Characterizing Operational Drift in Cement Manufacturing Process Data via Self-Supervised Representation Learning

A Multi-Year Operational Analysis Using PELT Change-Point Detection and SCARF Embeddings

Changgyun Kim
Department of Advanced AI Engineering, Samcheok, South Korea

***Abstract*—Cement finish milling generates large volumes of process variable data at hourly resolution, while quality measurements (Blaine fineness, 44 μm residue) are recorded only every two to four hours, producing a sparse-label regime with substantial unlabeled data accumulated over multi-year operation. In this study, we analyze 188,858 hourly records collected from three industrial cement mill units spanning 2017 to 2025 (approximately nine years) and characterize the operational drift structure embedded in the data. Using Pruned Exact Linear Time (PELT) change-point detection on three critical process variables, we identify between 12 and 21 detected drift events per mill, with Cohen's d effect sizes exceeding 1.0 for the majority of detected change points and reaching 4.3 in extreme cases. We then apply SCARF, a contrastive self-supervised learning method for tabular data, to learn 128-dimensional representations on the combined labeled (51,225 records) and unlabeled (111,506 records) data. Multi-seed training yields stable validation InfoNCE loss of 6.93 ± 0.02. Three clustering algorithms applied to the learned embedding space (K-means, Gaussian Mixture Model, HDBSCAN) consistently select 15 to 21 operational regimes, with silhouette scores between 0.36 and 0.41. The Adjusted Rand Index between embedding-space K-means and process variable space K-means is 0.35, indicating that the learned representation preserves coarse regime structure while resolving finer sub-regime variability. Cluster analysis further reveals strong mill specificity, with 14 of 15 embedding clusters dominated by a single mill, and temporal cluster evolution that aligns with PELT-detected change-point boundaries. These findings establish that long-term cement process data contains a richer operational regime structure than implied by raw process variable clustering, and that self-supervised pretraining can recover this structure, and that the resulting representation yields statistically significant gains in downstream quality prediction.**

***Keywords*—Cement manufacturing; self-supervised learning; change-point detection; operational drift; tabular data; representation learning; SCARF; PELT**

I. Introduction

Cement manufacturing is a globally significant industrial process responsible for an estimated 5 per cent of anthropogenic carbon dioxide emissions [1]. The finish milling stage, in which clinker is ground together with gypsum and supplementary cementitious materials, determines the principal product quality indices: Blaine fineness (cm²/g) and the residue retained on a 44 μm sieve (44 μmR, %). Stable control of these indices is essential because deviations propagate directly into customer-facing product specifications and into the energy efficiency of subsequent hydration and strength development.

Industrial cement facilities typically collect hourly process variable (PV) measurements from distributed control system (DCS) sensors, producing dense multivariate time series at the input side. However, quality measurements depend on laboratory sampling and are recorded only every two to four hours. This temporal asymmetry yields a structurally sparse-label regime in which only a small fraction of process records can be directly paired with quality labels for supervised learning. Recent benchmarks in the tabular learning literature [13] have reported that gradient-boosted tree ensembles remain competitive with or superior to state-of-the-art deep tabular architectures at sample sizes typical of industrial sparse-label settings, a pattern also observed in preliminary analysis of this facility's data.

Initial inspection of the full dataset revealed that pre-2023 Blaine measurements exhibit substantial distributional deviation relative to post-2023 records. Restricting analysis to the 2023 to 2025 period to avoid these artifacts reduced the available process records by approximately 87 percent (from 188,858 to 24,878) and limited the data span to roughly three years. The present analysis revisits this design choice by treating the pre-2023 period as a candidate source of unlabeled and partially labeled data, applying a wider valid-range filter that retains those pre-2023 records whose Blaine and 44 μmR fall within plausible operational bounds while excluding records exhibiting clear out-of-range artifacts. This study then investigates whether the retained pre-2023 records, combined with all available 2023 to 2025 records, contain a coherent operational regime structure that can be recovered through self-supervised representation learning.

The contributions of this study are threefold. First, we apply Pruned Exact Linear Time (PELT) change-point detection [2] with automatic penalty selection to three critical process variables across all three mills and characterize the temporal drift structure embedded in approximately nine years of operational data. Second, we apply SCARF [3], a contrastive self-supervised learning method specifically designed for tabular data, to learn 128-dimensional representations on the

combined labeled and unlabeled data. Third, we conduct a systematic clustering analysis on the learned embedding space using K-means, Gaussian Mixture Models, and HDBSCAN, and validate the resulting regimes against the change-point structure obtained independently through PELT.

The remainder of this study is organized as follows: Section II reviews related work on cement quality modeling, self-supervised learning for tabular data, and change-point detection. Section III describes the dataset and preprocessing pipeline. Section IV characterizes operational drift using PELT and K-means clustering on raw process variables. Section V describes the SCARF self-supervised pretraining method. Section VI presents embedding-space clustering and interprets the discovered regime structure with respect to mill identity and temporal evolution. Section VII discusses implications for downstream quality prediction and limitations of the present study. Section VIII concludes.

II. Related Work

A. Quality Prediction in Cement Manufacturing

Data-driven modeling in the cement and process industries has been surveyed extensively. Ge et al. [4] reviewed machine learning techniques applied across process-industry tasks, organizing approaches by supervision regime and identifying soft-sensor modeling as one of the principal application areas. Yuan et al. [5] introduced a deep stacked autoencoder with variable-wise weighting for soft-sensor modeling in nonlinear dynamic industrial processes. Pani and Mohanta [6] applied support vector regression, fuzzy inference, and adaptive neuro-fuzzy inference for online monitoring of cement fineness. Yuan et al. [7] developed a weighted linear dynamic system for feature representation in soft-sensor applications across nonlinear dynamic industrial processes. Zermane and Drardja [8] developed a random forest based monitoring system for cement production. Although the cited soft-sensor and process-industry studies do not directly compare tree-based and deep alternatives, recent benchmarks in the broader tabular learning literature [13] have reported that gradient-boosted tree ensembles remain competitive with or superior to the state-of-the-art deep tabular architectures at the sparse-label sample sizes typical of cement finish-milling operations.

B. Self-Supervised Learning for Tabular Data

Self-supervised learning (SSL) for tabular data has emerged as a methodology for leveraging the abundant unlabeled records present in industrial process data. VIME [9] learns representations through value reconstruction and masking. SCARF [3] adopts a contrastive framework in which positive views are generated through feature corruption (random replacement from the marginal distribution) and the encoder is trained with InfoNCE loss [10]. SubTab [11] partitions feature subsets and learns through cross-subset reconstruction. TabTransformer [12] applies attention-based encoders to tabular inputs. Recent benchmarks [13] show that SSL pretraining can narrow the performance gap between deep tabular models and gradient-boosted trees when downstream labeled data is limited, although these results have not been systematically validated on industrial datasets exhibiting operational drift, which is the setting of the present study.

C. Change-Point Detection and Operational Drift Identification

Change-point detection [14] aims to identify time points, where the statistical properties of a signal change. PELT [2] achieves linear time complexity through optimal pruning, with the segmentation penalty parameter controlling the number of detected change points. For multivariate signals or single signals with mean-shift transitions, the L2 cost model is commonly applied. Drift identification in industrial process monitoring has traditionally relied on statistical process control charts [15]; more recent approaches use machine learning-based concept drift detection [16]. In this study, we apply PELT with automatic penalty selection to recover the operational drift structure in cement process variables.

III. Dataset and Preprocessing

A. Data Source

The dataset consists of DCS records from three cement finish milling units, designated CM2, CM3, and CM4, at a commercial cement facility in the Republic of Korea, covering the period from January 2017 to October 2025. The raw dataset comprises 188,858 hourly records, each associated with 39 process variables (mill drive load, separator load, bag filter differential pressure, ball mill power, feed rates of clinker, gypsum, slag and fly ash, classifier and separator rotational speeds, and auxiliary roller and pressure measurements) and two quality target variables (Blaine fineness and 44 μmR). All variables were already aggregated to hourly resolution by the facility's historian system.

B. Labeled and Unlabeled Partitioning

Quality measurements at the finish milling stage are recorded approximately every two to four hours, while process variables are sampled hourly. To construct a valid labeled set, we apply a backward-aligned causal pairing rule: a process record at time t is paired with the most recent prior or simultaneous laboratory quality measurement that satisfies all of the following conditions: (1) the time gap is at most three hours, (2) Blaine fineness falls within [2,500, 5,000] cm^2/g, and (3) 44 μmR falls within [0, 25]%. Records that fail either quality range condition are excluded from the labeled set rather than being relegated to the unlabeled pool, because their associated quality values would otherwise contaminate downstream supervised training. The unlabeled set comprises records for which no laboratory quality measurement exists within the causal window.

This partitioning produces a labeled set D_1 of 51,225 records and an unlabeled set D_unlab of 111,506 records (see Table I). Relative to a period-restricted preliminary analysis of the same facility data, D_1 expands the labeled set through two mechanisms operating jointly: (1) the wider valid-range filter applied uniformly across the full 2017 to 2025 period admits pre-2023 records that pass both Blaine and 44 μmR range conditions, and (2) the wider range admits additional 2023 to 2025 records that were excluded under the stricter range. Approximately two-thirds of the increase in D_1 originates from the 2023 to 2025 period, while approximately one-third originates from pre-2023 records that satisfy both range conditions. The remaining pre-2023 records that fail either range

condition are excluded from both D_1 and D_unlab; only records lacking quality labels enter D_unlab.

The labeled set D_1 is partitioned into training (35,854 records, 70 per cent), validation (7,680 records, 15 per cent), and test (7,691 records, 15 per cent) subsets using a regime-stratified random split. A K-means clustering is fitted on D_1 in the 40-dimensional space formed by the 37 standardized process variables described in Section III C, concatenated with three mill one-hot indicator dimensions, with the number of clusters automatically selected by silhouette score over $k \in \{3, \ldots, 8\}$, yielding $k = 8$. A 70/15/15 random split is then performed within each of the eight clusters to produce regime-balanced train, validation, and test subsets. The same eight clusters are reused as the raw-PV reference clustering in Section IV B, so that the comparison between raw-PV and embedding-space clusterings is based on a single, fixed PV-space partition rather than two independently fitted K-means models. This protocol is adopted to ensure regime-balanced evaluation across the full 2017 to 2025 data span and to allow future comparison with downstream supervised tasks using the same partition. The fitted K-means model is subsequently applied to D_unlab via nearest-centroid assignment to provide regime labels for all 162,731 records used in the ARI computation of Section VI. We acknowledge that the random subset selection from a multi-year time-series introduces residual temporal contiguity between training and test records and is, in this sense, weaker than a strictly chronological split for assessing temporal generalization; this trade-off is discussed further in Section VII D.

C. Process Variable Preprocessing

From the 39 process variables, two with greater than 50 per cent missing values across the full 2017 to 2025 period are removed (POLYCOM_IMPELLER_CRUSHER and ETC_GRINDING_AID), yielding 37 retained PVs. Records with more than 50 per cent of PVs missing are treated as plant shutdown periods and are excluded from both D_1 and D_unlab. Remaining missing values within retained records are imputed using per-mill forward fill followed by backward fill. Multivariate outliers are then filtered using IsolationForest with contamination 0.02 and 200 trees, fitted jointly on the combined $D_1 \cup$ D_unlab feature matrix. The remaining 37 PVs, together with the mill identifier (one-hot encoded over CM2, CM3, CM4), constitute the 40-dimensional input feature space for both representation learning and downstream tasks.

TABLE I. DATASET SUMMARY

Subset	Mills	Records	Time range	Purpose
D_1 train	CM2/3/4	35,854	2017–2025	Future supervised training
D_1 val	CM2/3/4	7,680	2017–2025	Hyperparameter selection
D_1 test	CM2/3/4	7,691	2017–2025	Held-out evaluation
D_unlab	CM2/3/4	111,506	2017–2025	SSL pretraining
Total	CM2/3/4	162,731	2017–2025	—

[a] Splits are regime-stratified random subsets of D_1; time range refers to the data span, not subset-specific intervals. Total raw records: 188,858. Difference comprises records dropped by shutdown filtering, IsolationForest outlier filtering, and quality value range filtering (records failing either Blaine or 44 μmR range conditions are excluded entirely).

IV. OPERATIONAL DRIFT CHARACTERIZATION

Fig. 1 provides an overview of the complete analysis pipeline. The raw multivariate process data, after the preprocessing described in Section III, form a 40-dimensional input consisting of 37 standardized process variables and three mill one-hot indicators. This input feeds two complementary analyses. First, PELT change-point detection is applied to selected critical signals to characterize the temporal operational drift directly in the process-variable space (Section IV-A). Second, the same input is used for SCARF contrastive self-supervised pretraining, in which a shared encoder f and projection head g are trained under an InfoNCE objective on corrupted and uncorrupted feature views to produce a 128-dimensional representation (Section V). The learned embedding is then clustered to recover operational regimes, which are validated against the independently obtained change-point structure (Section VI). The remainder of this section begins with the PELT analysis.

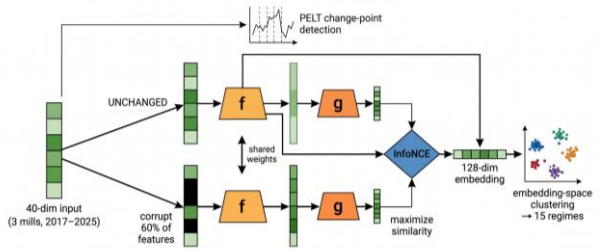

Fig. 1. Overview of the proposed pipeline (preprocessing, PELT drift detection, SCARF pretraining, and embedding-space clustering).

A. PELT Change-Point Detection

To characterize operational drift, we apply PELT change-point detection [2] to three operationally critical process variables: MILL_186_BF_PRESSURE (bag filter differential pressure, indicating grinding fineness), POLYCOM_WF_FEED_CLINKER (clinker feed rate, the dominant material input), and MILL_SEPOL_SEPOL (separator rotational speed). Each variable is aggregated to a daily resolution per mill, with a minimum of six valid hourly records required for valid daily aggregation. The L2 cost model is used with a minimum segment size of 60 days. The penalty parameter is selected automatically through grid search over {100, 300, 1,000, 3,000, 10,000, 30,000, 100,000, 300,000}, with a target range of 3 to 8 change points per signal per mill. When no penalty in the grid yields a change-point count within the target range, the penalty producing the count closest to the upper bound (8) is selected; this fallback applies to one signal-mill combination in the present analysis, CM4 POLYCOM_WF_FEED_CLINKER, which yields 10 change points at penalty 3,000 (the closest available value above the target upper bound). All other eight signal-mill combinations fall within the [3, 8] target range without fallback.

These three variables were selected to span the three principal and largely independent axes that govern finish-milling behavior rather than to exhaust the variable set: grinding fineness at the separation stage (MILL_186_BF_PRESSURE, the bag-filter differential pressure that reflects the attained product fineness), the dominant material input to the circuit

(POLYCOM_WF_FEED_CLINKER, the principal feed stream), and the classification control that sets the cut size (MILL_SEPOL_SEPOL, the separator rotational speed). This choice is further supported by an independent supervised quality-prediction analysis on data from the same facility, in which bag-filter differential pressure and separator-related variables emerge among the most influential predictors of both Blaine fineness and 44 μmR. Restricting the change-point analysis to these three complementary signals, therefore, targets the axes most likely to carry operationally meaningful drift while keeping the per-signal segmentation interpretable; a comprehensive multivariate change-point analysis over all 37 process variables is identified as future work in Section VII E.

Effect sizes are quantified using Cohen's d between adjacent segments, where $d = (\mu_2 - \mu_1) / s_{pooled}$ and the pooled standard deviation is computed across the two adjacent segments. By convention, $|d| > 0.8$ is interpreted as a large effect and $|d| > 1.2$ as very large. Table II summarizes the detected change points per mill and per process variable.

TABLE II. PELT CHANGE-POINT DETECTION RESULTS PER MILL

Mill	MILL_186_BF_PRESSURE	POLYCOM_WF_FEED_CLINKER	MILL_SEPOL_SEPOL
CM2	4 CPs (|d|max = 3.37)	3 CPs (|d|max = 1.50)	5 CPs (|d|max = 4.34)
CM3	7 CPs (|d|max = 2.93)	8 CPs (|d|max = 2.18)	6 CPs (|d|max = 2.85)
CM4	5 CPs (|d|max = 2.46)	10 CPs (|d|max = 2.05)	4 CPs (|d|max = 2.34)

[b.] CP = change point. |d|max denotes the maximum Cohen's d among all adjacent segment pairs for that variable and mill. The 10-CP count for CM4 POLYCOM_WF_FEED_CLINKER reflects the fallback rule described in the text.

The total number of detected change points per mill ranges from 12 (CM2) to 21 (CM3), with the majority of effect sizes satisfying $|d| > 1.0$ and several reaching $|d| > 3.0$. Importantly, the change points are distributed across all three process variables and all three mills rather than concentrated at a single temporal event, which would be the expected signature of a sensor recalibration or unit conversion failure. The pattern is instead consistent with progressive operational adjustments, equipment maintenance, and grinding aid composition changes that accumulate over multi-year operation. Fig. 2 illustrates the temporal pattern for MILL_186_BF_PRESSURE.

B. K-Means Regimes on Raw Process Variables

The eight K-means clusters introduced in Section III-B (fitted on D_1 in the 40-dimensional space of 37 standardized PVs plus three mill one-hot indicators) serve a second role in this section as the raw-PV reference clustering against which the embedding-space clusters in Section VI will be compared. Adopting a single, fixed clustering for both purposes avoids the ambiguity of comparing two independently fitted K-means models and ensures that the comparison in Section VI-B isolates the effect of the SCARF representation rather than any residual algorithmic variability. The silhouette score for this $k = 8$ partition is 0.16, computed on the same evaluation subset used for k selection. We note that $k = 8$ is the upper boundary of the search grid; while extension of the search range may yield higher silhouette values, the relatively low absolute silhouette score (a value of this magnitude indicates substantial cluster overlap, as silhouette scores below approximately 0.25 are commonly interpreted as evidence of weak cluster structure) reflects the overlapping nature of operating modes in the raw PV space. Fig. 2 overlays the K-means regime assignments as background shading, where each color corresponds to one regime.

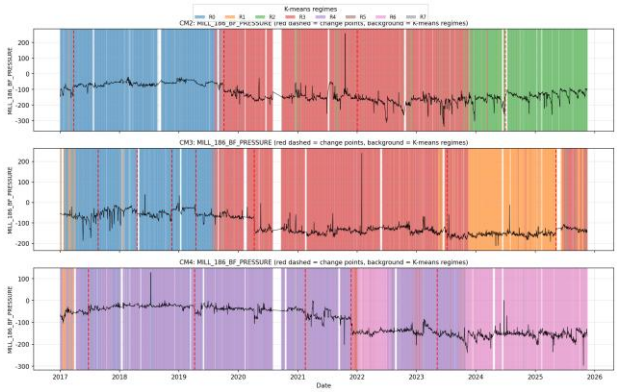

Fig. 2. MILL_186_BF_PRESSURE time series for CM2 (top), CM3 (middle), and CM4 (bottom). Red dashed lines: PELT change points. Background shading: K-means regimes on raw process variables.

Two observations from Fig. 2 are noteworthy. First, the K-means regime assignments are not temporally monotone: multiple regimes alternate across multiple time periods, consistent with operational cycling rather than a one-time global shift. Second, the K-means regime boundaries visually align with the PELT change-point locations, despite K-means having no temporal information. This convergence between the two independent methods (one purely temporal, one purely statistical in feature space) provides cross-method evidence that the discovered regimes correspond to genuine operational variability rather than algorithmic artifact.

V. SELF-SUPERVISED REPRESENTATION LEARNING

A. SCARF Architecture and Training

We adopt SCARF [3] for self-supervised representation learning on the combined dataset $D_1 \cup D\_unlab$ (162,731 records). The architecture comprises an encoder f mapping the 40-dimensional input (37 standardized PVs concatenated with three mill one-hot dimensions) through a three-layer multilayer perceptron (MLP) with a hidden size of 256 to a 128-dimensional embedding, followed by a projector g consisting of a two-layer MLP (128 to 128) used only during pretraining. After pretraining, the projector is discarded, and only the encoder is used to produce embeddings.

SCARF training optimizes the InfoNCE contrastive loss [10], in which each sample's positive view is generated through feature corruption (random replacement of 60 per cent of features with values sampled from the empirical marginal distribution of each feature) and the embedding is trained to maximize similarity between corrupted positive pairs versus dissimilarity to other samples in the batch. The InfoNCE loss for sample i in a batch of size N is defined as in Eq. (1):

$$L_{NCE}(i) = -\log \left[\exp(\mathrm{sim}(z_i, z_i^+)/\tau) / \Sigma_{j \in B(i)} \exp(\mathrm{sim}(z_i, z_j)/\tau) \right] \quad (1)$$

where, $z = g(f(x))$ is the projected embedding, $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity, $\tau = 0.5$ is the temperature parameter, and $B(i)$ denotes the set of all $2N - 1$ distinct in-batch views excluding

the anchor z_i itself (that is, the positive view z_i^+ and all other corrupted views from the same mini-batch of size N = 1,024). The implementation follows the standard SimCLR-style 2N-batch convention. Training uses AdamW optimizer with learning rate 1×10^{-3} and weight decay 1×10^{-5}, batch size 1,024, and maximum 100 epochs with early stopping (patience 15) on validation InfoNCE loss measured on a 10 percent random held-out split of $D_1 \cup D\_unlab$. The model is initialized from three random seeds (42, 123, 456), and the embedding with the lowest validation loss is selected for downstream analysis.

B. Pretraining Convergence

Multi-seed pretraining yields stable convergence behavior. The best validation InfoNCE losses across the three seeds are 6.92, 6.92, and 6.96, with a mean of 6.93 and a standard deviation of 0.02. For the 2N-batch convention used here with N = 1,024, the random-baseline expected loss under uniform negative sampling is $\log(2N - 1) = \log(2{,}047) \approx 7.62$ nats; the observed losses are below this baseline by approximately 0.7 nats. We acknowledge that the absolute InfoNCE value is sensitive to the choice of negative sampling convention, and the downstream evidence for meaningful representation learning is provided by the clustering quality and external validity analyses in Section VI rather than by the raw InfoNCE value alone. Fig. 3 shows training and validation loss curves across the three random seeds.

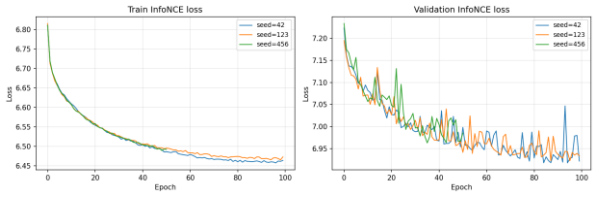

Fig. 3. SCARF training and validation InfoNCE loss curves across three random seeds.

VI. Embedding-Space Clustering

A. Clustering Algorithms and Selection

After pretraining, we apply L2 normalization to the 128-dimensional embeddings, projecting each sample onto the unit hypersphere. This is standard practice for contrastive embeddings, where cosine similarity (rather than Euclidean distance) was the training objective. Three clustering algorithms are then compared: K-means with silhouette-based k selection over {3, 5, 8, 10, 12, 15, 20}; Gaussian Mixture Model (GMM) with diagonal covariance and Bayesian Information Criterion (BIC) selection over {3, 5, 8, 10, 12, 15}; and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) with minimum cluster size selection over {200, 500, 1,000, 2,000} and minimum samples 50.

TABLE III. Embedding-Space Clustering Comparison

Algorithm	k	Silhouette	Davies–Bouldin	Calinski–Harabasz	ARI vs PV	Noise %
K-means	15	0.366	1.132	5,338	0.354	0.0%
GMM	15	0.363	1.072	5,245	0.378	0.0%
HDBSCAN	21	0.407	0.948	4,959	0.234	16.5%

[c.] Silhouette and Davies–Bouldin computed on a 30,000-sample subset for tractability. ARI is computed on full data.

Table III reports the clustering quality metrics. K-means selects k = 15 with silhouette 0.366, GMM selects k = 15 by BIC with silhouette 0.363, and HDBSCAN identifies 21 components at minimum cluster size 2,000 with silhouette 0.407 on non-noise samples (16.5 per cent of samples designated as noise). The Adjusted Rand Index (ARI) between each embedding-space clustering and the PV-space K-means clustering (Section IV) is 0.354 for K-means, 0.378 for GMM, and 0.234 for HDBSCAN. These ARI values indicate weak-to-moderate agreement, meaning that the embedding-space clusters preserve coarse PV-regime structure while introducing finer subdivisions. Fig. 4 visualizes the relative performance of the three algorithms across all four metrics.

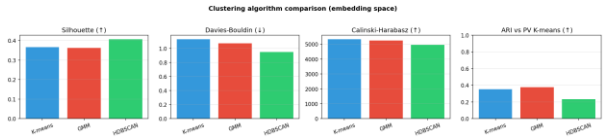

Fig. 4. Clustering algorithm comparison in the embedding space (silhouette, Davies–Bouldin, Calinski–Harabasz, and ARI versus PV K-means).

Three observations are worth emphasizing. First, the embedding-space silhouette score (0.37) substantially exceeds the raw PV-space K-means silhouette (0.16), indicating that the SCARF representation has produced a feature space where operating regimes are more cleanly separable; we acknowledge that the two silhouettes are computed in spaces of different dimensionality (128-dimensional L2-normalized embedding versus 40-dimensional standardized PV plus mill one-hot space) and that the contrastive objective directly optimizes for cluster-like compactness, so the quantitative ratio of the two silhouettes should be interpreted as a directional indicator rather than an equal-footing measurement. Second, K-means and GMM converge on the same number of components (k = 15) through independent selection criteria (silhouette and BIC, respectively). For K-means, k = 15 is an interior optimum of the wider search grid {3, 5, 8, 10, 12, 15, 20} (the silhouette score peaks at k = 15 and decreases at k = 20), which provides a stronger indication that k = 15 reflects an intrinsic count rather than a grid-boundary effect. For GMM, k = 15 coincides with the upper boundary of its search grid {3, 5, 8, 10, 12, 15}; BIC may favor an even larger number of components if the grid were extended, so the GMM result alone would be inconclusive. The agreement between the two algorithms — one finding an interior optimum and the other reaching its boundary at the same value — is consistent with k = 15 being an intrinsic count, though extending the GMM grid would strengthen the conclusion further. Third, HDBSCAN identifies 21 components with the highest silhouette, where the selected minimum cluster size of 2,000 was the largest value in its search range and yields 16.5 percent of samples designated as noise; for downstream pseudo-labeling applications this noise designation is operationally problematic. Based on these considerations and to ensure consistency with the visualizations in subsequent subsections, we adopt K-means with k = 15 as the primary representation for the remainder of this study.

To assess whether the contrastive representation is necessary, or whether a simpler and computationally cheaper pipeline would suffice, we additionally clustered the same 162,731-record input using linear dimensionality reduction. Table IV compares K-means (k = 15) applied to PCA-reduced

and standardized process-variable representations against the SCARF embedding. Two patterns emerge. First, the SCARF embedding produces markedly more separable clusters: its silhouette score (0.366) is roughly twice that of any PCA or standardized-PV baseline (0.16–0.24), indicating that linear projection preserves the diffuse, overlapping structure of the raw process-variable space rather than resolving it. Second, the PCA baselines exhibit higher ARI against the PV-space K-means partition (0.47–0.51 versus 0.354 for SCARF), which is expected: a linear projection largely reproduces the partition of the space it was derived from, whereas the SCARF embedding deliberately departs from it. This combination — substantially higher cluster separation together with lower agreement with the raw-PV partition — indicates that SCARF does not merely re-express the PV-space regimes more compactly but recovers a finer, better-separated regime structure that the simpler linear pipelines do not capture, justifying the additional cost of contrastive pretraining.

TABLE IV. SIMPLER DIMENSIONALITY-REDUCTION BASELINES VERSUS THE SCARF REPRESENTATION (K-MEANS, K = 15).

Representation	Dim.	Silhouette	ARI vs PV
Standardized process variables (no reduction)	40	0.161	0.474
PCA (95% variance retained)	32	0.174	0.513
PCA (compact)	10	0.236	0.482
SCARF embedding	**128**	**0.366**	**0.354**

[d] All methods cluster the same 162,731-record input (37 standardized process variables plus three mill one-hot indicators) with K-means (k = 15). Silhouette is computed on a fixed 20,000-record subsample; ARI is computed on all records against the k = 8 PV-space reference clustering of Section IV-B.

B. Relationship to PV-Space Regimes

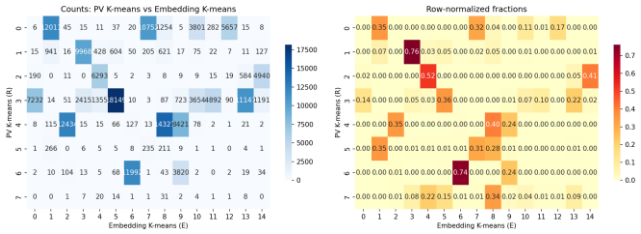

Fig. 5. Cross-tabulation between PV K-means regimes (rows) and embedding K-means clusters (columns). Left: counts; right: row-normalized fractions.

Fig. 5 shows the cross-tabulation between PV-space K-means regimes (R0–R7) and embedding-space K-means clusters (E0–E14), both as raw counts and row-normalized fractions. Two structural patterns are observed. First, certain PV regimes map cleanly to single embedding clusters. PV regime R1 has 76 percent of its samples in embedding cluster E3; PV regime R6 has 74 per cent in E6. These cases correspond to operating regimes that the raw PV space already separates well, and SCARF reinforces this separation. Second, other PV regimes split across multiple embedding clusters. PV regime R0, the largest in the raw PV space, distributes across four embedding clusters (E1, E7, E10, E12) with the largest fraction at 35 percent. PV regime R5 splits across three (E1, E7, E8). This finer subdivision indicates that within what raw PV clustering treats as a single coarse regime, the SCARF embedding identifies sub-regimes that share gross PV characteristics but differ in finer operational structure that the contrastive objective deemed discriminative.

C. Mill Specificity of Embedding Clusters

Fig. 6 displays the mill distribution within each embedding K-means cluster. The pattern is striking: 14 of the 15 clusters are dominated by a single mill (CM2 in clusters 0, 7, 11, 12, 13, 14; CM3 in clusters 1, 3, 5, 10; CM4 in clusters 2, 6, 8, 9). Cluster 4 is the only cluster with appreciable mill mixing, containing approximately 75 percent CM2 and 25 percent CM3 records. Cluster 8 is CM4-dominated (approximately 99 per cent CM4) with a small marginal CM3 component.

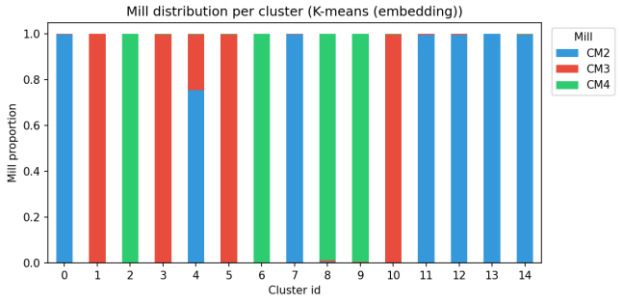

Fig. 6. Mill composition of each embedding K-means cluster. Nearly all clusters are dominated by a single mill, indicating strong mill-specific operating patterns captured by SCARF.

This pattern indicates that SCARF, despite receiving mill identifier only as a 3-dimensional one-hot feature alongside 37 process variables, has learned an embedding in which mill-specific operating patterns dominate. The contrastive objective evidently treats inter-mill differences as larger discriminative signals than within-mill regime differences. Cluster 4, the single mixed cluster, is operationally interpretable: it represents a region of the embedding space where CM2 and CM3 operating modes converge, plausibly indicating a shared operational state achievable on both mills (for example, similar feed composition and similar grinding fineness target).

D. Temporal Evolution of Embedding Clusters

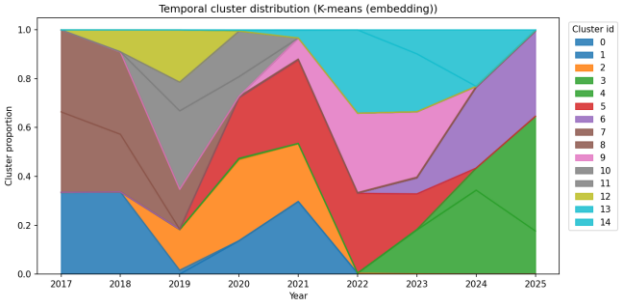

Fig. 7. Year-wise stacked area plot of embedding K-means cluster proportions, showing the progressive evolution of operating modes.

Fig. 7 displays a stacked area plot of cluster proportions by year. The visualization reveals a progressive temporal evolution of operational modes. In 2017 and 2018, clusters 7 and 8 dominate, together accounting for approximately 80 per cent of samples. From 2019 onward, additional clusters (notably clusters 11 and 12) gain prominence and the dominance of

cluster 7 begins to decline. From 2020 to 2021, mid-period clusters (clusters 2 and 5) become prominent. From 2022 onward, a further wave of clusters appears, including clusters 6, 9, 13, and 14. By 2024 and 2025, late-period clusters such as cluster 4 and cluster 6 dominate. The specific color–cluster mapping is shown in the figure legend.

The temporal pattern in Fig. 7 corresponds with the change points reported in Section IV. The first major transition, around 2019, coincides with PELT change points in multiple variables across all three mills. The second major transition, around 2022, coincides with PELT change points in CM2 and CM3. The most recent transition, around 2024, coincides with the latest PELT change point. The combined evidence from PELT (which uses purely temporal information) and SCARF clustering (which uses no temporal information except indirectly through the data distribution) thus converges on the same coarse drift structure, providing independent cross-method validation that the discovered structure reflects genuine operational evolution. We note that the alignment is presented qualitatively (visual coincidence in time) rather than through a formal alignment statistic; a quantitative alignment indicator is discussed as future work in Section VII F.

To quantitatively confirm that the embedding-derived regimes reflect genuine operational differences rather than artifacts of the contrastive learning process, we tested whether the 15 embedding regimes are statistically separable in the original process-variable space. A Kruskal–Wallis test across the regimes was significant for all 37 process variables ($p < 0.001$ in every case), with a mean effect size of $\eta^2 = 0.43$ and values reaching $\eta^2 = 0.79$ for bag-filter differential pressure (MILL_186_BF_DP) and 0.78 for bag-filter pressure (MILL_186_BF_PRESSURE). Because these regimes were discovered purely in the learned embedding space — without access to the raw process-variable distributions during clustering — their strong and uniform separability in process-variable space provides quantitative evidence that the discovered structure corresponds to physically distinct operating modes rather than to learning artifacts.

VII. Discussion

A. Implications for Pre-2023 Data Treatment

The analysis above directly addresses the central methodological question raised in the introduction. The initial restriction to the 2023 to 2025 period was motivated by observations of substantial distributional deviation in pre-2023 Blaine measurements relative to post-2023 records. The present analysis shows that, under a wider valid-range filter applied uniformly across the full 2017 to 2025 period, a meaningful fraction of pre-2023 records does pass both Blaine and 44 μmR range conditions and can be admitted to the labeled set. More importantly, the records that lack valid quality labels (and therefore enter D_unlab) still contain coherent operational structure: PELT detects multiple change points within the pre-2023 period itself (for example, CM3 MILL_SEPOL_SEPOL has six change points distributed across 2017 to 2023), and the SCARF embedding identifies multiple distinct operational clusters within this period (clusters 7, 8, 11, 12 dominate 2017 to 2020). The pre-2023 records thus contribute distinct operational regimes that, if discarded entirely, would represent a substantial loss of information for downstream tasks that benefit from operational diversity, even when those records cannot directly serve as supervised labels.

B. Validity of Embedding-Derived Regimes

The validity of the 15-cluster embedding-derived regime structure is supported by three lines of evidence, each with its own caveat. First, two clustering algorithms with different selection criteria (K-means with silhouette, GMM with BIC) converge on the same number of components ($k = 15$). For K-means, $k = 15$ is an interior optimum of the search grid {3, 5, 8, 10, 12, 15, 20}, which provides direct evidence for $k = 15$. For GMM, $k = 15$ coincides with the upper boundary of its search grid; the agreement is consistent with an intrinsic count but extending the GMM grid would further strengthen this conclusion. Second, the embedding-space silhouette score (0.37) substantially exceeds the raw PV-space score (0.16), although as noted in Section VI A, this comparison is across spaces of different dimensionality and the embedding directly optimizes for contrastive compactness. Third, the temporal evolution of embedding clusters aligns with the PELT change-point locations obtained from a completely independent algorithm using only daily-aggregated time series of three variables. Of the three lines, the third provides the strongest independent corroboration because PELT and SCARF use disjoint information (PELT uses time only on three signals, SCARF uses no time information on 37 signals). The convergence of these assessments provides reasonable, though not definitive, confidence that the discovered regime structure is operationally meaningful.

C. Mill Specificity and Cross-Mill Generalization

The strong mill specificity in Fig. 6 has implications for downstream prediction tasks. If 14 of 15 operating regimes are mill-specific, then models that share parameters across mills must implicitly learn mill-conditional behavior. Cluster 4, the one mixed cluster (CM2/CM3 = 75:25), suggests that some operational states are reachable across mills and that cross-mill transfer is plausible in those regions of the operating space. Future work should investigate whether cross-mill pseudo-labeling within shared operational clusters (such as cluster 4) yields better generalization than facility-wide pseudo-labeling that ignores mill structure.

D. Downstream Quality Prediction Benefit

While the central analysis of this study concerns operational regime structure, we additionally verify that the learned representation provides measurable benefits for the intended downstream task. Using the same labeled set D_l and the SCARF embeddings characterized above, we trained two state-of-the-art deep tabular regressors (FT-Transformer and SAINT) to predict Blaine fineness and 44 μmR, comparing models that receive the raw 40-dimensional input against models that additionally exploit the SCARF representation together with cluster-conditional pseudo-labeling of the unlabeled records. As summarized in Table V, incorporating the learned representation improves test R^2 for every target–model pair, raising Blaine R^2 from 0.524 to 0.567 (FT-Transformer) and from 0.483 to 0.557 (SAINT), and 44 μmR R^2 from 0.836 to 0.852 and from 0.825 to 0.844, respectively. The corresponding reductions in prediction error range from 9.1% to 14.3% and are statistically

significant for all four pairs under both the Wilcoxon signed-rank test and a paired t-test ($p < 0.01$; $n = 7{,}691$). The larger relative gain on Blaine, the harder chemically-mediated target, indicates that the representation is most valuable precisely where raw process variables are least informative. These results confirm that the regime structure recovered through self-supervised pretraining is not merely descriptive but translates into improved predictive performance. A complete end-to-end evaluation of the prediction system — including temporal (chronological) generalization and a systematic study of the pseudo-labeling strategy — is beyond the scope of the present characterization study and is pursued in a separate work.

TABLE V. DOWNSTREAM QUALITY PREDICTION WITH AND WITHOUT SCARF REPRESENTATION

Target	Model	Test R^2 (raw PV)	Test R^2 (+SCARF)	ΔR^2	RMSE (raw→+SCARF)	Wilcoxon p
Blaine	FT-Transformer	0.524	0.567	+0.043	100.7 → 96.0	$<10^{-63}$
Blaine	SAINT	0.483	0.557	+0.074	105.0 → 97.2	$<10^{-89}$
44 µmR	FT-Transformer	0.836	0.852	+0.015	0.521 → 0.496	$<10^{-66}$
44 µmR	SAINT	0.825	0.844	+0.019	0.538 → 0.508	$<10^{-44}$

[e.] Three-seed means: "raw PV" = 40-dimensional input used directly, "+SCARF" = same input with the SCARF representation and cluster-conditional pseudo-labeling.

E. Limitations

Seven limitations should be noted. First, the analysis is conducted on a single facility with three mills; multi-facility validation is essential for assessing generalizability across cement plants with different equipment vendors and operational practices. Within this single facility, however, the three mills (CM2, CM3, and CM4) differ substantially in their operating behavior — as evidenced by the strong mill specificity reported in Section VI C, where 14 of 15 embedding clusters are dominated by a single mill — so the present study already evaluates the method across three distinct operating environments rather than a single homogeneous process. The contribution of this work is methodological: to demonstrate that a coherent operational regime structure can be recovered from multi-year data of a given facility, rather than to claim cross-plant universality of the specific regimes discovered. Validation on independent plants with different equipment vendors remains essential future work to establish broader generalizability. Second, our PELT analysis is restricted to three process variables; a more comprehensive multivariate change-point analysis covering all 37 variables jointly may reveal additional regime transitions not visible in any single variable. Third, the choice of the L2 cost model in PELT assumes mean-shift transitions; transitions characterized by variance changes or covariance shifts may be missed. Fourth, the SCARF contrastive objective treats all feature corruptions as equally informative; feature-specific corruption rates (for example, higher corruption for derived variables and lower for raw sensor readings) may produce more informative embeddings. More broadly, the SCARF hyperparameters used here — the 128-dimensional embedding, the 60 per cent corruption rate, the three-layer encoder, and the contrastive temperature — follow the standard configuration recommended for SCARF rather than values tuned on this dataset; a systematic ablation over these design choices is a natural next step but lies outside the scope of the present characterization study, whose aim is to establish whether a coherent operational regime structure can be recovered rather than to optimize representation-learning hyperparameters. Relatedly, this study evaluates a single self-supervised method (SCARF), which was chosen as a representative contrastive approach well suited to the heterogeneous, mixed-scale process variables of finish milling; a controlled benchmark of SCARF against other self-supervised tabular paradigms — such as the reconstruction-based VIME [9] and SubTab [11], or the attention-based TabTransformer [12] — on the same industrial dataset would help establish which self-supervised inductive bias best captures operational-regime structure and is left to future work. Fifth, while Section VII-D confirms that the learned representation improves downstream quality prediction, a complete end-to-end predictive evaluation — including strictly chronological temporal generalization and a systematic analysis of the pseudo-labeling strategy — is left to separate work. Sixth, the SSL pretraining set described in Section V combines D_1 (including its test split) with D_unlab, which is acceptable for the unsupervised analysis presented in this study but would create a label-free form of test exposure if the same embeddings were reused without modification for downstream supervised evaluation; subsequent supervised studies will retrain the SSL encoder with the supervised test split excluded. Seventh, the regime-stratified random split of D_1 described in Section III-B places temporally adjacent records into different subsets, so any held-out test split constructed from D_1 provides an optimistic estimate of temporal generalization relative to a strictly chronological split; this protocol is adopted in the present study to ensure regime-balanced evaluation across the full 2017 to 2025 data span, and a strictly chronological split is recommended for any downstream supervised study that aims to assess temporal generalization on its own merits.

F. Toward Quantitative Alignment Indicators

The PELT–SCARF alignment reported in Section VI D is presented qualitatively. A quantitative indicator could be constructed by computing, for each PELT change-point date, the local Jensen–Shannon divergence between cluster assignment distributions in fixed-width windows before and after the change point, or by computing the ARI between cluster assignments in different temporal segments. Development of such indicators is left as future work.

VIII. CONCLUSION

This study has characterized the operational drift structure in approximately nine years of cement finish milling DCS data (188,858 hourly records across three industrial mill units) and demonstrated that self-supervised pretraining can recover this structure for downstream applications. PELT change-point detection identifies between 12 and 21 detected drift events per mill, with effect sizes (Cohen's d) consistently satisfying $|d| > 1.0$ for the majority of transitions. SCARF contrastive learning

on the combined labeled and unlabeled set (162,731 records) yields stable 128-dimensional embeddings (validation InfoNCE 6.93 ± 0.02 across three random seeds). Three independent clustering algorithms on the embedding space (K-means, GMM, HDBSCAN) consistently identify 15 to 21 operational regimes with silhouette scores between 0.36 and 0.41, substantially exceeding the raw PV-space silhouette of 0.16 (with the caveats on across-space comparison noted in Section VI-A). The embedding-space K-means regimes exhibit strong mill specificity (14 of 15 clusters dominated by a single mill) and show temporal evolution that aligns with the independently detected PELT change-point boundaries. These findings establish that long-term cement process data contains a richer operational regime structure than implied by raw PV clustering, and that contrastive self-supervised pretraining recovers this structure efficiently from unlabeled records, providing a representation suitable for use in downstream quality prediction tasks (downstream supervised performance is the subject of separate ongoing work).

Two directions for future work follow directly from these results. First, the cluster-conditional pseudo-labeling of the unlabeled subset, using the embedding-derived regimes as conditioning variables, is currently being investigated as a means of enabling state-of-the-art deep tabular models (FT-Transformer, SAINT) to close the performance gap with tree-based ensembles documented in the tabular learning literature; full empirical results will be reported in a forthcoming study. Second, the multi-facility generalization of the discovered regime structure remains an open question; comparison of embeddings across cement facilities with different equipment vendors would clarify whether the SCARF representation captures facility-specific or process-universal regime information.

REFERENCES

[1] A. Hasanbeigi, L. Price, and E. Lin, "Emerging energy-efficiency and CO_2 emission-reduction technologies for cement and concrete production: a technical review," Renewable and Sustainable Energy Reviews, vol. 16, no. 8, pp. 6220–6238, 2012.

[2] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," Journal of the American Statistical Association, vol. 107, no. 500, pp. 1590–1598, 2012.

[3] D. Bahri, H. Jiang, Y. Tay, and D. Metzler, "SCARF: self-supervised contrastive learning using random feature corruption," in Proc. International Conference on Learning Representations (ICLR), 2022.

[4] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: the role of machine learning," IEEE Access, vol. 5, pp. 20590–20616, 2017.

[5] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3235–3243, 2018.

[6] A. K. Pani and H. K. Mohanta, "Soft sensing of particle size in a grinding process: application of support vector regression, fuzzy inference, and adaptive neuro-fuzzy inference techniques for online monitoring of cement fineness," Powder Technology, vol. 264, pp. 484–497, 2014.

[7] X. Yuan, Y. Wang, C. Yang, Z. Ge, Z. Song, and W. Gui, "Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes," IEEE Transactions on Industrial Electronics, vol. 65, no. 2, pp. 1508–1517, 2018.

[8] H. Zermane and A. Drardja, "Development of an efficient cement production monitoring system based on the improved random forest algorithm," International Journal of Advanced Manufacturing Technology, vol. 120, pp. 1853–1866, 2022.

[9] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "VIME: extending the success of self- and semi-supervised learning to tabular domain," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.

[11] T. Ucar, E. Hajiramezanali, and L. Edwards, "SubTab: subsetting features of tabular data for self-supervised representation learning," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021.

[12] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678, 2020.

[13] I. Rubachev, A. Alekberov, Y. Gorishniy, and A. Babenko, "Revisiting pretraining objectives for tabular deep learning," arXiv preprint arXiv:2207.03208, 2022.

[14] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," Signal Processing, vol. 167, p. 107299, 2020.

[15] D. C. Montgomery, Introduction to Statistical Quality Control, 6th ed. Hoboken, NJ: Wiley, 2009.

[16] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: a review," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2346–2363, 2019.