Novel Segmentation Methodology for Robust Feature Engineering of Time Series Data in Prognostics and Health Management

Dai-Yan Ji[1], Jay Lee[2]

[1,2]*Center for Industrial Artificial Intelligence, Department of Mechanical Engineering, University of Maryland, College Park, MD, 20742, USA*
jidn@umd.edu
leejay@umd.edu

ABSTRACT

Time series segmentation plays a critical role in feature engineering for prognostics and health management (PHM), yet most existing approaches rely on domain-specific rules or fail to preserve meaningful transient patterns. This research proposes a segmentation-driven framework that leverages a greedy Perceptually Important Point (PIP) algorithm to identify informative structural regimes in sensor signals without prior domain knowledge. A global reference signal is constructed from class-level Euclidean-barycenter averages, and consistent segment boundaries are applied across all samples. Segment-level statistical features are then extracted and used for classification. Evaluation on a chemical gas sensor dataset demonstrates that the proposed method significantly outperforms traditional whole-signal summary statistics, achieving improved robustness to drift and unit variability. Future work includes parameter optimization of the PIP algorithm, exploration of class-sensitive segmentation strategies, and extension of the framework to remaining useful life (RUL) prediction and anomaly detection tasks.

1. PROBLEM STATEMENT

Time series data have long played a critical role in PHM, serving as the foundational information source for health assessment, fault diagnosis, and remaining useful life prediction. Figure 1 illustrates the end-to-end modeling process—from time series data collection and segmentation strategies to feature engineering and prognostic model outputs. Traditionally, the extraction of features from time series data relies heavily on summary statistics, such as mean, standard deviation, kurtosis, and skewness. While these statistical measures are straightforward, computationally efficient, and broadly applied, they frequently fail to capture crucial transient and dynamic behaviors inherent in complex systems. As a result, traditional statistical approaches may overlook valuable diagnostic information, significantly limiting the accuracy and robustness of subsequent health monitoring models. Alternative methods rely heavily on domain knowledge, segmenting signals based on predefined operational steps (e.g., recipe steps in semiconductor manufacturing). However, this reliance limits scalability and applicability to new or unknown operational contexts. Therefore, there is a critical need for a robust, automated segmentation method capable of effectively identifying meaningful signal segments without prior domain-specific knowledge, thus significantly enhancing feature extraction quality and prognostic accuracy.

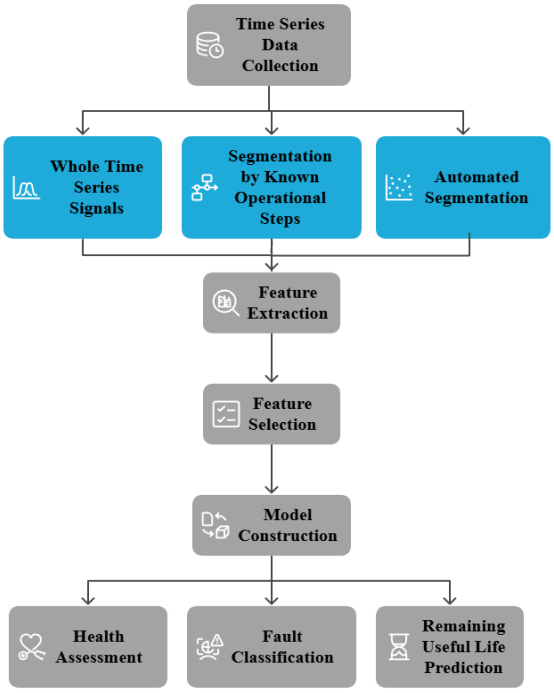

Figure 1. Comprehensive Workflow of Time Series Modeling for PHM Applications.

Recent literature highlights a variety of segmentation methods employed in time series-based PHM frameworks. Common approaches include:

- **Piecewise Approximation Methods**: This category includes techniques (Wilson, 2017) such as Piecewise Aggregate Approximation (PAA), Piecewise Linear Approximation (PLA), Piecewise Constant Approximation (PCA), and Symbolic Aggregate Approximation (SAX), which divide signals into equal-length or fixed-structure segments. These methods are computationally efficient and suitable for large-scale or streaming applications. However, they typically ignore the underlying signal dynamics, often producing segments that span multiple operational states, which limits interpretability and reduces the quality of extracted features.
- **Event- and threshold-based segmentation**: This includes techniques that segment signals based on predefined domain rules or threshold crossings (Yang, 2016), such as changes in valve positions or setpoint levels. It allows alignment with meaningful process events but depends heavily on domain knowledge and is not generalizable across systems.
- **Statistical changepoint detection**: These methods, such as Binary Segmentation and PELT(Killick et al., 2012), detect abrupt changes in distribution (e.g., mean, variance) and are useful for capturing transitions in operational modes. However, they struggle with noisy or gradual transitions and require tuning.
- **Pattern-based or structure-aware segmentation**: This category includes methods such as spike detection using wavelets (Quiroga et al., 2004), ramp and oscillation pattern fitting (Olszewski, 2001; Srinivasan & Rengaswamy, 2012), or shapelet-based matching. These techniques capture functional primitives and yield segments that are semantically meaningful for diagnostics but often require careful parameterization and are computationally more demanding.

Despite the diversity of existing segmentation methods—including piecewise approximation, event/threshold-based, statistical changepoint, and pattern-based techniques—they all share critical limitations. Many require manual rule-setting, are sensitive to noise or hyperparameters, or are computationally intensive. Current approaches fall short when faced with complex, variable-length behaviors typical of real-world PHM data. These deficiencies underscore the fundamental importance of accurate segmentation. When time series data are treated uniformly—ignoring inherent structures such as ramps, spikes, or oscillations—critical degradation cues may be diluted or lost entirely. Poor segmentation directly weakens the relevance and separability of extracted features, leading to compromised model accuracy and reduced interpretability. In contrast, precise and meaningful segmentation can expose latent degradation mechanisms, enhance signal clarity, and significantly improve downstream fault diagnosis and RUL prediction. Therefore, there is an urgent need for a segmentation-driven feature engineering methodology that is both automated and robust, capable of adaptively identifying and isolating dynamic regimes without strong assumptions or prior domain knowledge. Such an approach offers strong potential to improve the diagnostic quality and reliability of PHM systems across a wide range of industrial applications.

2. EXPECTED NOVEL CONTRIBUTIONS

This research proposes a segmentation-driven framework that directly addresses the limitations identified in existing segmentation methods, such as rigidity, noise sensitivity, and reliance on expert knowledge. The expected contributions are as follows:

1. Design and implementation of an automated, data-driven segmentation methodology capable of adaptively identifying meaningful structural patterns—such as ramps, spikes, and oscillations—without requiring prior knowledge of operational steps or predefined rules.
2. Advancement of time series feature engineering in PHM by isolating homogeneous signal regimes, enabling more precise and context-aware extraction of diagnostic and prognostic features.
3. Development of a generalizable and scalable framework applicable across diverse industrial datasets, overcoming the specificity constraints of traditional rule-based or fixed-window methods.
4. Empirical validation of the framework using industrial sensor data (e.g., gas sensor calibration), demonstrating improved signal interpretability, enhanced feature relevance, and increased classification and prediction accuracy relative to conventional techniques.
5. Contribution to PHM system interpretability by producing segment-aware insights that can inform domain experts and operators, thus supporting more effective maintenance strategies and improving system reliability.

3. PROPOSED RESEARCH PLAN

The proposed research plan includes several structured phases:

1. Development and validation of a segmentation-driven preprocessing methodology capable of accurately identifying and isolating key patterns within complex time series signals.
2. Implementation of robust and adaptive feature extraction techniques tailored to each identified segment, ensuring high-quality, pattern-specific diagnostics and prognostics.

3. Evaluation of the developed methodology using gas sensor calibration data as a practical case study, comparing the performance of the proposed approach with conventional methods through metrics such as accuracy, sensitivity, and specificity.
4. Integration of the segmentation-driven framework into broader PHM systems, assessing its scalability and applicability across various industrial scenarios, thus ensuring comprehensive validation and broad relevance.

3.1. Work Performed and Preliminary Results

To evaluate the effectiveness of the proposed segmentation-driven framework, we apply it to the problem of chemical gas classification using uncalibrated sensor data, a challenging task that involves significant sensor variability, drift, and multi-regime operating conditions. The dataset is sourced from the University of California, Irvine Machine Learning Repository (Fonollosa et al., 2016). The dataset includes signals from five sensor units exposed to four gas types at ten concentration levels. Each test lasts 600 seconds and is recorded at 100 Hz. Raw signals are first interpolated to a fixed time base of 60,000 points and then downsampled to 0.5 Hz using linear interpolation followed by downsampling. All sensor channels were individually normalized using z-score normalization to ensure comparability across different units.

We use data from Unit 2, based on four testing days, each covering 4 gases × 5 selected concentration levels, totaling 80 training samples. The remaining 560 samples—including the other 5 concentration levels from Unit 2 as well as all samples from Units 1, 3, 4, and 5—are reserved for testing. This configuration ensures both manageable training complexity and comprehensive evaluation of model generalization under unit variation, concentration shift, and temporal drift. We compare two approaches: (1) a traditional whole-signal summary statistics method, and (2) the proposed segmentation-based feature extraction method. In the summary statistics approach, statistical features such as mean, max, min, standard deviation, skewness, and kurtosis are extracted from each sensor signal over the full duration. In contrast, our segmentation-based approach first groups sensor signals by gas type and identifies the most representative response channel using peak-to-peak amplitude. Figure 2 visualizes class-level representative signals computed using Euclidean barycenter for each gas type. These class-wise signals are then averaged to construct a global reference signal, which serves as the foundation for universal segmentation.

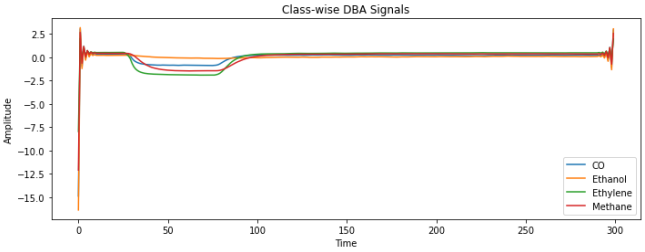

Figure 2. Class-Level Representative Signals Computed Using Euclidean barycenter.

A greedy Perceptually Important Point (PIP) segmentation algorithm, shown in Table 1, is applied to the global signal to determine key turning points, which serve as cut points. These fixed segment boundaries are then used to consistently segment all other signals. Within each segment and channel, we extract statistical descriptors such as mean, standard deviation, max, min, skewness, and kurtosis. This structured segmentation, shown in Figures 3 and 4, enables consistent and interpretable feature extraction across varying sensor dynamics. Classification is performed using a support vector machine (SVM) with an RBF kernel. The segmentation-driven approach achieves a significantly higher classification accuracy of 96.79%, as illustrated by the confusion matrix in Figure 5, outperforming the 84.64% achieved by the summary statistics baseline. The segmentation method also demonstrates greater consistency across units and days, highlighting its robustness to sensor drift and operational variability. This case study supports the effectiveness of structure-aware segmentation in enhancing feature relevance and improving PHM classification performance.

Table 1. Greedy PIP Selection for Time Serie Segmentation.

```
Algorithm 1: GreedyPIPSegmentation(signal, k, min_gap)

Input:
Signal // A 1-D array of length N representing the input time series
k   // Desired number of PIP points to select (including start and end)
min_gap // Minimum index distance allowed between any two selected PIP
points
Output:
    pip_indices // A sorted list of k perceptually important point (PIP)
indices

1:  pip_indices ← {0, N − 1}     ▷ Initialize with first and last point
2:  while |pip_indices| < k do
3:      max_dist ← −∞
4:      max_idx  ← −1
5:      for i ← 0 to N − 1 do
6:          if i ∈ pip_indices then
7:              continue
8:          if ∃ j ∈ pip_indices such that |i − j| < min_gap then
9:              continue
10:         left  ← max({j ∈ pip_indices | j < i}, default = 0)
11:         right ← min({j ∈ pip_indices | j > i}, default = N − 1)
12:         if left = right then
13:             continue
14:         x0 ← left;  y0 ← signal[left]
15:         x1 ← right; y1 ← signal[right]
16:         x  ← i;     y  ← signal[i]
17:         y_proj ← y0 + (y1 − y0) × (x − x0) / (x1 − x0)
18:         dist ← |y − y_proj|
19:         if dist > max_dist then
20:             max_dist ← dist
21:             max_idx  ← i
22:     if max_idx ≠ −1 then
23:         pip_indices ← pip_indices ∪ {max_idx}
24: return sort(pip_indices)
```

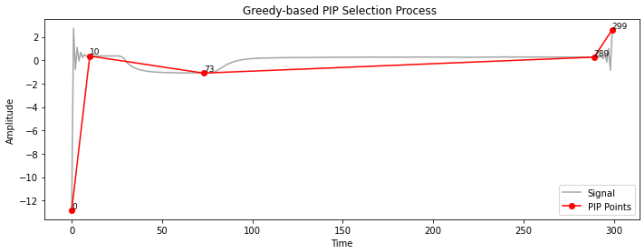

Figure 3. Greedy-based PIP selection process.

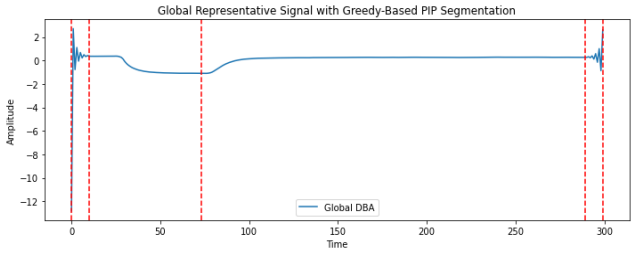

Figure 4. Representative points.

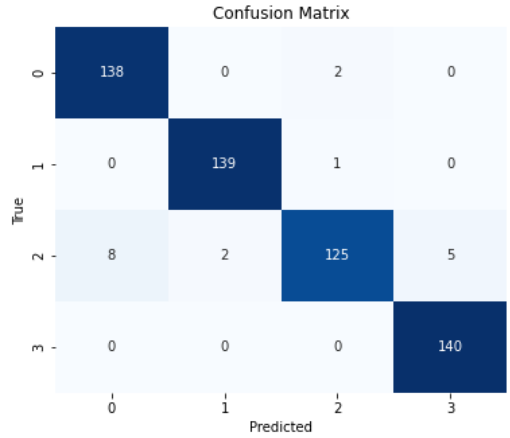

Figure 5. Confusion matrix for the proposed segmentation-based feature extraction method.

3.2. Remaining Work

Several important tasks remain to be addressed in the next step.

- First, although the current segmentation strategy based on a greedy Perceptually Important Point (PIP) algorithm has demonstrated strong performance, its effectiveness is sensitive to the choice of parameters—specifically, the number of PIP points (k) and the minimum gap between selected points (min_gap). Future work will explore systematic methods for parameter optimization, such as cross-validation, heuristic search, or Bayesian optimization, to adaptively tune these hyperparameters for different signal types or domains.
- Second, the current segmentation relies on a single global reference signal derived from class-wise Euclidean barycenter. While this promotes consistency, it may obscure class-specific segment boundaries that are informative for classification. An extension of the framework could incorporate hybrid segmentation strategies that combine global consistency with class-level sensitivity.
- Third, additional robustness testing will be conducted using other public or industrial time series datasets with different signal characteristics, drift behaviors, and fault types. This will help assess the generalizability and scalability of the proposed approach.
- Finally, the integration of the segmentation-driven framework into broader PHM pipelines—including data fusion, prognostics, and uncertainty quantification—remains an open area for development. By extending the framework beyond fault classification to tasks such as RUL prediction and anomaly detection, the full potential of structured segmentation in industrial AI systems can be realized.

REFERENCES

Fonollosa, J., Fernandez, L., Gutiérrez-Gálvez, A., Huerta, R., & Marco, S. (2016). Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization. *Sensors and Actuators B: Chemical*, *236*, 1044–1053.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598.

Olszewski, R. T. (2001). *Generalized feature extraction for structural pattern recognition in time-series data*. Carnegie Mellon University.

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*(8), 1661–1687.

Srinivasan, B., & Rengaswamy, R. (2012). Automatic oscillation detection and characterization in closed-loop systems. *Control Engineering Practice*, *20*(8), 733–746.

Wilson, S. J. (2017). Data representation for time series data mining: Time domain approaches. *Wiley Interdisciplinary Reviews: Computational Statistics*, *9*(1), e1392.

Yang, S. (2016). *An adaptive prognostic methodology and system framework for engineering systems under dynamic working regimes* [PhD Thesis]. University of Cincinnati.