates a condition requiring a response.

5. Methods

5.1. Simulation framework

Field records combining continuous multi-species gas monitoring with independently confirmed heating status are not publicly available, and the label circularity that arises when risk states are defined from the same indices being evaluated would invalidate the comparison in any case. A simulation environment was therefore constructed in which the ground truth is known by construction and is causally upstream of the observations.

The latent state is the temperature T of the heating zone, evolved by a lumped self-heating balance between Arrhenius heat generation and convective removal by ventilation:

$$\frac{dT}{dt} = q \varphi(t) \exp \left[-\frac{E_a}{R} \left(\frac{1}{T} - \frac{1}{T_{ref}} \right) \right] v(t) - h v(t) (T - T_{amb}) \quad (1)$$

where q is a per-episode heat-generation scale, $\varphi(t)$ encodes an incubation delay and an optional stalling phase, $v(t)$ is the relative ventilation airflow, and h is the cooling coefficient. Equation (1) has a bifurcation in q : below a critical value the system settles a few degrees above ambient, above it the temperature runs away. Episodes were generated with q sampled on both sides of that bifurcation, so that the dataset contains genuine event-free episodes as well as developing fires. Of 210 episodes, 85 reached the active-fire threshold, 9 peaked within the suspected-heating band, and 116 remained normal.

Gas concentrations are generated as observations of T , each species with its own activation energy, so that the ratios between species evolve with temperature rather than being fixed multiples of one another; this is the physical basis on which index gases work at all. Carbon monoxide, carbon dioxide, hydrogen and ethylene are produced with increasing activation energy in that order, reproducing the established observation that ethylene appears only at advanced stages. Methane is generated by seam desorption and is largely independent of temperature. Oxygen deficiency is generated with its own kinetics, and nitrogen is obtained by difference, which closes the mass balance: across the test set, the measured composition sums to between 99.3 % and 100.7 %. Products are diluted by the fluctuating ventilation airflow. Supplementary Table S2 lists the sampled parameter ranges.

A sensor model then adds independent multiplicative and additive noise, a slow random-walk drift, and occasional spike artefacts to each channel separately. Because the noise is independent per channel and the ventilation

term acts multiplicatively, the resulting observations are correlated through their common physical driver without being degenerate (Supplementary Fig. S3).

Risk labels follow the characteristic temperatures used in the self-heating literature: normal below 50 °C, suspected heating between 50 and 120 °C, and active fire at or above 120 °C. Because the transit through the intermediate band is intrinsically rapid once runaway begins, the suspected-heating class is genuinely rare, comprising 1.7 % of test windows; this imbalance is a property of the physics rather than a sampling artefact, and is handled by class weighting.

In addition to the simulated episodes, the study draws on field monitoring records described in Section 2.6. Those records comprise 18 monitoring episodes recorded at goaf, return-airway and sealed-area measurement points in bituminous and sub-bituminous coal seams across three underground mines, and they are used for validation only; no field record contributed to model training or to the selection of any hyperparameter.

5.2. Pre-processing and features

Each channel is smoothed with an exponential moving average, $y_t = \alpha x_t + (1 - \alpha)y_{t-1}$, with $\alpha = 0.08$ chosen to suppress spike artefacts while preserving the multi-day trend. From the smoothed channels the classical indices of Table 3 are computed. Channels spanning several orders of magnitude (CO, H₂, C₂H₄, ΔO_2 , G₁, CO/CO₂) are log-compressed before standardisation; the same transformation is applied to both model variants to ensure a fair comparison.

Table 3 states the Jones-Trickett ratio in its established form. It is worth emphasising, because it is sometimes misapplied, that this ratio identifies the fuel involved rather than the severity of the event: values of 0.4-0.5 indicate methane, 0.5-1.0 coal, oil or conveyor belting, 0.9-1.6 timber, and values above 1.6 are conventionally taken to indicate a sampling or analytical error rather than an intense fire. It is included here as an input feature and reported for fuel confirmation, but it is not treated as a severity threshold.

Table 3. Classical gas indices used as model features and as single-channel baselines.

Index	Definition	Established interpretation	Role in this study
Oxygen deficiency, ΔO_2	$\Delta O_2 = 0.265 N_2 - O_2$ (%)	Oxygen consumed relative to the inert fraction of intake air	Input feature and single-channel baseline
Graham's index, G_1	$G_1 = 100 \frac{CO}{\Delta O_2}$ (CO in %)	Severity of coal oxidation; rises with heating	Input feature and single-channel baseline
Jones-Trickett ratio, TR	$TR = \frac{CO_2 + 0.75 CO - 0.25 H_2}{\Delta O_2}$ (all in %)	Fuel identification, NOT severity: 0.4-0.5 methane, 0.5-1.0 coal/oil/belt, 0.9-1.6 timber, >1.6 sample invalid	Input feature; reported for fuel confirmation only
CO/CO ₂ ratio	$\frac{CO}{CO_2}$ (both in %)	Rises as oxidation intensifies	Input feature and single-channel baseline

5.3. Models

The classifier is a single-layer LSTM³³ with 64 hidden units, followed by an additive attention layer over the hidden-state sequence³⁴ and a softmax output over the three risk states. Writing h_t for the hidden state at step t :

$$e_t = v^T \tanh(W h_t + b), \quad \alpha_t = \text{softmax}(e_t), \quad c = \sum_t \alpha_t h_t \quad (2)$$

Inputs are 24-hour windows sampled with a stride of 10 hours. Two variants are trained with identical architecture, hyperparameters and random seed, differing only in their input channels: a physics-informed variant receiving the six measured gases together with the four derived indices, and a raw-gases-only ablation receiving the six measured gases alone. Training minimises class-weighted cross-entropy with the Adam optimiser, learning rate 1.5×10^{-3} with plateau decay, gradient clipping at 3.0, dropout 0.2, and early stopping on validation macro-F1.

The dataset is partitioned by episode rather than by sample, in the ratio 70:15:15, so that no window from a test episode appears in training in any form. Standardisation statistics are computed on the training partition only. Supplementary Table S3 reports the resulting partition sizes.

5.4. Baseline methods

Four single-channel baselines are evaluated alongside the two neural models: Graham's index, the CO concentration, the CO/CO₂ ratio and the oxygen deficiency ΔO_2 . Each is used directly as an alarm score, so that a threshold sweep over the score reproduces the family of fixed-threshold rules used in field practice, including the conventional $G_1 > 1.0$ criterion as one point on that family.

5.5. Evaluation protocol

The central methodological proposal of this article is the following protocol. Each method is reduced to a continuous alarm score $s(t)$. For a threshold τ , an alarm is declared at the first time the score exceeds τ for six

consecutive samples; the persistence requirement suppresses isolated spikes and is applied identically to every method. Two quantities are then measured:

- False-alarm rate: the fraction of episodes that never reach the active-fire threshold in which an alarm is nevertheless declared at any point.
- Detection rate and lead time: among episodes that do reach the active-fire threshold, the fraction in which an alarm is declared before the reference time t_{ref} at which the latent temperature first reaches 120 °C, and the median value of $t_{\text{ref}} - t_{\text{alarm}}$ over those episodes.

Sweeping τ across the full range of the score traces an operating characteristic for each method. Methods are then compared at a matched false-alarm rate, that is, at the threshold that maximises detection subject to the false-alarm rate not exceeding a stated ceiling. Uncertainty is quantified by bootstrap resampling of episodes, 2000 replicates, reported as 95 % percentile intervals. Because both quantities are defined per episode, this procedure is equally applicable to field records and requires only that monitoring periods without a confirmed fire be retained rather than discarded.

Data availability

The physics-based simulator, the trained models and the complete evaluation code required to reproduce every simulated result reported here will be deposited in a public repository upon acceptance. The field monitoring records analysed in this study are provided in full as Supplementary Data S1; mine, panel and seam identifiers have been anonymised at the request of the participating operators.

Author contributions

U.C. is the sole author. U.C. conceived the study, designed the protocol, wrote the simulator and analysis code, curated and analysed the field data, produced the figures and wrote the manuscript.

Additional information

Competing interests: The author declares no competing interests.

Supplementary information is available for this paper.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

1. Stracher, G. B. & Taylor, T. P. Coal fires burning out of control around the world: thermodynamic recipe for environmental catastrophe. *Int. J. Coal Geol.* 59, 7-17 (2004).
2. Kuenzer, C. & Stracher, G. B. Geomorphology of coal seam fires. *Geomorphology* 138, 209-222 (2012).
3. Deng, J., Shi, B., Torero Cullen, J. L., Xue, Y. & Zhou, F. Quantification of fugitive carbon emissions from uncontrolled combustion in coal fire area: the Sandaoba coalfield, Xinjiang, China. *Environ. Res.* 269, 120923 (2025).
4. Kim, A. G. Greenhouse gases generated in underground coal-mine fires. *Rev. Eng. Geol.* 18, 1-13 (2007).
5. Carras, J. N., Day, S. J., Saghafi, A. & Williams, D. J. Greenhouse gas emissions from low-temperature oxidation and spontaneous combustion at open-cut coal mines in Australia. *Int. J. Coal Geol.* 78, 161-168 (2009).
6. Mohalik, N. K., Lester, E., Lowndes, I. S. & Singh, V. K. Estimation of greenhouse gas emissions from spontaneous combustion/fire of coal in opencast mines: Indian context. *Carbon Manag.* 7, 317-332 (2016).
7. Singh, A. K., Singh, R. V., Singh, M. P., Chandra, H. & Shukla, N. P. Mine fire gas indices and their application to Indian underground coal mine fires. *Int. J. Coal Geol.* 69, 192-204 (2007).
8. Liang, Y., Zhang, J., Wang, L., Luo, H. & Ren, T. Forecasting spontaneous combustion of coal in underground coal mines by index gases: a review. *J. Loss Prev. Process Ind.* 57, 208-222 (2019).
9. Liang, Y., Yang, Y., Guo, S., Tian, F. & Wang, S. Combustion mechanism and control approaches of underground coal fires: a review. *Int. J. Coal Sci. Technol.* 10, 24 (2023).
10. Onifade, M. & Genc, B. A review of spontaneous combustion studies: South African context. *Int. J. Min. Reclam. Environ.* 33, 527-547 (2018).
11. Onifade, M. & Genc, B. Spontaneous combustion liability of coal and coal-shale: a review of prediction methods. *Int. J. Coal Sci. Technol.* 6, 151-168 (2019).
12. Onifade, M. & Genc, B. A review of research on spontaneous combustion of coal. *Int. J. Min. Sci. Technol.* 30, 303-311 (2020).
13. Onifade, M., Genc, B. & Wagner, N. Influence of organic and inorganic properties of coal-shale on spontaneous combustion liability. *Int. J. Min. Sci. Technol.* 29, 851-857 (2019).
14. Onifade, M., Genc, B., Gbadamosi, A. R., Morgan, A. & Ngoepe, T. Influence of antioxidants on spontaneous combustion and coal properties. *Process Saf. Environ. Prot.* 148, 1019-1032 (2021).
15. Gbadamosi, A. R., Onifade, M., Genc, B. & Rupprecht, S. Spontaneous combustion liability indices of coal. *Combust. Sci. Technol.* 193, 2659-2671 (2020).
16. Qin, Y., Song, Y., Liu, W., Wei, J. & Lv, Q. Assessment of low-temperature oxidation characteristics of coal based on standard oxygen consumption rate. *Process Saf. Environ. Prot.* 135, 342-349 (2020).

17. Zhang, L. et al. Prediction of coal self-ignition tendency using machine learning. *Fuel* 325, 124832 (2022).
18. Shukla, U. S., Mishra, D. P. & Mishra, A. Prediction of spontaneous combustion susceptibility of coal seams based on coal intrinsic properties using various machine learning tools. *Environ. Sci. Pollut. Res.* 30, 69564-69579 (2023).
19. Qi, Y., Xue, K., Wang, W., Cui, X. & Liang, R. Prediction model of borehole spontaneous combustion based on machine learning and its application. *Fire* 6, 357 (2023).
20. Zhuo, H. et al. Prediction model for spontaneous combustion temperature of coal based on PSO-XGBoost algorithm. *Sci. Rep.* 15, 2752 (2025).
21. Wang, K. et al. Research on prediction model of coal spontaneous combustion temperature based on SSA-CNN. *Energy* 290, 130158 (2024).
22. Yang, G. et al. Method and validation of coal mine gas concentration prediction by integrating PSO algorithm and LSTM network. *Processes* 12, 898 (2024).
23. Ding, H. et al. Research on coal spontaneous combustion temperature prediction model based on BO-LSTM. *Process Saf. Environ. Prot.* 201, 107570 (2025).
24. Shao, X. et al. Deep learning framework based on ITOC optimization for coal spontaneous combustion temperature prediction: a coupled CNN-BiGRU-CBAM model. *Sci. Rep.* 15, 26700 (2025).
25. Xing, W., Yang, Y., Zhang, Y. & Yang, Y. Gas concentration prediction based on SSA algorithm with CNN-BiLSTM-attention. *Sci. Rep.* 15, 34214 (2025).
26. Su, P., Zhang, J. & Guo, L. A PINN-LSTM model for real-time gas concentration prediction in mine conveyor belt fires. *Fire* 8, 450 (2025).
27. Danish, E. & Onder, M. Application of fuzzy logic for predicting of mine fire in underground coal mine. *Saf. Health Work* 11, 322-334 (2020).
28. Zhao, H. et al. Research on early warning model of coal spontaneous combustion based on interpretability. *Sci. Rep.* 15, 18847 (2025).
29. Deng, J. et al. A review on coal spontaneous combustion prediction based on machine learning. *J. China Coal Soc.* 50, 336-360 (2025).
30. IEC. IEC 62682:2014, Management of Alarm Systems for the Process Industries (International Electrotechnical Commission, 2014).
31. ISA. ANSI/ISA-18.2-2016, Management of Alarm Systems for the Process Industries (International Society of Automation, 2016).
32. EEMUA. Alarm Systems: A Guide to Design, Management and Procurement, EEMUA Publication 191, 3rd edn (Engineering Equipment and Materials Users' Association, 2013).
33. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* 9, 1735-1780 (1997).
34. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd Int. Conf. on Learning Representations* (2015).
35. Zhu, H., Chang, M. & Wang, H. Study on primal CO gas generation and emission of coal seam. *Int. J. Min. Sci. Technol.* 27, 973-979 (2017).

Figure legends

Figure 1. Study design. Ground-truth risk labels derive from the latent coal temperature, independently of the gas-derived indices supplied to the models.

Figure 2. Simulation operating characteristics. (a) Detection rate and (b) median warning lead time against false-alarm rate over the full threshold sweep of each method. The shaded band marks false-alarm rates at or below 0.10. Detection and lead time at the matched operating point are given in Table 2 and shown graphically in Supplementary Fig. S6.

Figure 3. Range of Graham's index in each field episode. Confirmed fires are shown in red. The highest values occur in event-free episodes and two of the three fires remain below unity throughout.

Figure 4. Two confirmed field fires. Upper panels: coal temperature. Lower panels: Graham's index on a logarithmic scale, with the conventional alarm level shown dashed. The sealed-area fire is not detectable by the index at any threshold usable in the goaf.

Figure 5. Field operating characteristics. (a) Detection rate and (b) median lead time against false-alarm rate measured on the fifteen event-free field episodes. Detection and lead time at the matched operating point are given in Table 2 and shown graphically in Supplementary Fig. S7.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPDATA.zip](#)