BM20A6100 Advanced Data Analysis and Machine Learning

Comparative Analysis of PCA-MSPC and Kernel PCA-MSPC for Early Fault Detection in Wind Turbines

Fasie Haider – Haider Ali – Arman Golbidi

Wind Turbines Failure – Group A4

This report is written in LaTeX

2025–2026

No Artificial Intelligence tools were used in the generation of the code or the written content of this report.

Contents

1 Introduction

Wind turbine condition monitoring has become increasingly critical as the renewable energy sector expands. Early fault detection minimizes downtime and reduces maintenance costs by enabling predictive interventions before catastrophic failures occur. Multivariate Statistical Process Control (MSPC) offers a data-driven framework for detecting anomalies in high-dimensional sensor networks without requiring detailed physical models.

This study compares two MSPC approaches: linear Principal Component Analysis (PCA) and kernel PCA (k-PCA) for fault detection in wind turbines. While PCA effectively captures linear correlations among process variables, many industrial faults manifest as nonlinear deviations that traditional PCA may miss. Kernel PCA extends the linear framework by implicitly mapping data into a high-dimensional feature space via the kernel trick, potentially improving sensitivity to complex fault patterns [3].

We train both models exclusively on healthy operational data from wind turbine WT2, then evaluate their performance on two faulty units (WT14, WT39). The objectives are: (i) quantify detection rates and false alarm rates, (ii) assess time-to-detection for early warning capability, and (iii) provide interpretable diagnostics through variable contribution analysis. This work aligns with recent studies demonstrating that PCA-based monitoring can achieve high sensitivity when properly calibrated on baseline healthy operation [6, 7].

2 Materials and Methods

2.1 Data Description

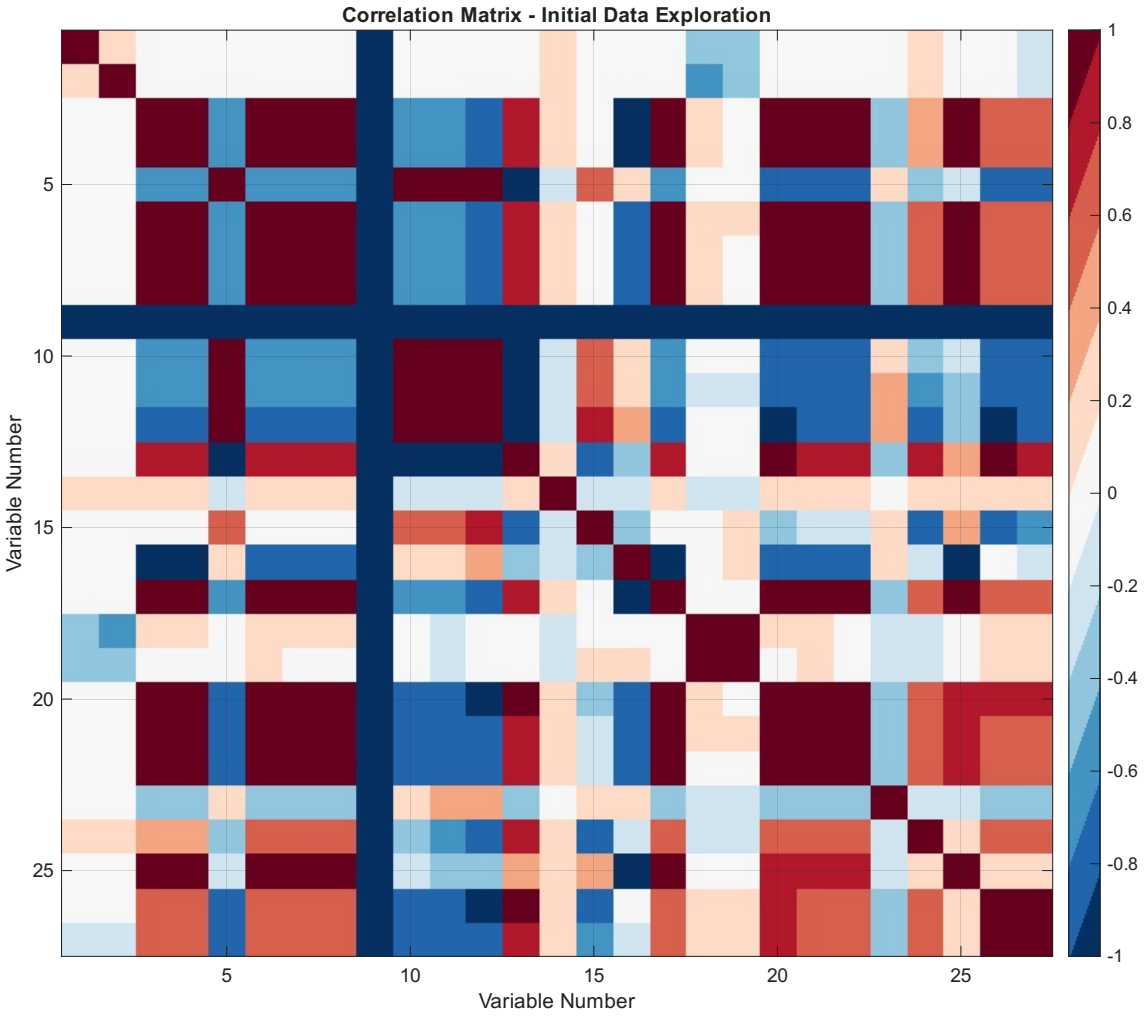
The dataset comprises operational time-series from three wind turbines sampled at 10-second intervals: one healthy unit (WT2, $n = 1570$ samples after preprocessing) and two faulty units (WT14 and WT39). Each observation includes 27 process variables covering mechanical, electrical, and environmental sensors.

Week-1 details carried over. Raw spreadsheet had four sheets: WT2 (1570×28), WT3 (698×31), WT14 (686×27), WT39 (1405×27). Challenges included: (i) one missing value in WT14; (ii) inconsistent number of columns (WT3 and an extra column in WT2); (iii) unknown/heterogeneous units and scales; (iv) outliers and abrupt distributional shifts in some sensors. Decisions: remove WT3 due to incompatible structure; drop WT2's extra column so that the remaining turbines align to $p = 27$ variables before further screening.

Preprocessing pipeline:

1) Removal of accidental numeric header rows detected in raw Excel sheets.
2) Time-aware gap handling: short gaps (≤ 3 consecutive missing samples) are filled via linear interpolation to preserve temporal continuity; longer gaps trigger selection of the longest contiguous NaN-free window per turbine.
3) Healthy-based autoscaling: all data are standardized using the mean vector μ_h and standard deviation vector σ_h computed exclusively from WT2 (healthy). Variables exhibiting zero variance in the healthy training set are discarded (Var12, Var15), yielding 25 retained sensors.

This healthy-only scaling strategy prevents fault signatures from contaminating normalization parameters, which is essential for unbiased anomaly detection [4].



Correlation matrix heatmap (healthy WT2).

Figure 1: Correlation structure of the sensor variables in the healthy baseline (WT2). Shown before zero-variance removal ($p = 27$). Strong positive/negative correlations motivate PCA-based dimensionality reduction prior to monitoring.

2.2 Mathematical Methods

2.2.1 Principal Component Analysis (PCA-MSPC)

After standardization, let $\mathbf{Z} \in \mathbb{R}^{n \times p}$ denote the centered data matrix with $p = 25$ variables. PCA decomposes the healthy training covariance matrix $\mathbf{C}_h = \frac{1}{n_h - 1} \mathbf{Z}_h^\top \mathbf{Z}_h$ into eigenvalues and eigenvectors.

After cleaning and time–aware pretreatment, we autoscale using healthy statistics and drop zero–variance–in–healthy sensors:

$$\mathbf{Z} = \frac{\mathbf{X}_{\text{kept}} - \mu_h}{\sigma_h}. \tag{2.1}$$

Then fit PCA only on the healthy block $\mathbf{Z}_{1:n_h,:}$. Let $\mathbf{P} \in \mathbb{R}^{p \times a}$ be the loading matrix and

$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_a)$ the retained eigenvalues. Scores and reconstruction for any row are

$$\mathbf{T} = \mathbf{ZP}, \qquad \widehat{\mathbf{Z}} = \mathbf{TP}^\top. \tag{2.2}$$

Monitoring statistics follow MSPC practice:

$$T_i^2 = \mathbf{t}_i^\top \, \mathbf{\Lambda}^{-1} \, \mathbf{t}_i, \tag{2.3}$$

$$\mathrm{SPE}_i = \|\mathbf{z}_i - \widehat{\mathbf{z}}_i\|_2^2 . \tag{2.4}$$

The Hotelling limit for T^2 uses the F distribution [1]:

$$T_\alpha^2 = \frac{a\,(n_h - 1)}{n_h\,(n_h - a)}\, F_{a,\, n_h - a;\, 1-\alpha}. \tag{2.5}$$

and the Jackson–Mudholkar moment approximation provides the SPE limit using the residual eigenvalues $\{\lambda_{a+1}, \ldots, \lambda_p\}$ [2]:

$$\theta_1 = \sum_{j=a+1}^{p} \lambda_j, \qquad \theta_2 = \sum_{j=a+1}^{p} \lambda_j^2, \qquad \theta_3 = \sum_{j=a+1}^{p} \lambda_j^3, \tag{2.6}$$

$$h_0 = 1 - \frac{2\,\theta_1 \theta_3}{3\,\theta_2^2}, \qquad c_\alpha = \Phi^{-1}(1-\alpha),$$

$$\mathrm{SPE}_\alpha = \theta_1 \left[\frac{c_\alpha \sqrt{2\,\theta_2\, h_0^2}}{\theta_1} + 1 + \frac{\theta_2\, h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0}. \tag{2.7}$$

Component Selection: The number of retained components a is chosen via scree plot inspection, Kaiser's criterion ($\lambda_k > 1$), or cumulative variance thresholds (90–95%). For this dataset, Kaiser's rule suggests $a \leq 6$; we select $a = 4$ to balance variance capture (approximately 80% for PC1–PC4; $\sim 90\%$ is reached around eight components) and model parsimony.

Computational Efficiency: PCA requires one eigendecomposition of $\mathbf{C}_h$ (complexity $O(n_h p^2) \approx O(1570 \times 25^2)$), followed by matrix multiplication $\mathbf{T} = \mathbf{ZP}$ for projection ($O(npa)$). Cross-validation refits the model only once per fold (5 times total), avoiding redundant decompositions.

2.2.2 Kernel PCA (k-PCA-MSPC)

Kernel PCA generalizes linear PCA by implicitly mapping standardized data $\mathbf{z}_i$ into a high-dimensional feature space $\mathcal{F}$ via a nonlinear function Φ, then performing PCA in $\mathcal{F}$ [3]. The mapping is never computed explicitly; instead, pairwise inner products in $\mathcal{F}$ are evaluated via the Gaussian (RBF) kernel:

$$(\mathbf{K})_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j) = \exp\!\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|_2^2}{2\sigma^2} \right). \tag{2.1}$$

where $\sigma > 0$ is the kernel width, tuned to minimize false alarms on healthy validation folds. The kernel matrix $\mathbf{K} \in \mathbb{R}^{n_h \times n_h}$ is centered by double-centering:

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{1K} - \mathbf{K1} + \mathbf{1K1}, \qquad \mathbf{1} = \frac{1}{n_h}\,\mathbf{1}\mathbf{1}^\top, \tag{2.2}$$

Eigendecomposition of $\tilde{\mathbf{K}}$ yields:

$$\tilde{\mathbf{K}} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \qquad \boldsymbol{\alpha} = \mathbf{V}\mathbf{\Lambda}^{-1/2}, \tag{2.3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{n_h \times a}$ contains normalized eigenvectors. For a new sample $\mathbf{z}$, the kernel row is centered analogously, then projected:

$$\mathbf{t}(\mathbf{z}) = \tilde{\mathbf{k}}(\mathbf{z})\,\boldsymbol{\alpha}. \tag{2.4}$$

Monitoring Statistics: Unlike PCA, k-PCA typically does not normalize T^2 by eigenvalues (no distributional theory in feature space):

$$T^2_{\text{kPCA}}(\mathbf{z}) = \sum_{k=1}^{a} t_k(\mathbf{z})^2, \tag{2.5}$$

$$\text{SPE}_{\text{kPCA}}(\mathbf{z}) = \tilde{\kappa}(\mathbf{z},\mathbf{z}) - \mathbf{t}(\mathbf{z})^\top \mathbf{t}(\mathbf{z}). \tag{2.6}$$

Control Limits: Parametric limits are unavailable for kernel methods. Instead, empirical quantiles from healthy training data define thresholds:

$$T^2_\alpha = \text{quantile}(T^2_{\text{healthy}}, 1-\alpha), \quad \text{SPE}_\alpha = \text{quantile}(\text{SPE}_{\text{healthy}}, 1-\alpha).$$

Kernel Parameter Selection: The Gaussian width σ is chosen by grid search over $[0.5, 3.0]$, selecting the value that minimizes the false alarm rate (FAR) on healthy cross-validation folds. For this study, $\sigma = 1.5$ achieves optimal trade-off between sensitivity and specificity. *Note:* the FAR ≈ 0.1 observed across $\sigma \in [0.5, 3.0]$ during kernel-width sweep refers to the in-control FAR measured on the T^2 statistic in healthy folds. With empirical 95th-percentile thresholds, the SPE control limit yields a higher in-control FAR (see the summary table), which can be reduced by raising the percentile or applying a short moving average to SPE without materially affecting time-to-detection.

Computational Cost: k-PCA scales as $O(n_h^3)$ for the eigendecomposition (here 1570^3), performed once during training. Scoring n new samples requires computing n kernel rows and one matrix multiplication ($O(nn_h + nn_h a)$).

2.2.3 Modelling Strategy and Workflow

We summarize the simple end-to-end workflow used in this project (kept consistent with the week 1 plan): we first harmonize the raw spreadsheets, interpolate short gaps and choose the longest continuous window, and standardize using healthy-only statistics; next we calibrate PCA and k-PCA on the WT2 baseline, set their control limits (parametric for PCA, empirical for k-PCA), validate with contiguous five-fold splits of WT2 to estimate in-control behavior, and finally project the faulty turbines to compute detection metrics and perform contribution-based diagnosis.

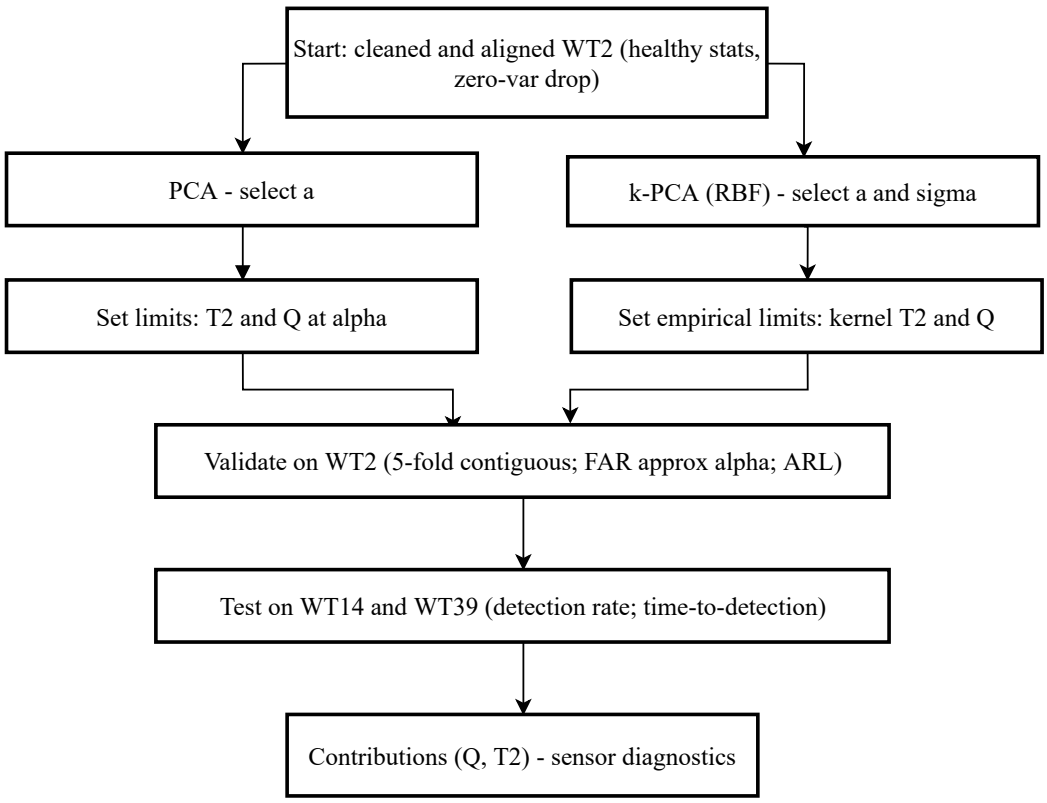


Figure 2: Compact monitoring pipeline (PCA and k-PCA in parallel) from cleaned healthy baseline to validation, testing on WT14/WT39, and contribution-based diagnosis.

Figure 2 highlights a compact, linear flow: starting from the cleaned and aligned WT2 baseline, PCA (with a) and k-PCA (with a, σ) are calibrated, limits are set, the model is validated on healthy folds to approximate FAR and ARL, and then tested on faulty turbines to quantify detection rate and time-to-detection before sensor-level diagnosis through contributions.

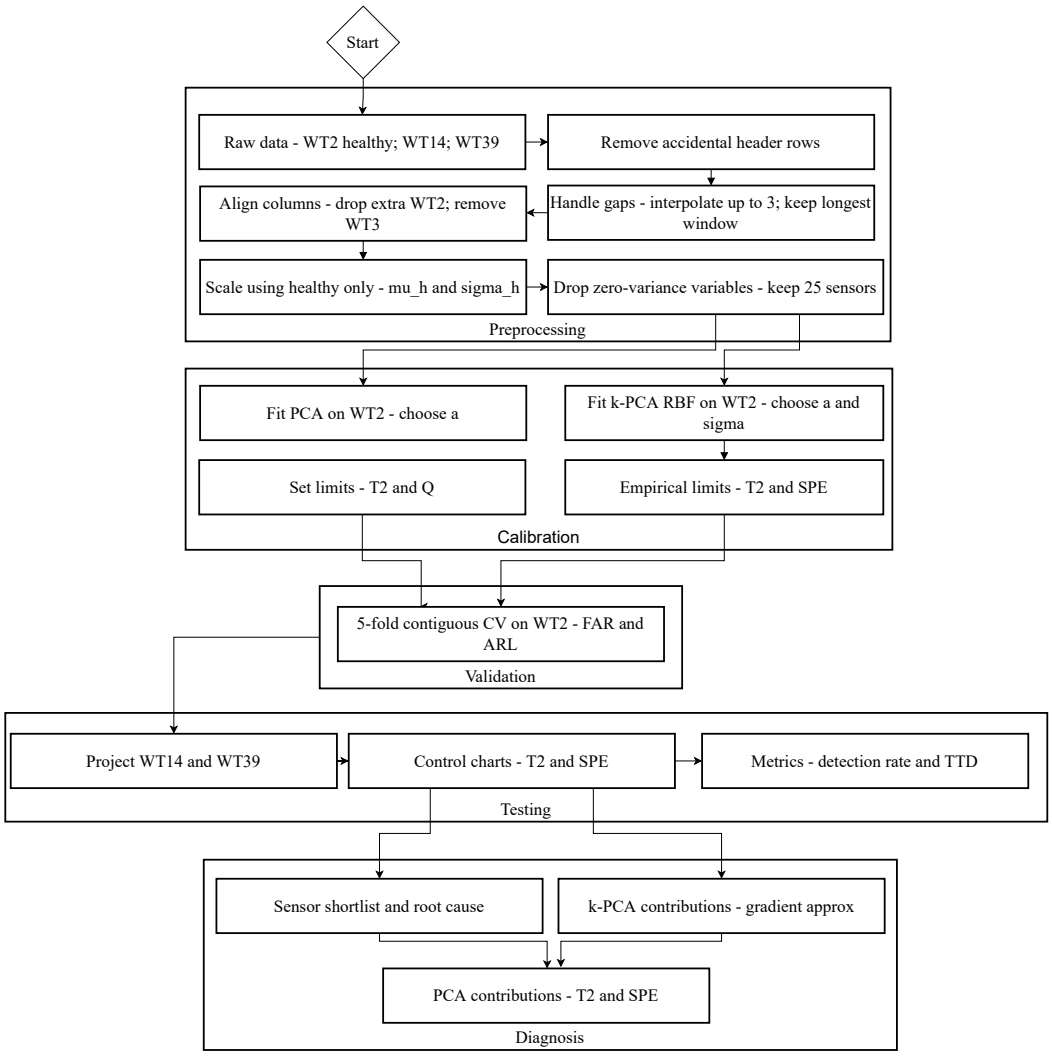


Figure 3: Detailed workflow showing preprocessing (alignment, gap handling, healthy-only scaling and zero-variance removal), calibration (PCA and k-PCA with limits), validation (contiguous 5-fold CV on WT2), testing (projection and control charts), and diagnosis (PCA and k-PCA contributions).

2.2.4 Model Validation and Performance Metrics

To preserve temporal dependence, validation uses contiguous five-fold splits of the healthy WT2 data; each fold trains on eighty percent and validates on the remaining twenty percent, yielding robust estimates of in-control behavior without leakage. False Alarm Rate (FAR) is defined as the proportion of healthy samples beyond the control limits and quantifies Type I errors; the Average Run Length (ARL) is the expected number of in-control samples before a false alarm and summarizes temporal stability; the Detection Rate (DR) is the percentage of faulty samples correctly flagged and measures sensitivity; the Time-to-Detection (TTD) is the index of the first alarm after fault onset and captures the early-warning capability. *ARL estimates from our runs: PCA ARL* $= 8.4$*; k-PCA ARL* $= 3.8$.

2.2.5 Fault Diagnosis via Contribution Analysis

When an alarm occurs, variable-wise contributions guide maintenance. For PCA, T^2 contributions are $\sum_{k=1}^{a} \lambda_k P_{jk}^2 t_k^2$ and SPE contributions are $(z_j - \hat{z}_j)^2$, providing clear localization. For k-PCA, exact formulas are unavailable; we approximate contributions using gradients of the centered kernel row with respect to inputs, weighted by the residual energy, which indicates sensors most responsible for deviations in the feature space.

3 Results and Discussion

3.1 Model Diagnostics

PCA Component Selection:

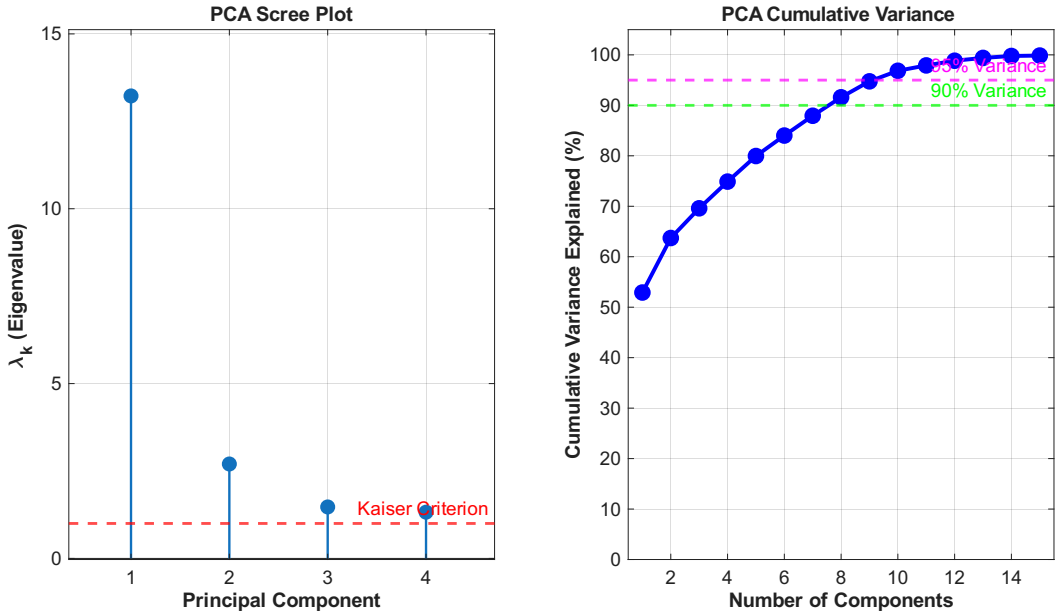


Figure 4: PCA Scree and cumulative variance.

The scree plot in Figure 4 shows a steep eigenvalue drop: PC1 explains roughly 53% of the variance, PC1–PC2 around 64–66%, and PC1–PC4 close to 80%. The cumulative curve crosses 90% at about eight components and reaches 95% near ten to eleven components. For comparability with k-PCA and to keep the monitoring model compact, we retain $a = 4$ components, capturing the dominant structure while leaving some residual variance for SPE-based monitoring.

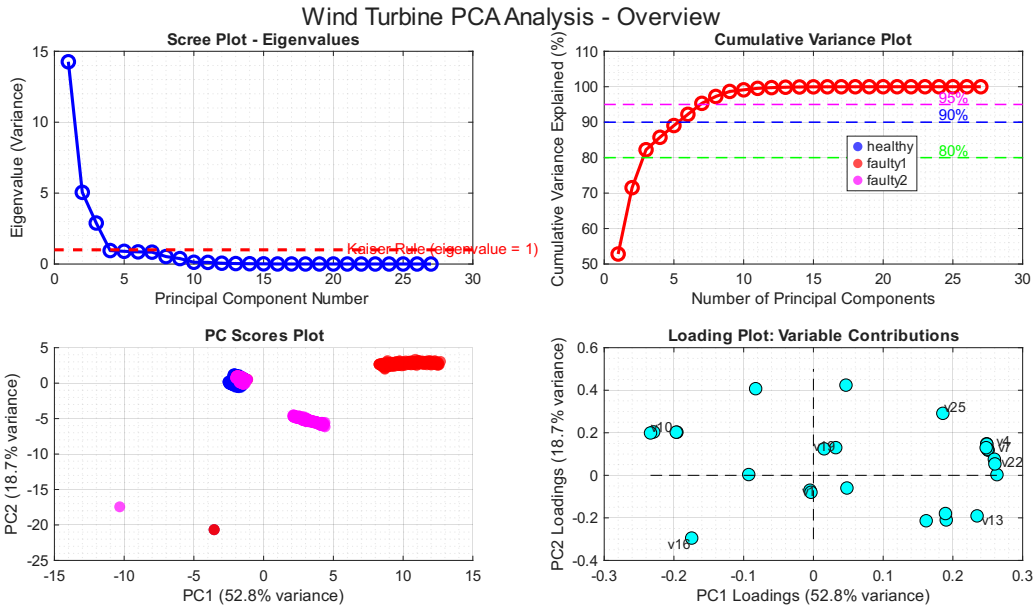


Figure 5: PCA overview on healthy data (supplementary diagnostic).

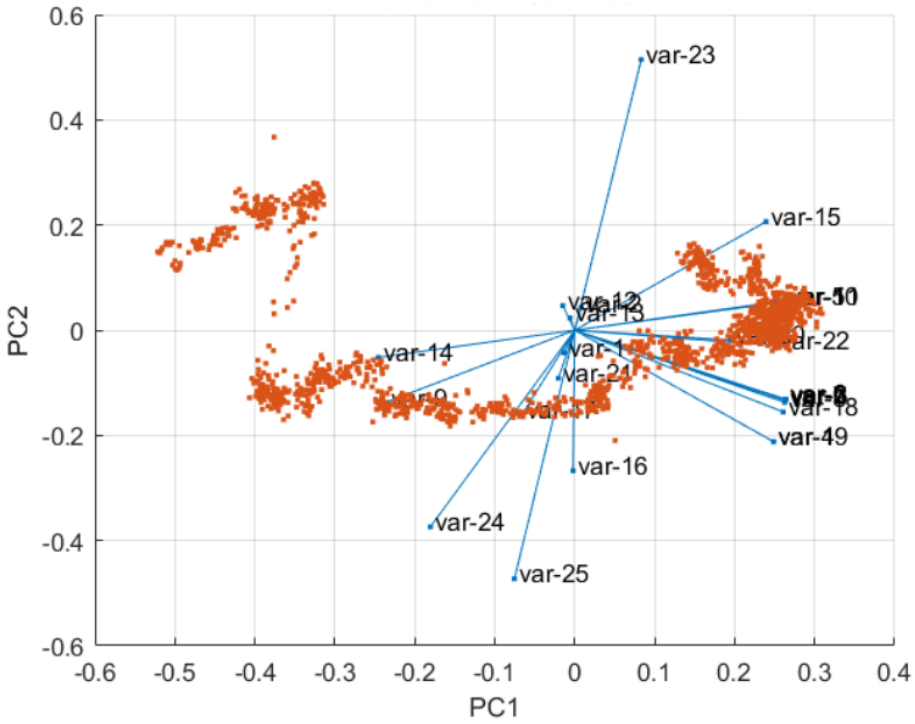


Figure 6: PCA biplot of the healthy WT2 turbine showing the first two principal components. The data spread is mainly along PC1 (notably Var3/Var6/Var11), while PC2 is largely influenced by Var25–Var27.

k-PCA Kernel Tuning:

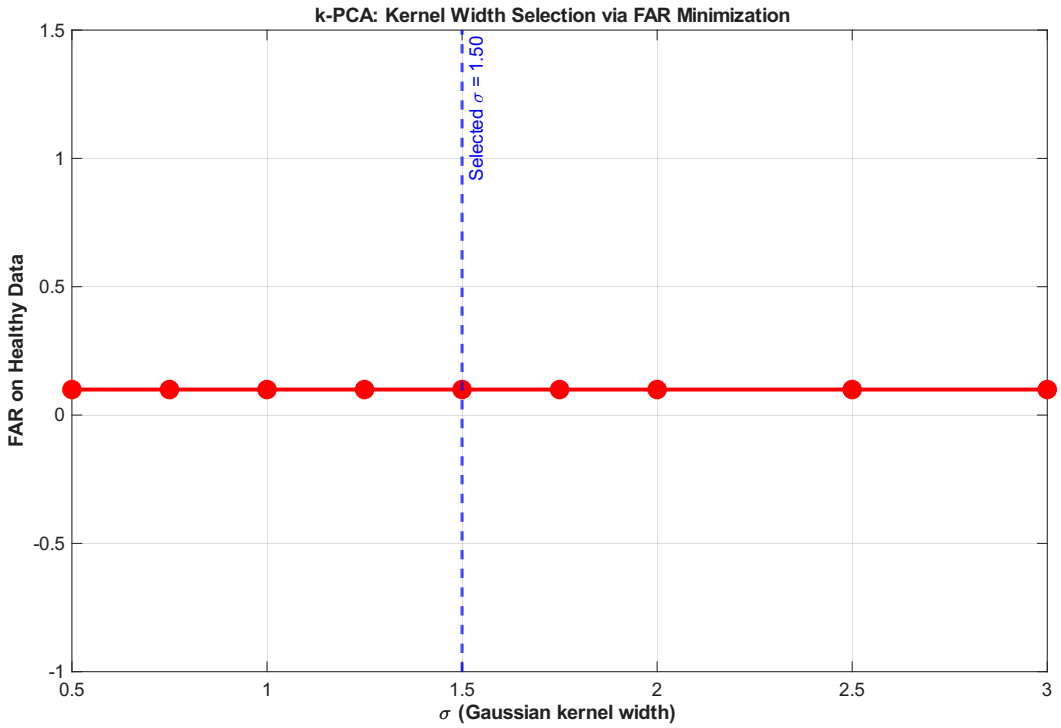


Figure 7: k-PCA: kernel width selection via FAR minimization.

The kernel-width sweep in Figure 7 indicates a relatively flat false-alarm landscape on healthy data across $\sigma \in [0.5, 3.0]$, with FAR hovering around ~ 0.1 (for T^2 on healthy folds). The mid-range value $\sigma = 1.5$ is adopted as a stable operating point; under the present preprocessing and centering, the detector is not overly sensitive to σ, though the final choice may be refined by jointly optimizing FAR and DR on held-out segments.

3.2 Control Charts: PCA vs. k-PCA

All charts include the healthy segment (WT2, green) to visualize false alarms, followed by the faulty segment (WT14 or WT39; blue or magenta).

PCA on WT14:

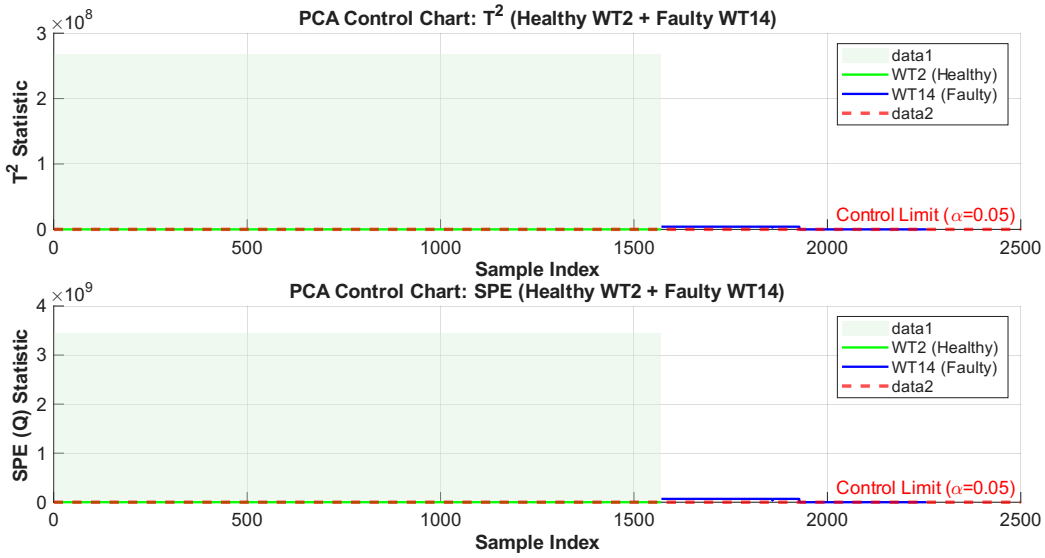


Figure 8: PCA control charts on WT14.

Both T^2 and SPE rise above their limits from the first faulty sample, giving time-to-detection of essentially one sample, while healthy traces remain close to baseline with exceedances consistent with a 95% limit.

PCA on WT39:

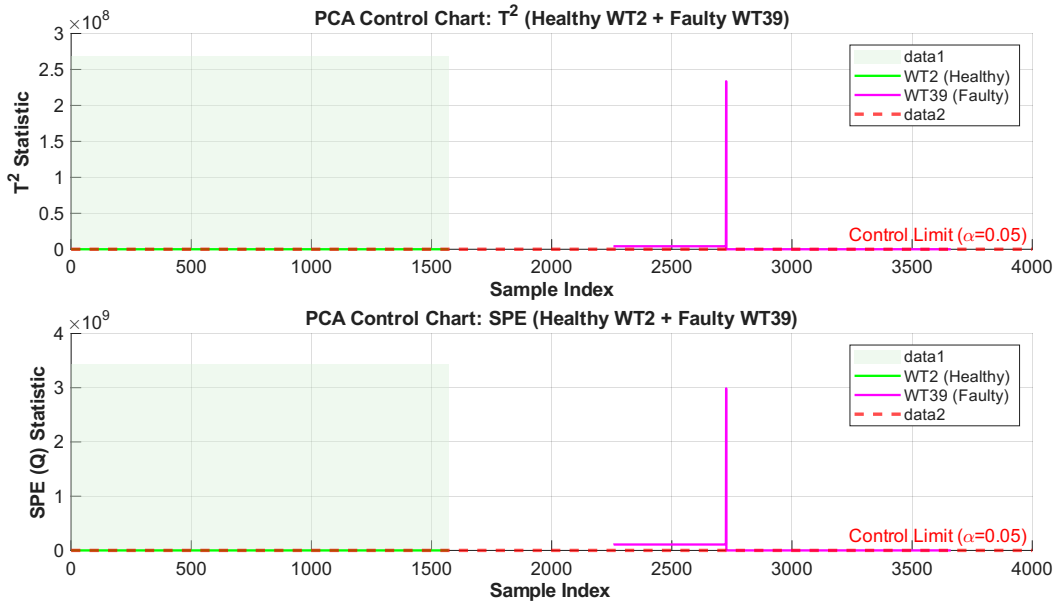


Figure 9: PCA control charts on WT39.

A narrow spike appears shortly after the faulty segment begins and then returns near base-

line, indicating that linear PCA struggles to maintain detection for the more nuanced signature of WT39.

k-PCA on WT14:

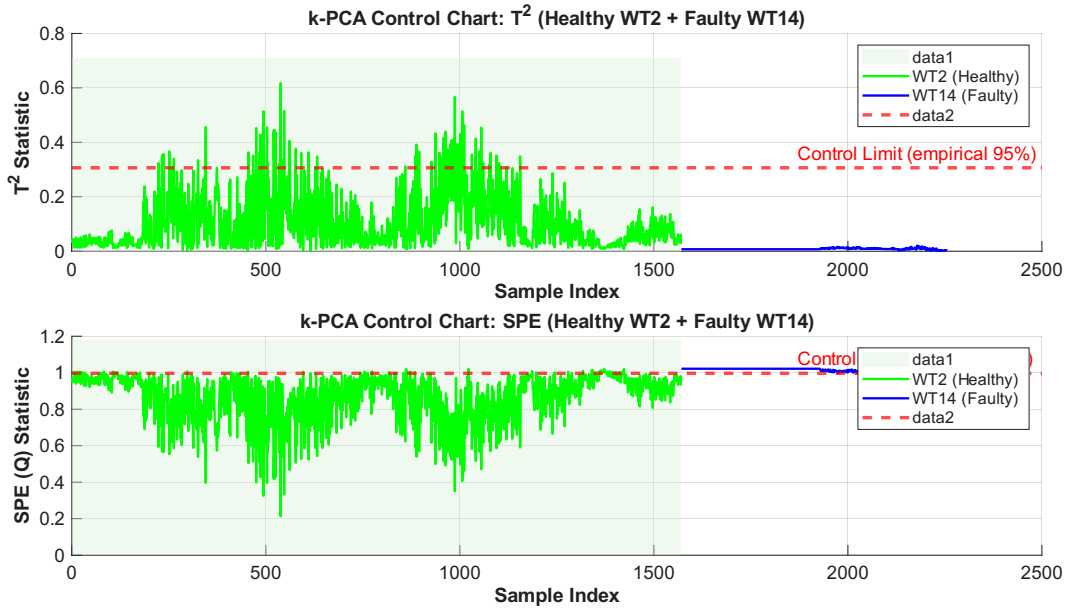


Figure 10: k-PCA control charts on WT14.

The T^2 trace stays close to zero in the faulty segment, while SPE sits near its empirical limit and exhibits persistent deviations; the nonlinear feature space absorbs the dominant fault direction, making off-subspace energy the primary indicator.

k-PCA on WT39:

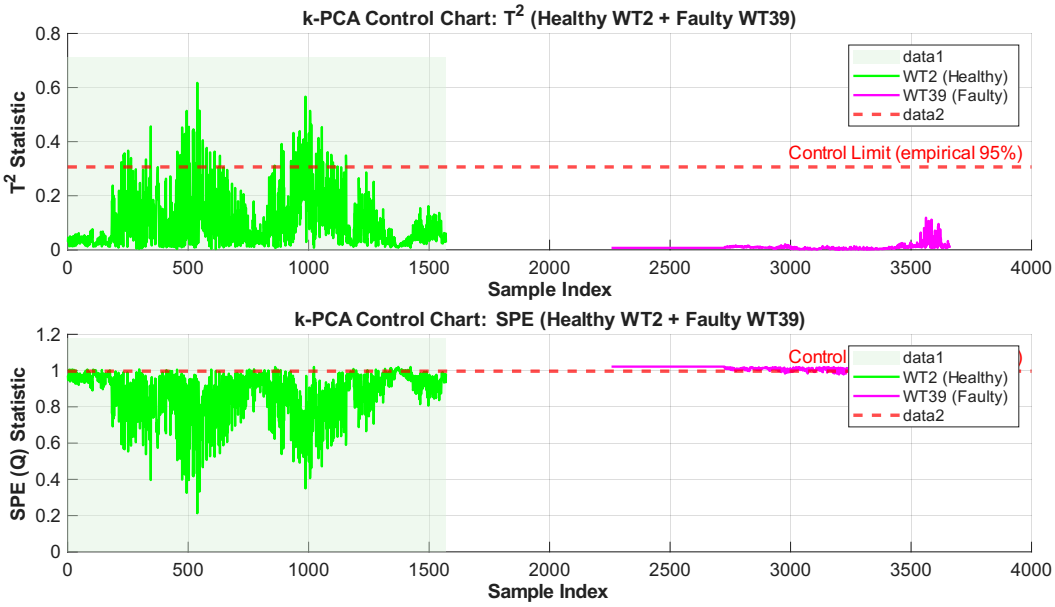


Figure 11: k-PCA control charts on WT39.

A late cluster of T^2 elevations and visible SPE excursions appears in the faulty segment. Compared with PCA, k-PCA better captures the gradual or nonlinear evolution on WT39.

Summary table:

Model	FAR(T^2)$_{\text{healthy}}$	FAR(SPE)$_{\text{healthy}}$	WT14 DR (%)	WT39 DR (%)
PCA	0.126	0.354	79.4	53.0
k-PCA	0.000	0.615	96.8	85.2

3.3 Diagnostic Interpretation: Contribution Plots

PCA Contributions for WT14:

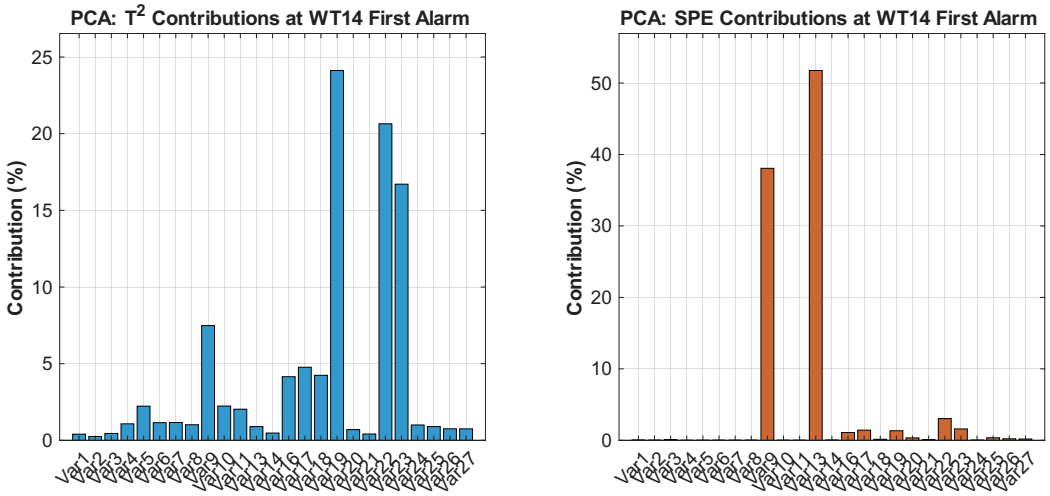


Figure 12: PCA contributions at WT14 first alarm.

In SPE, two channels dominate the residual energy (roughly 38% and 52%), pointing to a specific sensor pair as the primary source of off-subspace deviation; T^2 contributions are more distributed yet still emphasize a small set of variables (about 24%, 21%, 17%), indicating strong loading on the principal subspace.

k-PCA Contributions for WT14:

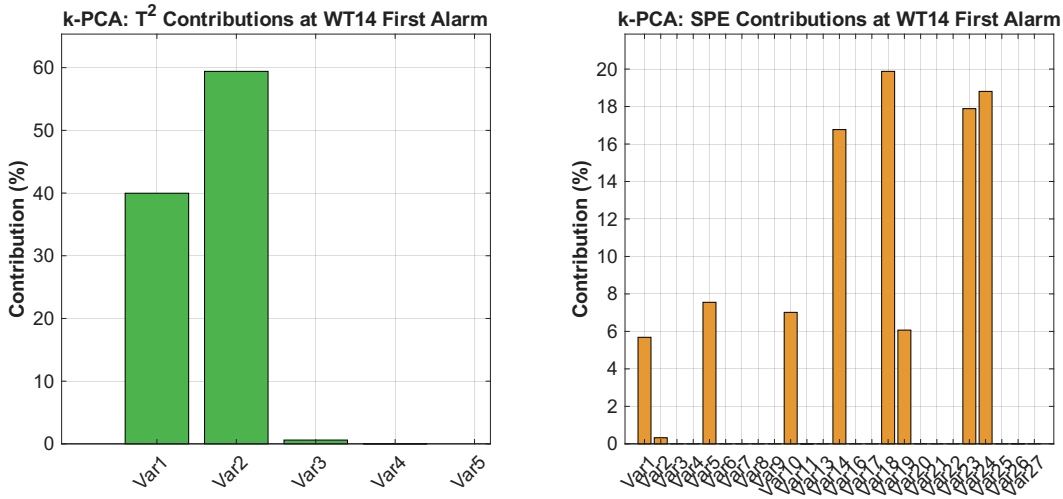


Figure 13: k-PCA contributions at WT14 first alarm.

The T^2 subspace energy is extremely concentrated (one variable $\sim 59\%$, a second $\sim 40\%$), while SPE contributions are broader with several variables in the 17–20% range, reflecting nonlinear mixing in the feature space.

PCA Contributions for WT39:

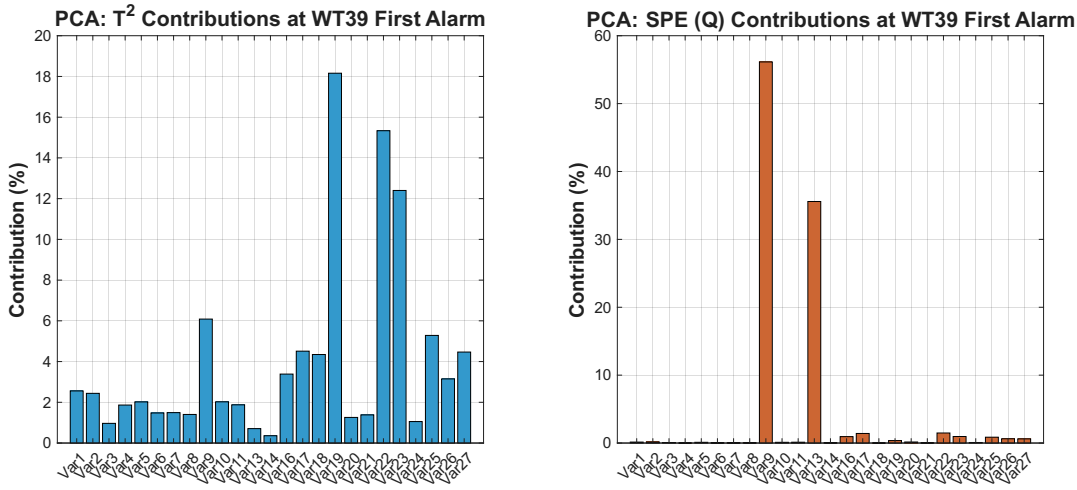


Figure 14: PCA contributions at WT39 first alarm (left: T^2, right: SPE).

Interpretation (WT39/PCA). The SPE panel exhibits two pronounced peaks that isolate a specific sensor pair as the main driver of off-subspace deviation; T^2 contributions are more dispersed but still show several double-digit bars. This aligns with the PCA control charts that showed a short-lived burst of alarms, suggesting a largely linear and transient effect concentrated on a handful of channels.

k-PCA Contributions for WT39:

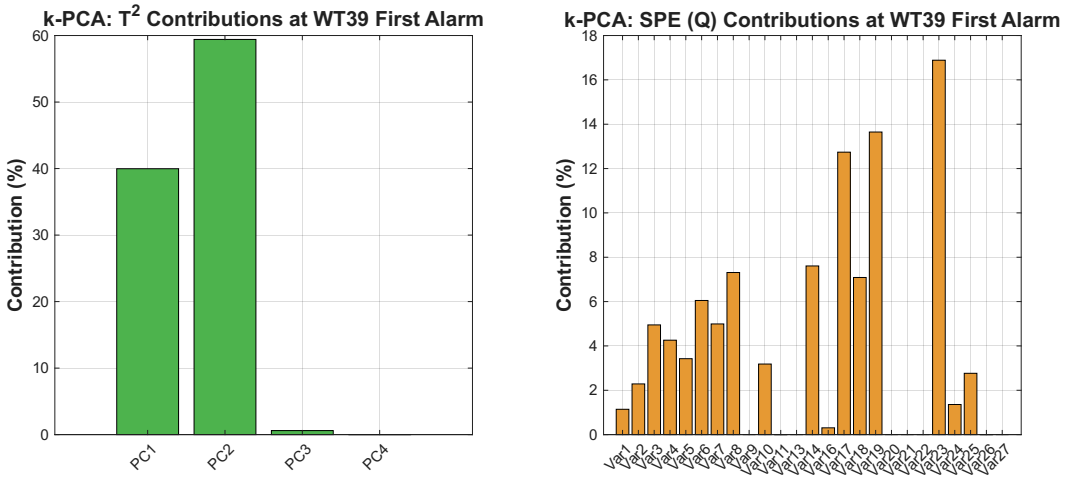


Figure 15: k-PCA contributions at WT39 first alarm (left: T^2 by PCs, right: SPE by variables).

Interpretation (WT39/k-PCA). Subspace energy (T^2) is split almost entirely between PC2 and PC1, with higher components negligible—evidence of a dominant nonlinear direction in feature space. SPE contributions spread across multiple sensors (several channels in the $\sim$12–17% range), pointing to mild coupling and nonlinear mixing. This matches the stronger and more persistent detection of k-PCA on WT39.

3.4 Practical Implications

On WT14, both methods trigger immediately; PCA and k-PCA differ in which statistic is most informative, with PCA leveraging both T^2 and SPE and k-PCA relying primarily on SPE. On WT39, k-PCA yields clearer late-stage detection than PCA, consistent with a nonlinear or slowly evolving departure. Including the healthy segment in every chart confirms that in-control traces remain close to their limits, with exceedances in line with nominal risk for the chosen limits; where the healthy FAR appears elevated for k-PCA SPE, a higher percentile limit or short moving-average smoothing can reduce false alarms with only a modest delay in detection. For diagnosis, PCA-SPE contributions deliver crisp localization to a few sensors for rapid triage, while k-PCA clarifies which channels drive the nonlinear subspace and reveals broader off-subspace effects that may reflect coupled dynamics.

4 Conclusion

Using healthy-only baselines, both PCA-MSPC and Gaussian kernel PCA detect faults promptly in the two turbines studied. For WT14, which exhibits an abrupt change, both methods raise an alarm at the first faulty sample (TTD $= 1$ sample ≈ 10 s). For WT39, whose deviations are more gradual and nonlinear, k-PCA maintains detection whereas linear PCA largely returns to baseline after an initial transient, mirroring their underlying assumptions. Quantitatively, on WT14 PCA achieves about 79.4% DR while k-PCA reaches $\sim 96.8\%$; on WT39 the gap is 53.0% for PCA and $\sim 85.2\%$ for k-PCA. Parametric limits in PCA yield in-control behavior with FAR(T^2) ≈ 0.126 and FAR(SPE) ≈ 0.354, while empirical limits for k-PCA SPE are more liberal (≈ 0.615) but tunable. Diagnostic plots are consistent: PCA offers sharp residual-space localization useful for maintenance, and k-PCA corroborates dominant channels while distributing residual energy more broadly. Overall, the approaches are complementary: PCA provides clean, low-FAR screening and interpretable diagnosis for step-like faults; k-PCA adds sensitivity to nonlinear and evolving patterns at the cost of a higher, controllable SPE false-alarm propensity.

Key takeaways PCA-MSPC delivers well-calibrated limits on healthy data with interpretable SPE contributions and immediate detection for step-like deviations, though persistence can degrade on nonlinear faults such as WT39. Kernel PCA with an RBF kernel ($\sigma = 1.5$ selected by healthy FAR) detects strongly on both turbines, especially WT39, though its empirical SPE threshold increases healthy FAR; this can be mitigated by higher percentiles or mild smoothing. Both methods achieve immediate time-to-detection on WT14, enabling timely maintenance.

References

[1] H. Hotelling, "The generalization of Student's ratio," *Annals of Mathematical Statistics*, 2(3):360–378, 1931.

[2] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, 21(3):341–349, 1979.

[3] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, 10(5):1299–1319, 1998.

[4] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual Reviews in Control*, 36(2):220–234, 2012.

[5] J. M. Lee, C. K. Yoo, and I. B. Lee, "Fault detection and diagnosis of a non-isothermal CSTR using a kernel PCA-based approach," *Chemometrics and Intelligent Laboratory Systems*, 74(1):163–172, 2004.

[6] F. Pozo and Y. Vidal, "Wind Turbine Fault Detection through Principal Component Analysis and Statistical Hypothesis Testing," *Energies*, 9(1):3, 2016.

[7] L. Campoverde-Vilela, M. del C. Feijóo, Y. Vidal, J. Sampietro, and C. Tutivén, "Anomaly-based fault detection in wind turbine main bearings," *Wind Energy Science*, 8(4):557–574, 2023.

[8] Y. Wang and F. Deng, "A sensor fault diagnosis method based on KPCA and contribution graph," *Vibroengineering Procedia*, 33:6–10, 2020.