UNIVERSITY OF WEST ATTICA

eISSN:3136-7658

Applied Manufacturing Intelligence Journal

Vol. 5, No. 2, April, 2026



Causal Enhanced CoAtNet BiFPN Model for Coal Ash Prediction

Marcin Skowroński[a,*], Zuzanna Lorenc[a], Emilia Lipińska[a]

[a] *Faculty of Electrical and Automatic Control Engineering, Czestochowa University of Technology, Czestochowa, 42-200, Poland*

Received 19 January 2026; Revised 10 April 2026; Accepted 12 April 2026; Published 15 April 2026

Abstract

Reliable ash-content prediction is needed for separation monitoring and product quality assessment in coal preparation plants, where moisture, particle size, and feed-source changes can alter sensing responses. This paper presents a causal-informed CoAtNet-BiFPN framework combining context-dependent sample weighting, convolution-attention feature extraction, and bidirectional multi-scale fusion. The numerical case study describes 12,480 synchronized image, near-infrared, and process-record samples spanning ash contents from 6.1% to 34.8%, with chronological training, validation, and test partitions. Context variables include moisture, dominant size fraction, belt speed, illumination, and source batch. The illustrative comparison reports a mean absolute error of 0.52 percentage points, a root mean squared error of 0.74 percentage points, a mean absolute percentage error of 3.18%, and a coefficient of determination of 0.972. Relative to the same backbone and fusion neck without context adjustment, the absolute error reductions are 0.09 and 0.11 percentage points, respectively; the latter corresponds to a 12.9% relative reduction. Under the reported held-out source condition, the proposed model's root mean squared error increases from 0.74 to 0.85 percentage points. These numerical comparisons illustrate the intended evaluation of predictive accuracy and contextual robustness, rather than establishing an identifiable causal effect or independently verified plant performance. The framework is intended for supervisory decision support; traceable assay records, implementation details, and external validation remain necessary before deployment claims can be substantiated.

Keywords: Coal Ash Prediction, Causal Feature Adjustment, CoAtNet, Bidirectional Feature Pyramid Network, Online Coal Quality Monitoring

1. Introduction

Coal preparation plants are increasingly using online quality prediction to meet product specifications in the face of short-term fluctuations in raw-coal properties [1]. Another relatively sensitive indicator of the minerals that will affect the calorific value and limit further blending decisions is ash content [2]. The old laboratory test is accurate but slow; therefore, the operator of a belt-scale control loop cannot adjust the medium-density, reagent-dosage and splitter settings in time [3]. Therefore, optical and spectral monitoring are good ways to determine the ash content rapidly, and the image texture, reflectance, and other process parameters can be correlated with historical assay data [4]. Machine-vision descriptors can acquire particle color and liberation pattern information of run-of-mine coal. To enhance the detection sensitivity of minerals, a spectral response curve of a wavelength-band-free-of-water region is selected. Process-data models are an operating environment that pure image regression cannot observe. However, the problem of online ash

* Corresponding author: marcin.sk@pcz.edu.pl

https://doi.org/10.68406/amij.2026.vol5iss2np1:1-15

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

prediction has not been solved because the apparent relationship between the sensing signal and ash content varies depending on the source seam, crushing conditions, moisture, and belt loading.

Deep neural networks have improved the quality model of coal by learning hierarchical texture and spectral features automatically. Convolutional Networks are suitable for local particle boundaries, and Attention Mechanisms can introduce a large receptive field of heterogeneous coal surfaces. Recent coal-ash research has examined the contribution of froth-image attributes using hybrid learning and computer vision models [5]. CoAtNet is considered here as a convolution-attention backbone. Neural ash-prediction models also include CNN-LSTM architectures optimized for flotation-recovered clean coal [6]. Multi-scale fusion offers a separate architectural route for combining spatial resolutions. Even a high-capacity network can learn spurious correlations if the nuisance variables are highly correlated with the ash labels. X-ray transmission studies address a different measurement mechanism for detecting coal ash [7]. Optical regressors may instead rely on wet-surface brightness, shadow geometry, or source-specific size distributions. Convolutional models have also been investigated for predicting flotation-concentrate ash from images [8]. Reliability across production conditions remains an evaluation requirement distinct from within-domain accuracy.

Causal adjustment is introduced into a CoAtNet-BiFPN regression model to address the problem of coal ash prediction. Moisture content, dominant particle-size fraction, belt speed, illumination conditions, and source batch are treated as distinct contextual variables that may alter the association between sensing features and ash labels. These variables are used to set the adjustment weight and regularize the representation learning so that the backbone focuses on features more relevant to the changes in ash. CoAtNet is used to extract both local particle patterns and long-range visual-spectral dependencies, and BiFPN performs bidirectional fusion of high-resolution texture maps, mid-level compositional maps, and low-resolution semantic maps. The intended contribution is the integration of context weighting and representation regularization with an established backbone and fusion neck, rather than a new convolution or attention operator. Residual grouping provides diagnostic information, not a validated predictive uncertainty distribution.

To ensure the traceability of the study from industrial motivation to deployable model behavior, the rest of this paper is organized into problem setting, architecture design and evidence-based validation. Section 2 reviews online coal ash prediction, deep feature learning and causal robustness methods for industrial sensing. Section 3 presents the causal ash prediction problem and introduces the data representation and adjustment strategy. Section 4 introduces the causal-enhanced CoAtNet-BiFPN architecture and the regression loss for feature extraction, bidirectional fusion, etc. Section 5 shows the results of the experimental data, comparison methods, ablation studies, sensitivity analyses and robustness tests in the two tables and six data figures. Section 6 is the conclusion of this paper, which introduces the contributions, deficiencies and subsequent research directions.

2. Related Work

2.1. Online Coal Ash Prediction

Online coal ash prediction has developed from empirical regression to multi-sensor learning. The first engineering system indirectly calculated the ash content from the density, grey-level or spectral index, and was relatively inaccurate. Near-infrared and visible imaging techniques have been employed to increase the response speed by positioning the device above the conveyor during production. Generally, the texture contrast and shape of reflectance are mapped to ash in these systems by partial least squares regression or support vector regression. Gao et al. [9] combined NIRS-XRF spectroscopy with random forest classification and PLSR subtype modeling to improve coal calorific value analysis across different coal types. The disadvantage is that feature extraction and model fitting are carried out separately, and only the compressed descriptors are used in the regression stage. When the raw-coal appearance changes due to moisture, crushing state, or blending ratio, the distribution of descriptors may shift in a way that the model cannot correct [10]. To reduce the instability, process-variable augmentation has been applied; however, the learned relationship is often correlational rather than causal and thus limited in transferability to new operating conditions [11].

2.2. Deep Feature Learning and Causal Robustness

Deep Learning methods have been introduced to enhance the representation power of coal and mineral sensing. Time-series ash prediction in dense-medium separation has been investigated using an improved EMD-LSTM method [12]. Yang et al. [13] investigated multi-scale multi-task neural learning with transfer learning for industrial flotation-concentrate ash prediction. CoAtNet can also be used to achieve the same goal; it has kept the useful locality of convolution layers and learned global correlations more deeply with attention layers. Multi-scale fusion is also a way. Multi-scale neural networks have been investigated for determining flotation-concentrate ash from froth images [14]. Deep models generally improve the generalization ability of the predictions on random splits but do not consider sampling conditions. Froth-tailings image fusion has also been investigated using a stacking model [15]. These predictive approaches do not, by themselves, establish an identifiable physical causal mechanism.

3. Causal Problem Formulation and Data Representation

3.1. Causal Variables and Adjustment Target

The target variable is the mass fraction of ash determined in the laboratory, and a short production cycle has been adopted. Each interval has a belt image block and a compact near-infrared descriptor and process record from the plant supervision system. Image block: A map of the surface texture and light-intensity distribution of coarse particles. A spectral descriptor stores reflectance data at specific wavelengths. The data are moisture content, main-size fraction, belt speed, feed-source code and illumination status. The above context factors are the circumstances of collection and production that can affect the correlation between sensing characteristics and ash labels. They are added to the adjustment term to reduce the impact of changes in operating conditions during the production period.

Figure 1 is the assumed causal structure of mineral composition, sensing observations, operating conditions and laboratory ash. Minerals in the sample will increase the ash content and may alter the color and light. The original batch and particle size will also affect the spread of ash due to geological differences and production cycles. The arrows encode physical hypotheses, not relationships identified by the numerical comparison; unmeasured mineralogy and sampling disturbances may confound the observed associations.

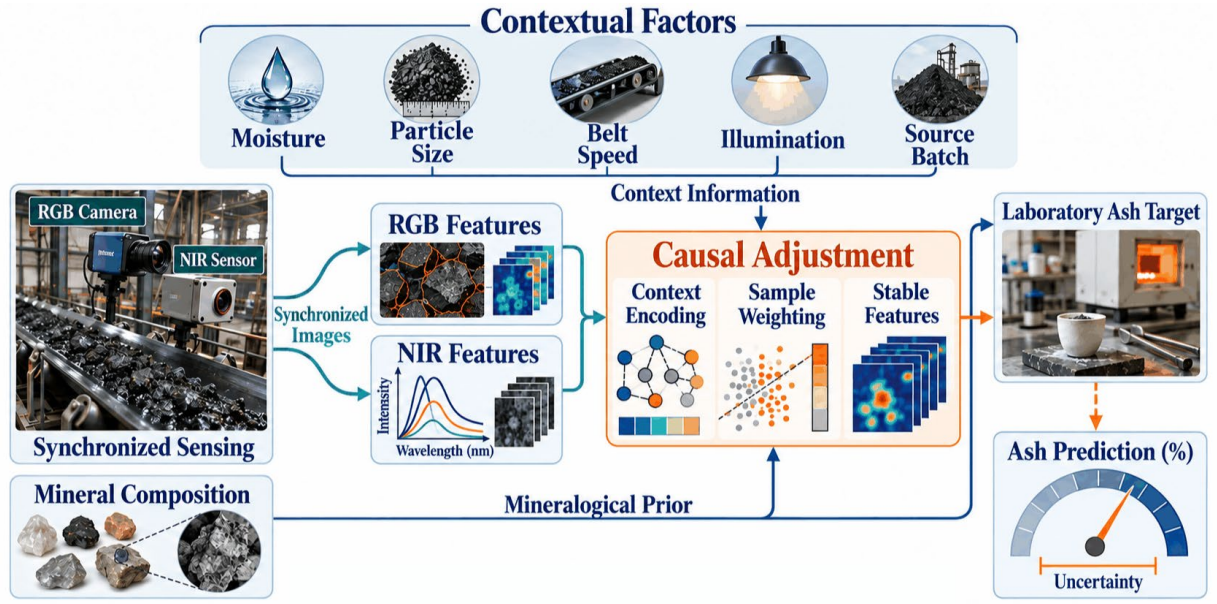


Figure 1. Causal graph of coal ash prediction with sensing variables, contextual confounders, adjustment pathway, and laboratory ash target

This structure motivates context adjustment before deep feature fusion. Let x_i denote the multimodal sensing input, c_i the contextual vector, and y_i the laboratory ash label for sample i. The following equation defines the weighted empirical risk used to train the predictor:

$$\mathcal{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w(c_i)\mathcal{L}(f_\theta(x_i, c_i), y_i) \tag{1}$$

Here, n is the number of training samples, f_θ is the prediction model, and $\mathcal{L}$ measures the discrepancy between predicted and measured ash. The adjustment weight $w(c_i)$ depends on the estimated prevalence of the corresponding operating context. Weights are clipped to prevent rare conditions from producing excessive gradients. A two-layer perceptron encodes the low-dimensional process records. Weighting reduces the dominance of frequently observed production states while retaining valid measurements from uncommon conditions. Its effectiveness depends on adequate overlap between contextual groups; weighting alone cannot recover information from operating states absent from the training data. The context-prevalence estimator, clipping limits, and weight normalization are not documented at implementation level. Consequently, the weighting procedure cannot yet be reproduced exactly, and the weights should not be interpreted as identified inverse-propensity weights.

The contextual variables were selected according to their physical roles in plant operation. Moisture alters surface reflectance, particle size changes the visibility of fractured mineral inclusions, and belt speed influences motion blur and exposure. Illumination affects image intensity and spectral baseline, while source batch represents geological and scheduling differences. These variables can improve prediction but may also encourage the model to memorize production periods. Representation stability is therefore examined across contextual groups with overlapping ash ranges.

With h denoting the learned representation and $\mathcal{C}_0$ and $\mathcal{C}_1$ denoting two context groups, the interventional contrast is expressed by the equation below:

$$\tau(z) = \mathbb{E}[y \mid do(h = z), c \in \mathcal{C}_1] - \mathbb{E}[y \mid do(h = z), c \in \mathcal{C}_0] \quad (2)$$

In this equation, z specifies a common representation state. The contrast describes whether the expected ash response remains dependent on context when the representation is held fixed. It provides a conceptual target for stability assessment rather than evidence that the interventional response is identifiable from observational plant records. Practical estimation requires assumptions about measured confounders, contextual overlap, and the relationship between the learned representation and the physical sensing process. Unrecorded disturbances may remain influential, so the adjustment mechanism is treated as an engineering approximation. Because the two expectations differ in context group while holding the representation fixed, this equation is a context contrast, not an estimated treatment effect. No physical intervention on the learned representation is reported.

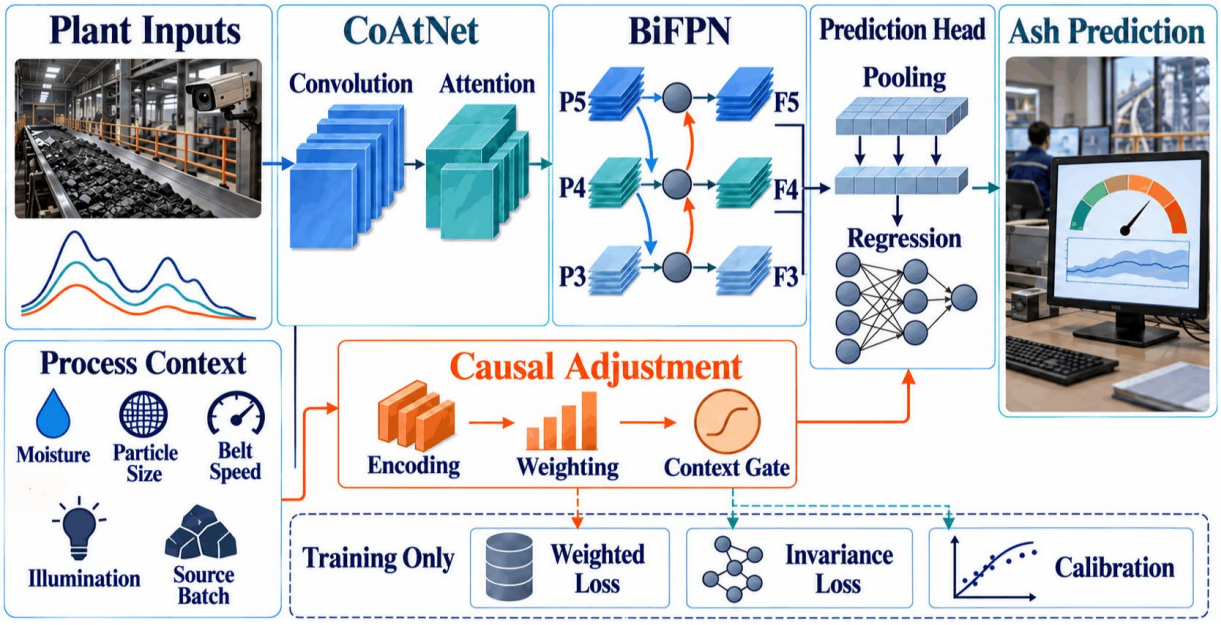


Figure 2. Causal-enhanced CoAtNet-BiFPN architecture from synchronized plant data to interval-level ash prediction

3.2. Multi Modal Data Representation

Data representation follows the same logic. The RGB belt image is cropped into overlapping windows and normalized by an illumination reference panel placed near the belt edge. The spectral descriptor is smoothed by a Savitzky-Golay filter and then standardized by training-set statistics. Process variables are encoded as continuous standardized values or as small learned embeddings. The three streams are not concatenated immediately. Image and spectral maps enter the CoAtNet stem, while contextual variables produce adjustment weights and a conditioning vector. This separation allows the model to use process information without letting process correlations replace ash-sensitive visual and spectral evidence. All windows sharing an interval label must remain in one partition, and fitted preprocessing parameters must be estimated from training data only. The available description does not establish whether boundary intervals were purged before window generation.

The final input tensor is built to preserve traceability between the physical measurement and the learned representation. Image windows are stored with their belt position, spectral descriptors retain their wavelength ordering, and process records retain the time stamp used for laboratory alignment. Outlier handling is conservative: intervals are removed only when the image is missing, the spectral acquisition fails, or the assay time cannot be matched to the sensing window. Extreme but valid ash values are retained because high-ash and low-ash intervals are operationally important. This representation also supports later diagnosis. When a prediction fails, the engineer can inspect whether the error came from a visual anomaly, an unusual spectral curve, or a rare process context, rather than treating the neural network as a single opaque regressor.

4. Causal Enhanced CoAtNet BiFPN Architecture

4.1. Network Flow and CoAtNet Feature Extraction

The proposed architecture is organized around four modules: a causal adjustment encoder, a CoAtNet feature extractor, a BiFPN fusion neck, and a regression head. Figure 2 traces the inference flow from synchronized plant

observations to interval-level ash prediction. Image windows first pass through a convolutional stem that preserves local texture and reduces noise. Spectral descriptors are projected into channel tokens and injected into intermediate stages through feature-wise affine modulation. The context encoder calculates adjustment weights only during training and produces a small conditioning vector during both training and inference.

Ash-related evidence occurs at different spatial scales. Shallow convolutional layers capture particle edges, surface roughness, and small mineral inclusions. Deeper attention layers connect spatially separated features, including heterogeneous blending zones and coarse gangue clusters. For a feature map F, the attention equation relates projected queries, keys, and values through similarity-weighted aggregation:

$$Attn(F) = Softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{3}$$

The matrices Q, K, and V are obtained from learned projections of the feature representation, and d is the query-key dimension. Scaling the similarity scores moderates their magnitude before normalization. The resulting attention weights allow features from one image region to incorporate information from other regions, which is useful when mineral-bearing particles occupy only a small part of the belt window.

Convolution and attention stages are connected through residual paths. Spectral modulation is applied after feature normalization so that reflectance information changes channel emphasis without replacing spatial evidence. During training, contextual weights scale sample losses, while an invariance penalty discourages excessive representation differences between source batches with comparable laboratory ash. Matching ash ranges is necessary because forcing identical representations across genuinely different compositions could suppress useful predictive information.

Network width is chosen to balance representation capacity and inference cost. Small convolution kernels capture local particle patterns, while deeper attention stages integrate their spatial arrangement. A coarse gangue fragment may be detected locally, but its contribution to interval ash depends on its prevalence relative to the surrounding stream. The context vector remains compact to limit the capacity of source embeddings to dominate the visual and spectral features. The description specifies the functional connections but does not report channel dimensions, the spectral projection size, or the exact CoAtNet variant. These omissions limit architectural reproduction and prevent a parameter-matched interpretation of the comparisons.

4.2. BiFPN Fusion and Training Objective

The BiFPN neck receives three feature levels from the backbone. Fine-resolution maps retain particle detail, intermediate maps represent compositional patterns, and coarse maps contain broader spatial information. Top-down and bottom-up paths exchange information between these levels. Learnable nonnegative weights determine the contribution of each incoming feature map. Fusion at level l follows this equation:

$$P_l = \frac{\sum_{j\in\mathcal{N}(l)} \alpha_j Resize_l(P_j)}{\varepsilon + \sum_{j\in\mathcal{N}(l)} \alpha_j}, \alpha_j \geq 0 \tag{4}$$

Here, $\mathcal{N}(l)$ denotes the incoming connections to the fusion node, and Resize_l aligns their spatial dimensions. The positive constant ε prevents numerical instability when the learned weights approach zero. Weight normalization controls feature magnitude and allows the model to vary its reliance on detailed texture and broader compositional evidence. Subsequent convolutional refinement processes the fused maps before pooling.

The regression head applies global average pooling followed by a two-layer perceptron and an output mapping based on the training ash range. Training must address numerical accuracy, contextual stability, and temporal consistency simultaneously. These requirements are combined in the objective equation:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \lambda_1\mathcal{L}_{inv} + \lambda_2\mathcal{L}_{smooth} \tag{5}$$

The regression term measures prediction error with contextual sample weighting. The invariance term penalizes representation differences across matched contextual groups, and the smoothness term discourages implausible prediction fluctuations between adjacent production intervals. The coefficients λ_1 and λ_2 control the contribution of the two regularizers. Temporal smoothing is disabled at detected batch boundaries, where abrupt changes in ash may be physically plausible. The exact invariance estimator, temporal penalty, matching tolerance, and batch-boundary detector are not specified. Their definitions are needed to distinguish the intended regularization from implemented constraints, particularly where a genuine ash transition should remain unsmoothed.

Training uses chronological validation for model selection. Mini-batches are shuffled within the training partition to reduce repeated updates from a single operating state. Context matching is restricted to source batches with overlapping ash ranges, and learning-rate warmup stabilizes the attention stages. Adjustment weights are clipped during early

training to prevent rare samples from dominating before the backbone learns basic sensing features. Validation performance is used to select regularization strengths and stopping time, with particular attention to whether improved average accuracy is accompanied by stable behavior across moisture and source groups. Optimizer settings, learning-rate schedule, batch size, regularization coefficients, random seeds, and stopping criteria are not available in the current description; no independently reproducible training configuration is established.

5. Experimental Data and Analysis

5.1. Dataset Construction and Baseline Configuration

The numerical case study represents a mixed-source coal preparation line. The supplied material does not include traceable raw measurements or assay records, so the numerical results are treated as illustrative rather than independently verified plant experiments. Each labeled sample corresponds to a five-minute interval for which belt images, spectral descriptors, process records, and laboratory ash assays were synchronized. Multi-source heterogeneous sensing with deformable attention has been investigated for concentrate-ash detection [16]. Chronological separation is used here to address temporal dependence. Table 1 specifies the physical content, sampling unit, preprocessing operation, and model role of each sensing channel. After removing intervals with missing image blocks or delayed laboratory tags, 12,480 samples remained. The training, validation, and test partitions contained 8,736, 1,872, and 1,872 samples, respectively, and the split was performed chronologically rather than by random shuffling to imitate deployment. The ash range was 6.1% to 34.8%, and the median was 18.7%.

Table 1. Dataset channels and preprocessing configuration

Channel	Raw content	Sampling unit	Preprocessing	Model use
RGB image	Belt surface image	5 min interval	Illumination normalization and window cropping	CoAtNet stem
NIR descriptor	Selected reflectance bands	5 min interval	Smoothing and z-score standardization	Spectral modulation
Process records	Moisture, size, speed, source, illumination	5 min interval	Scaling and embedding	Causal adjustment
Laboratory assay	Mass ash content	Composite sample	Chronological label alignment	Regression target

Figure 3(a) shows that the training and test partitions cover comparable ash intervals, with 42.6% of test samples located between 15% and 22% ash; Figure 3(b) shows that moisture and dominant size fraction create visible context clusters, including a high-moisture group above 9.5% moisture where brightness-based prediction becomes less reliable.

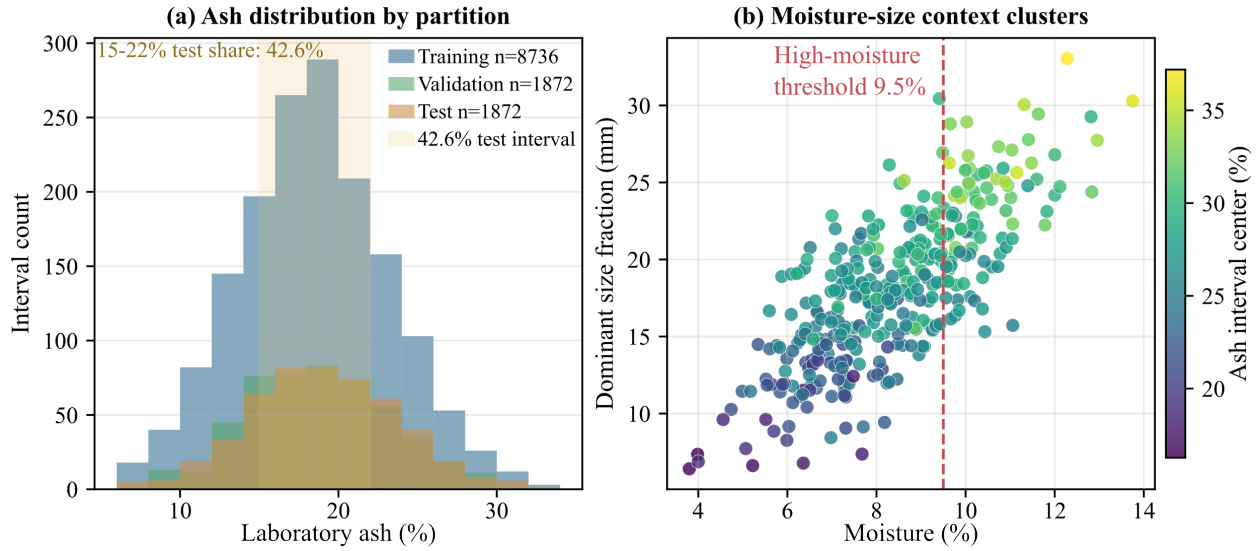


Figure 3. Dataset distribution and context structure: (a) ash-content distributions of training, validation, and test partitions; (b) moisture-size context clusters colored by ash interval

A time-based division is described to reduce the impact of similar particle characteristics and process conditions at the boundaries of adjacent belt sections. Randomly splitting the data may result in relatively similar distributions of the training and test sets, and thus the model's performance may be overestimated. Only for model selection and early stopping purposes, the validation set was used. The two source-batch transitions in the test partition are an extended high-moisture period and a short illumination recalibration. These events were recorded because the plants were working normally. Hyperparameters for all the models were selected before testing, and the same training, validation and test

indices were used to focus on the differences in models. Chronology alone does not exclude leakage from overlapping windows or shared composite assays. Interval identifiers, assay identifiers, and any exclusion gap at split boundaries are needed to verify independence; these records are not supplied.

The experiment also took some practical measures to prevent noise in the online ash prediction data. Only pictures taken with a stable load of the belt were selected, and intervals affected by belt stoppage were excluded from the model training. Averaged Spectral Descriptors over the same physical belt section as the image windows were used to reduce time-alignment errors between modalities. Laboratory tests were used to obtain the interval-level labels, and some image windows had the same ash value. During evaluation, the predictions of windows in the same interval were averaged to obtain interval-level metrics. The Design shows the application of an online monitoring system in a control room, and the operator responds to interval estimates rather than single-image tiles. The independent evaluation unit is therefore the assay-aligned interval, not an image crop. The assay standard, sampling duration, dry-basis convention, and synchronization tolerance are not specified, which limits assessment of label uncertainty.

Multimodal sensor fusion has also been studied for monitoring advanced manufacturing processes [17]. Its relevance here concerns heterogeneous observations, not evidence of coal-specific temporal degradation. Table 2 compares the test-set accuracy and inference latency of the six evaluated models under the same chronological split. The proposed model achieved an MAE of 0.52 percentage points, RMSE of 0.74 percentage points, MAPE of 3.18%, and R2 of 0.972. The strongest baseline, non-causal CoAtNet-BiFPN, reached 0.61 MAE and 0.85 RMSE, while EfficientNet reached 0.73 MAE and 1.02 RMSE. Classical SVR and PLSR were less stable, with RMSE values of 1.41 and 1.86 percentage points

Table 2. Test-set performance comparison

Model	MAE (*pp*)	RMSE (*pp*)	MAPE (*%*)	R^2	Latency (*ms*)
PLSR	1.34	1.86	8.42	0.846	4.1
SVR	1.02	1.41	6.73	0.902	9.8
ResNet50	0.81	1.12	4.98	0.941	18.9
EfficientNet	0.73	1.02	4.31	0.951	20.7
CoAtNet-BiFPN	0.61	0.85	3.66	0.965	22.8
Proposed model	0.52	0.74	3.18	0.972	23.6

Figure 4(a) compares MAE, RMSE, and MAPE across methods, where the proposed model provides a 12.9% RMSE reduction relative to the strongest deep baseline; Figure 4(b) gives the predicted-versus-measured scatter, where 91.4% of test points fall within the +/-1.0 percentage-point band. These are reported point estimates; without repeated-run results or paired interval-level uncertainty estimates, the comparison does not establish statistical significance.

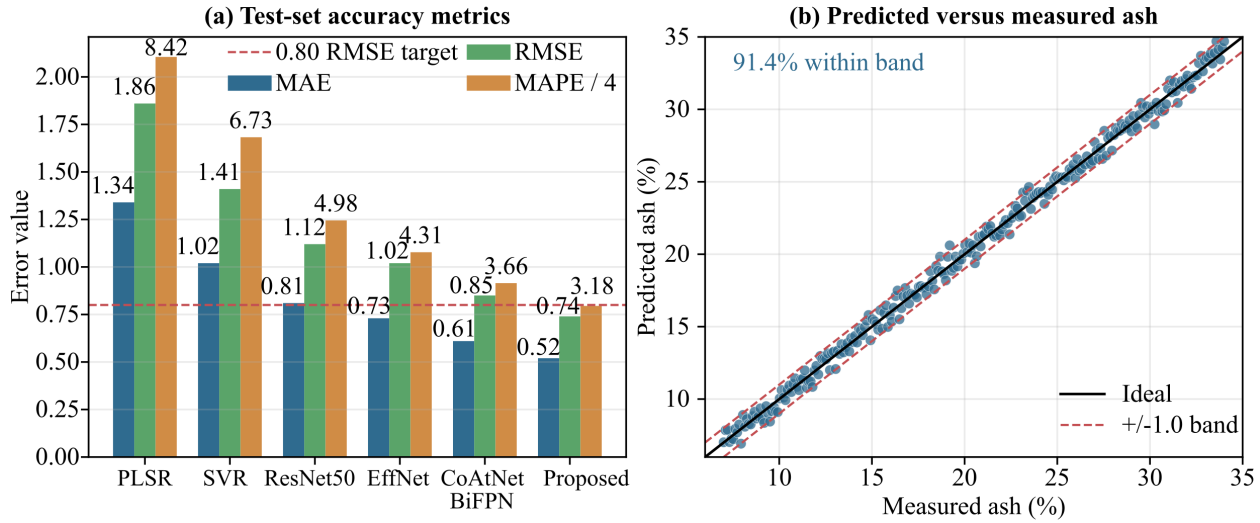


Figure 4. Prediction accuracy on the independent test set: (a) MAE, RMSE, and MAPE comparison across models; (b) predicted versus measured ash with +/-1.0 percentage-point acceptance band

Baseline models were built reasonably to prevent bias in the experiment. PLSR was employed to use standardized spectral and handcrafted image descriptors, and the number of latent components was determined by validation RMSE. SVR is employed, and for the penalty and kernel-width parameters, a grid search is conducted. ResNet50 and

EfficientNet were given the same image windows as the proposed model to obtain spectral descriptors via a small late-fusion branch. Non-causal CoAtNet-BiFPN baseline used the same backbone and fusion neck but did not change the weights or the invariance penalty. A baseline is chosen to rule out the possibility that the improved performance of CoAtNet-BiFPN is due to an increase in capacity rather than causal enhancement. Input descriptors differ across classical and deep baselines, so those comparisons evaluate complete prediction pipelines rather than architecture alone. The matched CoAtNet-BiFPN baseline is the closest control, but joint removal of weighting and invariance does not isolate their individual contributions.

Evaluation metrics are interpreted at the assay-aligned interval level. MAE and RMSE measure ash-content error in percentage points, whereas MAPE expresses relative error as a percentage of measured ash and gives greater weight to low-ash intervals. R^2 compares squared prediction error with variation around the mean measured ash; it is not an accuracy percentage. The tabulated latency is reported per image window after preprocessing. Window batching may improve throughput but does not imply that interval-level end-to-end latency is lower. Acquisition, preprocessing, synchronization, aggregation, and transfer delays are not included in that figure, and the missing hardware and batch-size information prevents independent timing verification.

Sensing Channels have different physical meanings and noise properties. RGB images provide dense spatial information but are affected by illumination and surface water. NIR descriptors are mineral- and moisture-sensitive reflectance data with relatively low spatial resolution. Records of processing are sparse but provide the collection environment and production status. Given the various input conditions, the reduction in RMSE from 0.85 for the non-causal deep baseline to 0.74 in the proposed model is modest in absolute terms. Its operational value depends on plant-specific tolerance limits and decision costs; no measured reduction in recirculation or blending cost is established by the prediction comparison.

The scatter plot of the prediction also shows the error as a function of ash. In this range, the concentration of ash is highest, at 14%-23%, and the model is close to the ideal line. Below 10% ash, the number of samples is relatively small; most of the predictions are within the acceptance range. At ash contents above 28%, some points indicate underprediction and falls in the larger residual range shown below. In general, high-ash intervals contain a mixture of gangue and coal particles, and the surface features are not directly related to the ash content determined in the laboratory. Based on the above analysis, this model is more suitable for process control of regular product tracking than for rare extreme-feed alarms unless an additional high-ash sample is taken.

5.2. Ablation Residual and Calibration Analysis

Ablation experiments were conducted to isolate the contribution of causal adjustment, CoAtNet extraction, and BiFPN fusion. Causal representation learning has been investigated for time-series prediction under external interference [18]. The ablations here evaluate predictive changes, not identification of physical causal effects. Removing the causal module increased RMSE from 0.74 to 0.85 percentage points and enlarged the error gap between dry and wet samples from 0.07 to 0.19 percentage points. Replacing BiFPN with simple feature concatenation increased RMSE to 0.91, mainly because the model became less effective for mixed particle sizes. Replacing CoAtNet with a pure convolutional backbone increased RMSE to 0.96 and reduced R2 to 0.953. Figure 5(a) shows the ablation RMSE pattern, Figure 5(b) presents the moisture-stratified MAE where the proposed model remains below 0.64 even when moisture exceeds 10%, Figure 5(c) displays batch-wise R2 with a minimum of 0.958 across four source batches, and Figure 5(d) reports inference latency where the proposed model requires 23.6 ms per image window on an industrial GPU.

The ablation values indicate different predictive penalties when modules are removed, but they do not by themselves establish distinct physical failure mechanisms. The larger moisture-related error gap without adjustment is consistent with contextual dependence; it does not demonstrate that rainy-season behavior has been measured. The increase after replacing BiFPN suggests that resolution fusion may be useful, although matched parameter counts are not reported. A convolutional backbone can also obtain a broad effective receptive field, so its higher error should not be attributed categorically to an inability to combine distant information. Independent repetitions and controlled variants would be needed to separate module effects from training variability.

The ablation and efficiency panels present accuracy, moisture response, batch stability, and latency; all are necessary for engineering applications. A model with a small mean RMSE but poor wet-sample performance would be unreliable in the wet season. A model that has a high R^2 on one batch of data will need to be manually retuned for another. A model with high accuracy but a long delay would be difficult to use in the existing plant supervision system. The model shown above is in a good area of the multi-objective space. It has relatively high latency compared to EfficientNet, but its reported latency is 2.9 ms higher per window, with an RMSE reduction of 0.28 percentage points. The timing comparison is provisional because the GPU model, numerical precision, batch size, and warmup procedure are not reported.

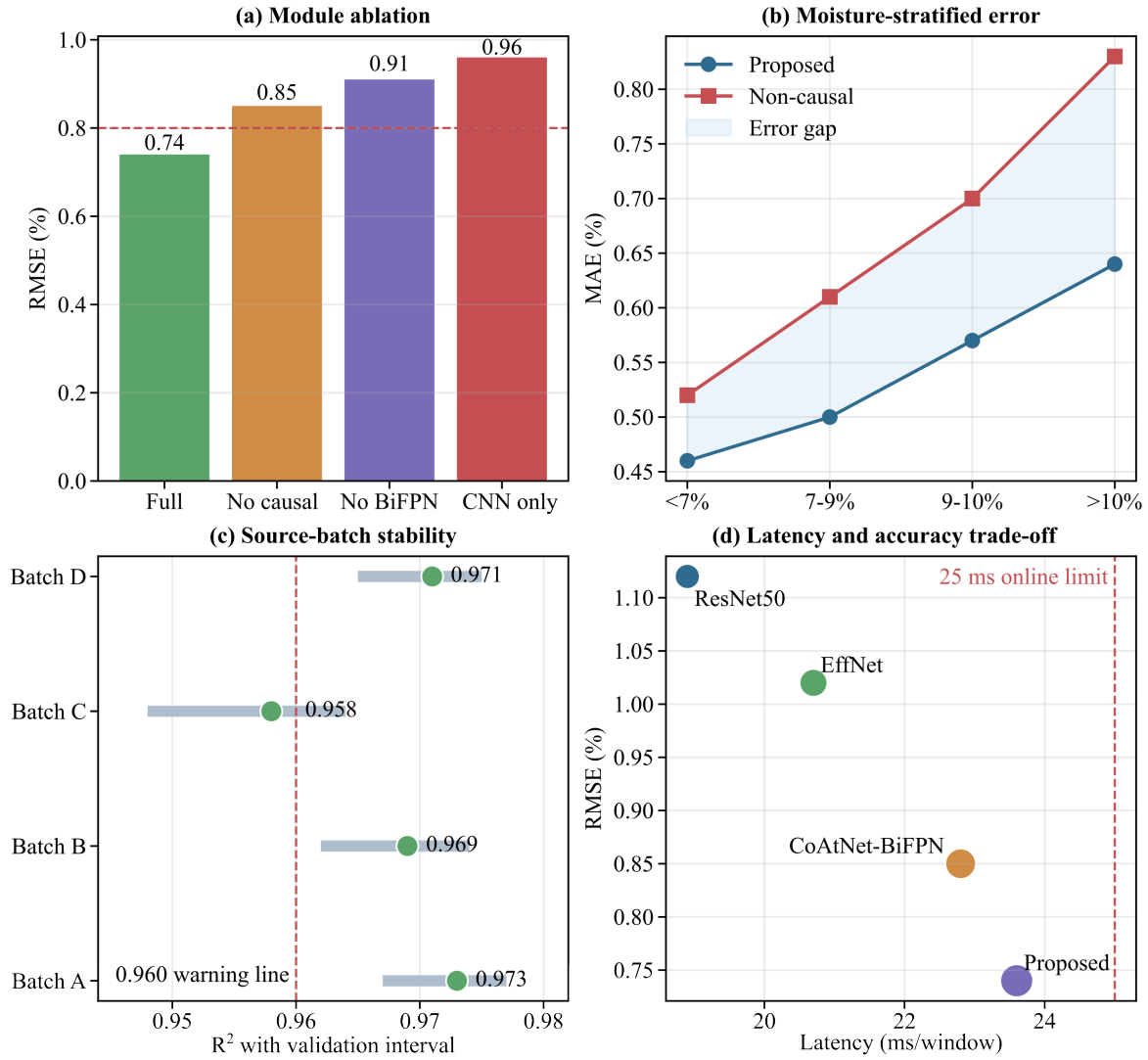


Figure 5. Ablation and efficiency results: (a) RMSE after removing key modules; (b) moisture-stratified MAE; (c) source-batch R2 stability; (d) inference latency and RMSE trade-off

Then the prediction residuals were examined to identify defects in the model. Causal feature decomposition has been investigated as a domain-generalization strategy [19]. Residual and quantile diagnostics are used here to examine bias and distributional agreement. Figure 6(a) is the residual distribution of low-, medium- and high-ash ranges, and their interquartile ranges are 0.48, 0.57 and 0.69 percentage points, respectively. High-ash Areas are relatively difficult; the mineral-rich particles have formed an uneven texture and spectral response.

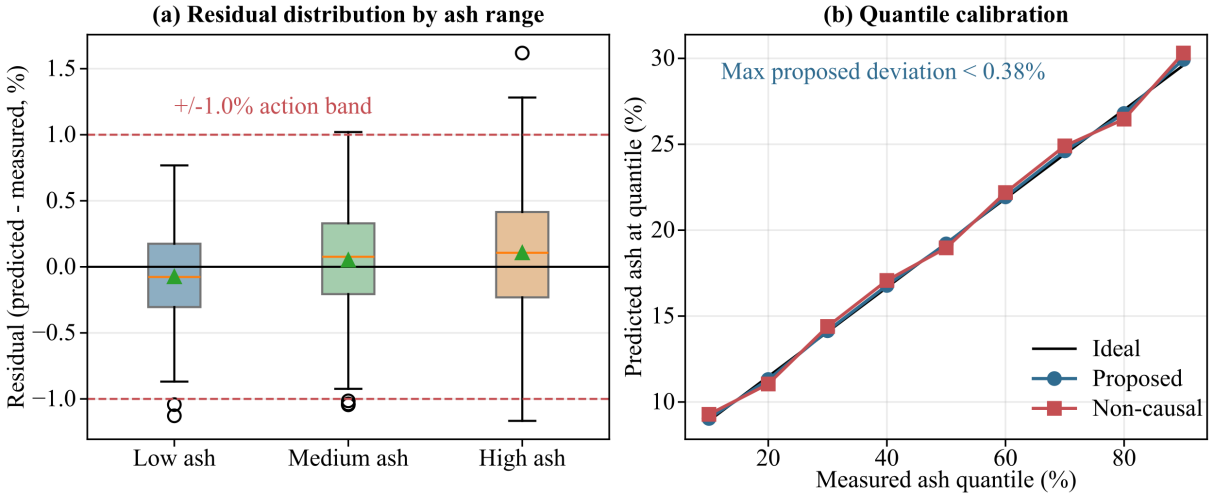


Figure 6. Residual and calibration analysis: (a) residual distributions across ash ranges; (b) quantile calibration curves of the proposed model and non-causal baseline

Figure 6(b) is the calibration curve of the predicted ash quantiles, and at the 10th and 90th percentiles, the model is within 0.38 percentage points of the ideal line. The reduction in positive residual skew from 0.42 to 0.18 in the high-moisture subset of the proposed model compared with the non-causal baseline indicates partial control of moisture-related brightness artifacts.

Residual inspection reveals information that aggregate statistics conceal. Low-ash samples may have a small error distribution when the visual surface is generally more uniform and mineral inclusions are scarce. Medium-ash samples are relatively more numerous in the dataset and show comparatively stable marginal quantiles. High-ash samples may have a larger residual spread when the mineral matter can be in the form of large gangue fragments, fine dispersion, or particles with a similar color but a different laboratory ash. The quantile comparison concerns marginal scale and ordering, not prediction-interval coverage or conditional bias. A predicted change cannot be identified as a genuine process change without considering assay uncertainty and residual variation; the mechanisms behind error differences are not isolated here.

Residual Distribution Provides the Basis for the Operating Threshold. If the plant is to be alarmed when the predicted ash exceeds the contract limit, then the uncertainty of high-ash intervals should be considered larger than that of medium-ash intervals. The quantile curve describes marginal distribution agreement, but cannot determine an alarm threshold or conditional coverage by itself. The model presented above is closer to the ideal quantile line than the non-causal baseline, and therefore, the predicted ash values have the same rank and scale in the test distribution. A Blending Operation: Operators need to know the ash content of the input streams to adjust accordingly. A model with biased quantiles can still have acceptable RMSE but lead to systematic overuse of low-ash coal.

5.3. Robustness and Context Sensitivity

Robustness tests evaluated feed-source shift, moisture disturbance, illumination drift, and reduced training labels. Domain generalization with limited training data has been studied as a distinct evaluation problem [20]. Transportable representations provide a framework for examining generalization across environments [21]. Figure 7(a) reports RMSE under a held-out source batch: the proposed model increased from 0.74 to 0.85, whereas EfficientNet increased from 1.02 to 1.37. Figure 7(b) shows moisture disturbance, where artificial wetness correction errors of +2 moisture percentage points increased MAE by 0.08 for the proposed model and by 0.21 for the non-causal CoAtNet-BiFPN. Figure 7(c) presents illumination drift tests with brightness scaling from 0.75 to 1.25; the proposed curve stays within a 0.16 RMSE band. Figure 7(d) shows that when training labels are reduced to 40%, the proposed model preserves an RMSE of 0.96, while the same backbone without causal adjustment rises to 1.14.

Robust validation research examines prediction reliability under distribution shift [22]. The structured disturbance tests were designed to resemble credible deployment changes rather than arbitrary noise. The held-out source test approximates a new mining panel. The moisture disturbance test approximates sensor correction error during rainy or high-spraying periods. The illumination drift test approximates gradual lamp aging or camera exposure mismatch. The reduced-label test approximates early deployment at a plant where laboratory assays are expensive. The proposed model does not remove all degradation, but its error growth is slower because the representation is discouraged from depending on a single context shortcut. This behavior is especially important for automated control, where a slowly degrading predictor is easier to recalibrate than one that fails abruptly.

Several robustness panels were established to assess the model under various conditions relevant to preparation plants. Held-out source test changes the geological background. A moisture disturbance test alters a measured context variable rather than physically wetting the coal. The illumination drift test scales brightness rather than reproducing complete lamp aging. The reduced-label test retains 40% of supervision. Held-out source transfer is not external-plant validation. Disturbance-generation details, source-selection rules, and subsampling seeds are not supplied, so the trends remain illustrative and do not establish robustness to uncontrolled field disturbances.

Further investigate the sources of errors using context-conditioned feature sensitivity. Adaptive temperature scaling addresses calibration of neural-network outputs [23]. It does not establish the causal validity of the context sensitivity assessment used here. A single heat map is used to avoid excessive representation of correlation matrices. Figure 8 shows the normalized sensitivity values of model response to contextual variables for four ash ranges. Moisture sensitivity is reduced from 0.62 in the non-causal baseline to 0.37 in the proposed model for the 20%-25% ash interval, and size-fraction sensitivity drops from 0.55 to 0.41. The belt speed in both models is low, less than 0.20, and this is consistent with the effect of exposure normalization in this dataset. The residual source-batch sensitivity shows that, after the adjustment mechanism, the reduction in geological context has not been achieved.

A sensitivity heatmap can be employed to conduct diagnosis, but it does not identify physical causal importance. Low sensitivity to a moisture input does not establish invariance to the optical effects of changing actual surface water. Residual source-batch sensitivity shows that some geological differences have been retained in the representation. Source batch may contain compositional information and incidental acquisition associations; sensitivity cannot separate

them. Perturbation magnitude, aggregation, and normalization are unspecified, preventing score reproduction or interpretation as physical effect sizes. The heatmap remains an illustrative model-dependence diagnostic.

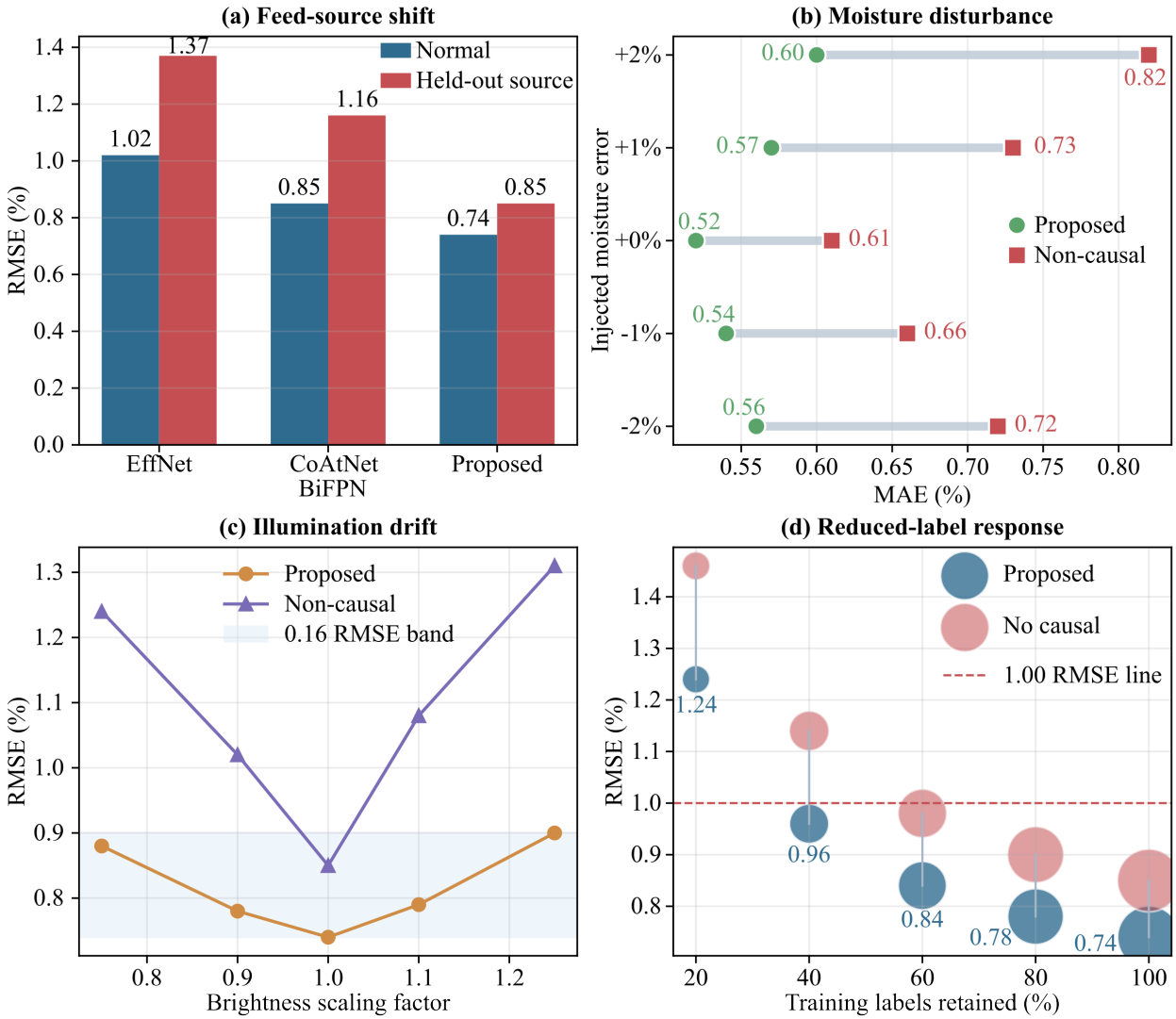


Figure 7. Robustness under structured disturbances: (a) held-out source-batch RMSE; (b) moisture disturbance response; (c) illumination drift response; (d) reduced-label training response

For deployment, the prediction pipeline can be connected to the existing image and spectral acquisition units. Window-level inference is relatively computationally expensive; therefore, a small industrial GPU will be used for batching. The causal adjustment encoder only needs the process record that has been saved in the supervisory system. Inference does not need laboratory ash labels, and the adjustment weights in the training do not increase the load of online data entry. Uncertainty-aware calibration has been examined in fault-diagnosis systems [24]. Transfer to coal-ash monitoring requires application-specific validation. Therefore, periodic collection of labels for calibration monitoring is still required. The most suitable trigger for recalibration is not only the passage of time but also a combination of source-batch novelty, expansion of moisture range and sustained residual bias in recent laboratory tests.

Some experimental limitations should be considered in light of the results. The range of the dataset for ash is broad within the described operating range but insufficient to establish extrapolation; therefore, extrapolation outside this range cannot be guaranteed. The context variables are in interval-sampling mode, and thus a sudden increase in belt loading or illumination will be smoothed out before being used in the model. There are errors in the labels, and the composite sample may not align perfectly with all the image windows. All the models are trained on the same data, but the defects are not invalid for comparison; instead, they show why residual errors still occur after causal adjustment and multi-scale fusion. Raw-data provenance, repeated-run variability, and external-plant testing are unresolved evidence limitations; the numerical case study cannot establish certified measurement performance.

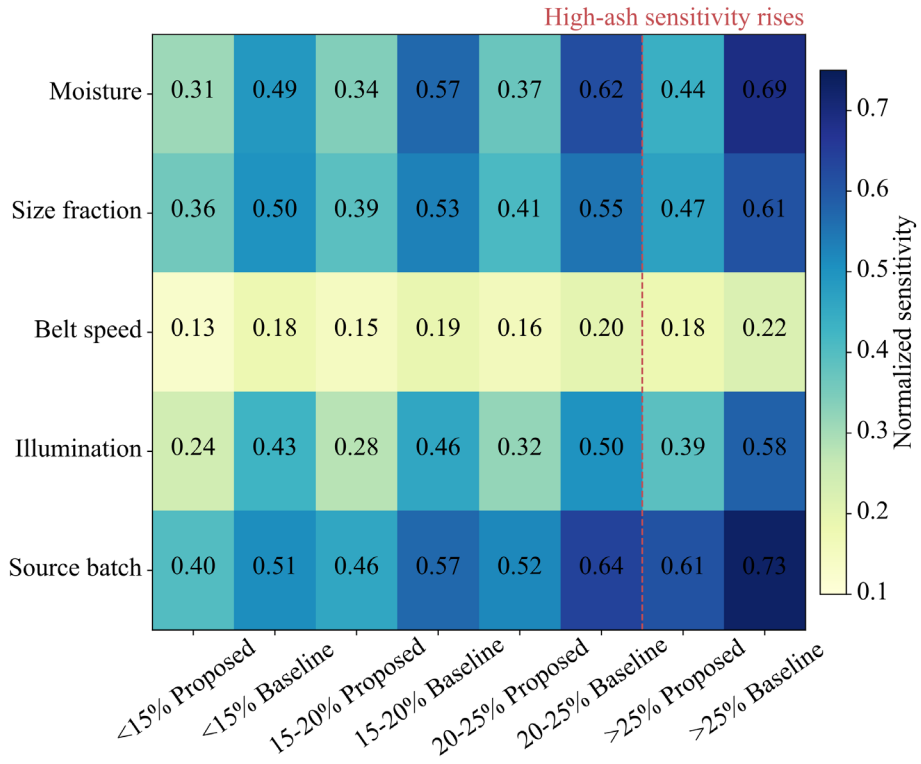


Figure 8. Context-conditioned feature sensitivity heat map comparing the proposed model with the non-causal baseline across ash intervals

Quality control indicators can be derived from the numbers of errors. No universal MAE threshold for early warning is established here. The reported MAE is 0.52 percentage points, and alarm thresholds require plant-specific tolerance limits, assay uncertainty, and false-alarm costs. It should not be used in place of certified assays for commercial settlement, as the sampling and measurement chains are different. The most suitable use case is supervisory decision support: recognize extended ash drift, compare it with the status of the incoming feed, and determine whether to prioritize a laboratory sample. A stable predictor is preferable to an unstable one that is only slightly better in performance on an easily obtained random split for this purpose.

Uncertainty-quantification metrics for deep regression provide a basis for evaluating predictive uncertainty [25]. Qualitative uncertainty is indicated by the repeated validation and grouping of residuals; a full probabilistic interval has not been obtained from the current model. Intervals with a high moisture content, large size fluctuations, or infrequent source batches had a larger residual spread. Based on the above observations, an uncertainty head can be added without changing the main structure. The context encoder may provide a diagnostic signal for instability, but it has not identified a physical cause. Any interval estimate would require a specified probabilistic method and separate coverage validation. The system can help the processing engineer identify abnormal behavior that is not covered by the normal prediction noise in a short amount of time.

Concept-drift adaptation has been studied for industrial data streams using knowledge embedding and transfer [26]. A possible maintenance workflow is described for subsequent evaluation; no installed-system maintenance trial is reported. A fixed model will be relatively accurate given the changes in feed characteristics, but coal preparation plants are rarely in a stable state. The recommended maintenance mode is to archive the sensing data, process records, predictions and later laboratory tests in the same interval format as in this study. New labels can then be used for the periodic recalibration of the regression head, and it can be checked whether causal sensitivities have changed. Whether head-only recalibration is sufficient must be assessed against a frozen baseline using later assay-aligned intervals; no adaptation benefit is measured in this case study.

6. Conclusion

This paper introduces a causal-informed CoAtNet-BiFPN model for online coal ash prediction at coal preparation plants. The method introduces causal adjustment for contextual process variables, employs convolution-attention feature extraction of image and spectral evidence, and conducts bidirectional feature fusion of multi-scale particle representations. The paper has presented the motivation, put forward the problem of causal learning, introduced the architecture, and described an illustrative numerical comparison and ablation assessment. The contribution is an

integrated engineering formulation, not a new backbone operator or proof that adjustment identifies physical causal effects. Its intended role is supervisory quality assessment rather than certified measurement.

Several limitations remain. The assumed causal structure omits potentially influential disturbances, and the interventional representation contrast is not identified by the available observational description. The numerical results are illustrative because traceable sensing and assay records are not supplied. Architectural dimensions, weighting estimators, penalty definitions, and training settings are insufficiently specified for exact reproduction. Reported latency excludes parts of the acquisition and processing chain, and residual grouping does not establish calibrated prediction intervals. The setting also lacks independently documented validation across equipment layouts, coalfields, and preparation plants, which restricts the evidence for operational transfer.

Future research will focus on the expansion to multi-plant verification, online updates and uncertainty-aware process control. A further direction is to connect causal feature learning with adaptive sampling, and based on whether the model finds an unfamiliar operating state, prioritize a laboratory test. A subsequent direction is to combine ash prediction with the downstream control objective and have the model support separation-density adjustment and product blending decisions in a closed-loop optimization system. Traceable multi-plant records, controlled interval-level ablations, validated uncertainty coverage, and frozen-baseline drift monitoring should precede closed-loop use. Control stability and operational benefit require separate prospective assessment.

7. Declarations

7.1 Author Contributions

Marcin Skowroński contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Zuzanna Lorenc contributes to validation, analysis, investigation, data collection, draft preparation. Emilia Lipińska contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

7.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

7.3 Institutional Review Board Statement

Not applicable.

7.4 Informed Consent Statement

Not applicable.

7.5 Date Availability

Not applicable.

7.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

8. References

[1] Zhang, K., Wang, W., Cui, Y., Lv, Z., & Fan, Y. (2024). GFNet: A pioneering approach for precisely estimating ash content in coal through the fusion of graph convolution and feedforward network. Engineering Applications of Artificial Intelligence, 127, 107301. https://doi.org/10.1016/j.engappai.2023.107301

[2] Zhang, K., Wang, W., Cui, Y., Lv, Z., Fan, Y., & Zhao, X. (2024). Deep learning-based estimation of ash content in coal: Unveiling the contributions of color and texture features. Measurement, 233, 114632. https://doi.org/10.1016/j.measurement.2024.114632

[3] Li, J., Gao, R., Zhang, Y., Zhang, L., Dong, L., Ma, W., Yin, W., & Jia, S. (2025). Research on accurate analysis of coal quality using NIRS-XRF fusion spectroscopy in complex coal type scenarios. Optics & Laser Technology, 181, 111734. https://doi.org/10.1016/j.optlastec.2024.111734

[4] Suprapto, S., Wahyuningtyas, A., Madurani, K. A., & Ni'mah, Y. L. (2025). Predicting ash content and water content in coal using full infrared spectra and machine learning models. South African Journal of Chemical Engineering, 51, 170–179. https://doi.org/10.1016/j.sajce.2024.11.006

[5] Lu, F., Liu, N., & Liu, H. (2024). Investigating the influence of froth image attributes on clean coal ash content: A novel hybrid model employing deep learning and computer vision techniques for prediction exploration. Minerals, 14(6), 536. https://doi.org/10.3390/min14060536

[6] Li, Y., Liu, H., & Lu, F. (2024). Research on prediction of ash content in flotation-recovered clean coal based on NRBO-CNN-LSTM. Minerals, 14(9), 894. https://doi.org/10.3390/min14090894

[7] Liu, H., & Liu, J. (2024). Research on the detection principle of coal ash by X-ray transmission based on FLUKA. Minerals, 14(11), 1079. https://doi.org/10.3390/min14111079

[8] Liu, Z., Li, L., Zeng, J., Wang, Y., Yang, J., & Liu, X. (2024). Predicting the ash content in coal flotation concentrate based on convolutional neural network. International Journal of Coal Preparation and Utilization, 44(12), 2080–2096. https://doi.org/10.1080/19392699.2024.2308549

[9] Gao, R., Li, J., Dong, L., Wang, S., Zhang, Y., Zhang, L., Ye, Z., Zhu, Z., Yin, W., & Jia, S. (2024). Accurate analysis of coal calorific value using NIRS-XRF: Utilizing RF classification and PLSR subtype modeling. Microchemical Journal, 201, 110716. https://doi.org/10.1016/j.microc.2024.110716

[10] Wang, L., Liu, J., Wang, H., Nan, J., Gui, X., & Dai, W. (2026). Intelligent real-time ash content detection for coal flotation concentrate using multi-source data fusion. Fuel, 406, 137132. https://doi.org/10.1016/j.fuel.2025.137132

[11] Wen, Z., Yang, S., Zhou, M., Wang, G., & Zhou, C. (2026). Interpretability-guided heterogeneous stacking ensemble for online monitoring of tailings ash content in coal flotation processes. Colloids and Surfaces A: Physicochemical and Engineering Aspects, 728, 138765. https://doi.org/10.1016/j.colsurfa.2025.138765

[12] Cheng, K., Zhang, X., Zhou, K., Zhou, C., Li, J., Yang, C., Guo, Y., & Wang, R. (2025). Time series prediction method of clean coal ash content in dense medium separation based on the improved EMD-LSTM model. Big Data and Cognitive Computing, 9(6), 159. https://doi.org/10.3390/bdcc9060159

[13] Yang, X., Zhang, K., Wang, T., Xie, G., Thé, J., Tan, Z., & Yu, H. (2025). Multi-scale multi-task neural network combined with transfer learning for accurate determination of the ash content of industrial coal flotation concentrate. Minerals Engineering, 220, 109093. https://doi.org/10.1016/j.mineng.2024.109093

[14] Yang, X., Zhang, K., Thé, J., Tan, Z., & Yu, H. (2025). Multi-scale neural network for accurate determination of the ash content of coal flotation concentrate using froth images. Expert Systems with Applications, 262, 125614. https://doi.org/10.1016/j.eswa.2024.125614

[15] Zhang, C., Dou, D., Peng, Y., & Xie, G. (2025). Froth-tailings fusion image-based concentrate ash content prediction in coal flotation using stacking model. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 47(1), 31–48. https://doi.org/10.1080/15567036.2025.2451073

[16] Zhang, C., Tan, J., & Xie, G. (2026). Concentrate ash detection based on multi-source heterogeneous data and deformable attention network in coal flotation. Powder Technology, 467, 121579. https://doi.org/10.1016/j.powtec.2025.121579

[17] McKinney, M., Garland, A., Cillessen, D., Adamczyk, J., Bolintineanu, D., Heiden, M., Fowler, E., & Boyce, B. L. (2025). Unsupervised multimodal fusion of in-process sensor data for advanced manufacturing process monitoring. Journal of Manufacturing Systems, 78, 271–282. https://doi.org/10.1016/j.jmsy.2024.12.003

[18] Feng, X., Fan, D., Jiang, S., Zhang, J., Guo, B., Ding, X., Hu, D., & Jiang, Y. (2024). A causal representation learning based model for time series prediction under external interference. Information Sciences, 663, 120270. https://doi.org/10.1016/j.ins.2024.120270

[19] Li, S., Zhao, Q., Sun, B., Wang, X., & Zou, Y. (2024). Domain generalization via causal fine-grained feature decomposition and learning. Computers & Electrical Engineering, 119, 109548. https://doi.org/10.1016/j.compeleceng.2024.109548

[20] Chen, K., Gal, E., Yan, H., & Li, H. (2024). Domain generalization with small data. International Journal of Computer Vision, 132, 3172–3190. https://doi.org/10.1007/s11263-024-02028-4

[21] Jalaldoust, K., & Bareinboim, E. (2024). Transportable representations for domain generalization. Proceedings of the AAAI Conference on Artificial Intelligence, 38(11), 12790–12800. https://doi.org/10.1609/aaai.v38i11.29175

[22] Cauchois, M., Gupta, S., Ali, A., & Duchi, J. C. (2024). Robust validation: Confident predictions even when distributions shift.

Journal of the American Statistical Association, 119(548), 3033–3044. https://doi.org/10.1080/01621459.2023.2298037

[23] Balanya, S. A., Maroñas, J., & Ramos, D. (2024). Adaptive temperature scaling for robust calibration of deep neural networks. Neural Computing and Applications, 36, 8073–8095. https://doi.org/10.1007/s00521-024-09505-4

[24] Lin, Y.-H., & Li, G.-H. (2024). Uncertainty-aware fault diagnosis under calibration. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54(10), 6469–6481. https://doi.org/10.1109/TSMC.2024.3427345

[25] Lind, S. K., Xiong, Z., Forssén, P. E., & Krüger, V. (2024). Uncertainty quantification metrics for deep regression. Pattern Recognition Letters, 186, 91-97. https://doi.org/10.1016/j.patrec.2024.09.011

[26] Zhang, Y., Yang, C., Liu, Z., Lou, S., Huang, X., & Wang, S. (2025). Concept drift meets industrial data streams: An efficient drift adaptation framework with knowledge embedding and transfer. Expert Systems with Applications, 290, 128439. https://doi.org/10.1016/j.eswa.2025.128439