1 Feedback Incorporation & Pretreatment Adjustment (0.5p)

We audited our pipeline against the instructor's feedback and updated the pretreatment accordingly. **Already satisfied:** retain correlated sensors (no pruning); drop only constant (zero-variance) variables; avoid clipping genuine extremes; use a true biplot (scores+loadings); include core PCA visuals (scree, cumulative variance, loading bars, grouped score plots) with minimal clutter. **Now fixed:** switch from all-data scaling to *healthy-based autoscaling* (fit μ, σ on Healthy WT only and apply to all WTs); replace indices by sensor names; color scores by time/cycle; add evidence plots (pre/post scaling boxplots, post-scaling correlation heatmap, key sensor time-series). **Now implemented:** time-aware imputation for short gaps and longest complete window for long gaps *before* PCA; alignment to the least-frequent sampling rate; contiguous blocked validation on Healthy; testing on both Faulty units per A-level task.

Numeric top-row labels in Excel (fixed). A numeric header row was being ingested as data. We added a defensive check in `load_turbine_data` to detect and remove an integer-only top row (few distinct values or exactly $1{:}n_{\text{vars}}$), so correlations/PCA use only sensor measurements.

Time-aware preprocessing functions (added). `time_aware_preprocess` orchestrates: `impute_short_gaps` (linear interpolation up to `MaxGap`); `select_longest_complete_window` (trim to the longest NaN-free segment); `longest_true_run` (index helper). This prevents unjustified row deletion, preserves temporal order, and ensures PCA uses time-consistent data.

Healthy-based variance filtering. Two sensors (`Var12`, `Var15`) had $\sigma_h{=}0$ on Healthy and were removed ($27{\to}25$), avoiding singularities.

Effect of healthy-based scaling. Switching to *healthy-based scaling* and dropping zero-variance sensors clarified the PCA structure (PC1+PC2 explain nearly all variance) and strengthened Healthy vs. Faulty separation in score space.

2 Data Centering & Scaling (0.5p)

In `pca_implementation(...)` we apply *healthy-based autoscaling*: compute μ_h, σ_h on the Healthy block (first n_h rows) and standardize all rows,

$$z_{i,j} = \frac{x_{i,j} - \mu_{h,j}}{\sigma_{h,j}}, \quad \mu_{h,j} = \tfrac{1}{n_h}\sum_{i=1}^{n_h} x_{i,j}, \quad \sigma_{h,j} = \sqrt{\tfrac{1}{n_h-1}\sum_{i=1}^{n_h}(x_{i,j} - \mu_{h,j})^2}.$$

Variables with $\sigma_{h,j}{=}0$ are removed. This centers Healthy at ≈ 0 with unit variance, avoids leakage from faults, and improves conditioning (logs: `kept 25/27; dropped 2 (sd_h==0)`; rank/condition reported).

3 Outliers & Missing Data—Mitigation, Synchronization & Sampling (0.5p)

Extremes (outliers). We *retain* fault-driven extremes (no clipping/winsorizing) to preserve diagnostic signal for PCA and T^2/SPE; interpretation/localization is done later via top-Q/T^2 contribution plots.

Missing values & mitigation. We avoid dropping time rows. Short gaps ($\leq$3 samples) are linearly interpolated with bounded ends; if longer gaps remain, we keep the longest contiguous NaN-free window per unit. In this run no rows were removed (`Healthy 1570`, `Faulty1 686`, `Faulty2 1405`) and the PCA input had **0** missing values.

Synchronization / sampling. The sheets (`No.2WT`, `No.14WT`, `No.39WT`) contain *no timestamps*. Per instructor guidance we assume equal sampling and mark explicit re-synchronization as **N/A (with justification)**. If timestamps become available, we will convert to timetables, `retime` to the least-frequent common clock (mean/median aggregation), align variables/ordering, then re-apply the same gap handling and scaling.

Evidence (this run). Matrix entering PCA: $n = 3661$ rows; $p = 25$ kept variables after dropping `Var12`, `Var15` ($\sigma_h = 0$); missing values at PCA entry: **0**.

4 Visualizing Pretreated Data (0.5p)

After time-aware cleaning (short-gap interpolation + longest NaN-free window) and before PCA, we visualize the pretreated matrix to verify class separability, scaling effects, and inter-variable structure (with `Var12`, `Var15` removed; others retained).

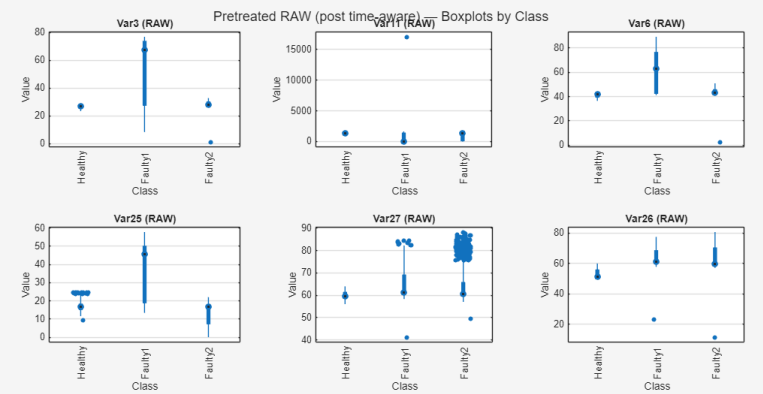


Figure 1: Pretreated RAW (post time-aware) — boxplots by class.

Raw units: Healthy is compact; Faulty1 elevated/volatile; Faulty2 near-healthy with outliers—underscoring the need for scaling.

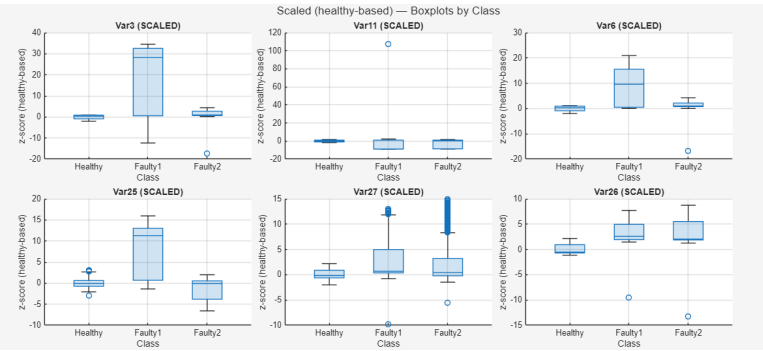
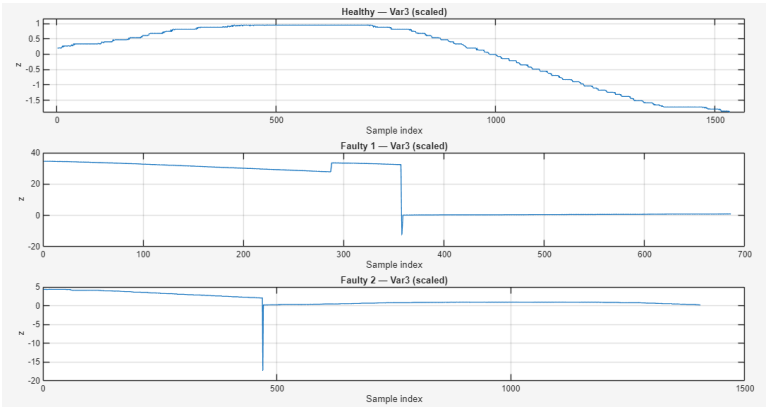


Figure 2: Scaled (healthy-based) — boxplots by class.

After healthy-based z-scoring: Healthy ≈ 0; Faulty1 shows many-σ shifts; Faulty2 mild offsets—classes now directly comparable.



(a) Scaled time profiles for a key variable across units.

Figure 3: Inter-variable structure and temporal behavior after pretreatment.

Interpretation of Var3 time profiles. The healthy unit shows a mild drift within $\pm 2\sigma$. *Faulty 1* starts with a strong bias ($+30$–35σ) and shifts abruptly near sample ~ 350 to about -10σ, indicating a sensor offset/miscalibration. *Faulty 2* stays near healthy levels but shows a sharp transient ($\sim -17\sigma$) around sample ~ 480. Such extremes explain why Var3 dominates PC1 loadings and charts detect immediate faults in Faulty 1 and transients in Faulty 2.

5 Modelling Goal (0.5p)

Detect and monitor **faults** by learning the nominal behavior from the *healthy* unit (WT2) and projecting two *faulty* units (WT14, WT39) into the same latent space. We will: (i) calibrate **PCA**, **Robust-PCA**, and **k-PCA (RBF)** on healthy data; (ii) deploy **Hotelling's** T^2 and Q**/SPE** charts; (iii) **diagnose** alarms via contribution analysis. *Upstream pretreatment:* drop WT3 (inconsistent p), drop the extra-quality variable on one faulty unit, align common variables & order.

6 Model Calibration: Tools, Methodology, Data (0.5p)

Data & scaling. Train only on cleaned/aligned WT2 (common $p{\approx}27$). Remove zero-variance-in-healthy variables. Use *healthy-based* z-scores (WT2 μ_h, σ_h) and apply unchanged to WT14/WT39 (no leakage; interpretable limits).
Models.

- **PCA (linear):** SVD on scaled $X_h \Rightarrow$ loadings P, scores $T{=}X_hP$, residuals $E{=}X_h{-}\hat{X}_h$.
- **Robust-PCA:** robust (μ_r, Σ_r) via MCD/ROBPCA on WT2; eigendecompose Σ_r to obtain $P^{(r)}, \Lambda^{(r)}$ (down-weights outliers; often different loadings vs PCA).
- **k-PCA (RBF):** $k(x,x'){=}\exp(-\|x{-}x'\|^2/(2\sigma^2))$; center Gram matrix, eigendecompose; new-sample scores via kernel trick with centered kernel vector.

Choosing #PCs (a). Start from **90–95%** cumulative variance (scree+Kaiser) and confirm with validation that in-control false alarms meet α.
Control limits (healthy-only).

$$T^2_\alpha \approx \frac{a(n-1)}{n-a} F_{a,n-a,(1-\alpha)}, \qquad Q_\alpha = \theta_1\Big(\frac{z_\alpha\sqrt{2\theta_2}\,h_0}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}\Big)^{1/h_0}$$

with $\theta_k = \sum_{j>a}\lambda_j^k$, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $z_\alpha = \Phi^{-1}(1-\alpha)$. For k-PCA: center the kernel, tune σ on WT2 to keep FAR$\approx\alpha$; initial limits can also be set from healthy-score/residual empirical quantiles (95th/99th) and then verified against FAR on WT2.
Why both T^2 and Q? T^2 flags unusual *score-space* behavior (within-subspace), while Q flags *residual-space* energy (off-subspace). Using both improves sensitivity to different fault modes.

7 Validation (0.5p)

Blocked CV on WT2 (contiguous 5-fold). Train on 4 adjacent folds, fix (a, σ) & limits; evaluate held-out fold **FAR** at $\alpha \in \{0.01, 0.05\}$ for T^2 and Q. Report fold-wise FAR and stability of a (and σ).
Optional stability. Estimate in-control **Average Run Length (ARL)** on WT2.
Sensitivity. Inspect impact of variance target (85–95%), α, pruning near-collinear vars, and kernel width σ.

8 Testing & Metrics (0.5p)

Test data. WT14 & WT39 projected with the *same* WT2-based scaling/order; no refitting.
Primary metrics.

- **Detection rate:** % with $T^2{>}T^2_\alpha$ or $Q{>}Q_\alpha$.
- **Time-to-detection:** index of first exceedance (earliest alarm).
- **(Optional) Stability:** ARL on held-out healthy segments.

Diagnostics (contributions). For out-of-control sample i:

$$c^{(Q)}_{i,j} = e^2_{i,j}\ (\sum_j c^{(Q)}_{i,j} = Q_i), \qquad c^{(T^2)}_{i,j} = \sum_{k=1}^{a}\frac{(p_{j,k}\,t_{i,k})^2}{\lambda_k}\ (\sum_j c^{(T^2)}_{i,j} = T^2_i).$$

Deliverables note. *For each model–unit pair, we will provide paired T^2/Q charts on WT14/WT39 and* ***Top-Q contribution*** *barplots for the first out-of-control samples to localize sensors.*

9 Mathematical Summary (0.5p)

For scaled $X \in \mathbb{R}^{n \times p}$:

$$X = TP^\top + E, \quad P^\top P = I, \quad T = XP, \quad \hat{X} = TP^\top.$$

With $\Lambda_a = \mathrm{diag}(\lambda_1, \ldots, \lambda_a)$ and scores t_i:

$$T_i^2 = t_i^\top \Lambda_a^{-1} t_i, \qquad Q_i = \|e_i\|_2^2,\ e_i = x_i - \hat{x}_i.$$

Robust-PCA: use (μ_r, Σ_r) (MCD) $\Rightarrow$ robust $P^{(r)}, \Lambda^{(r)}$, then $T^{2(r)}, Q^{(r)}$. **k-PCA:** with centered kernel $\tilde{K} = V\Lambda V^\top$, scores $t = \Lambda_a^{-1/2} V_a^\top \tilde{k}$, and $Q^{(\mathrm{k})} = \tilde{k}^\top \tilde{k} - \tilde{k}^\top V_a V_a^\top \tilde{k}$.

9.1 PCA vs Robust-PCA — Differences

- **Calibration:** PCA uses classical mean/covariance; Robust-PCA uses robust estimates $\Rightarrow$ less influenced by outliers in *healthy* data.
- **Loadings/PCs:** Robust-PCA can rotate PCs away from directions defined by few extreme points; cumulative variance and selected a may differ.
- **Charts:** Robust T^2/Q typically yield *lower* false alarms on healthy data with occasional spikes, and can delay/advance detection depending on fault resemblance to prior outliers vs new behavior.

10 Model Diagram / Operations Flowchart (0.5p)

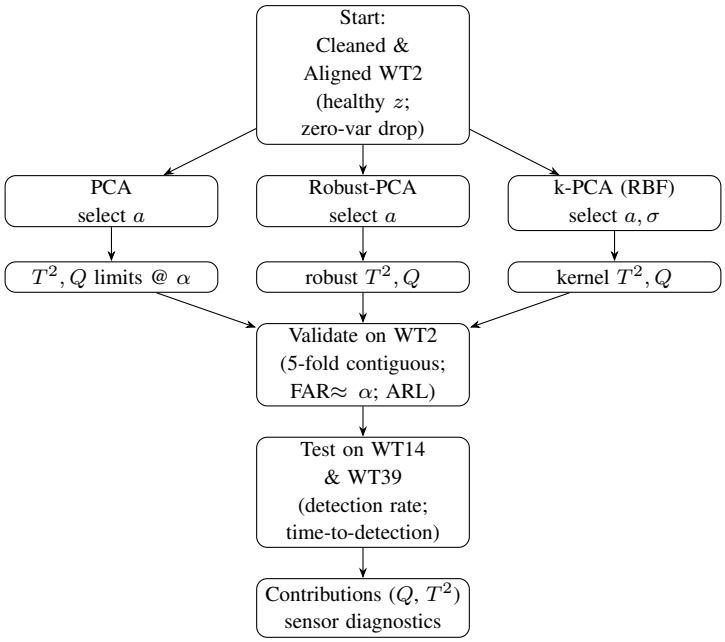


Figure 4: From cleaned healthy data to detection and diagnostics across PCA variants.

11 Team Roles & Responsibilities

Fasie Haider handles pretreatment and column alignment (applying dataset hints), healthy-based scaling on WT2 and zero-variance pruning, and QC; *Haider Ali* leads linear PCA, selects a (90–95% variance), sets T^2/Q limits, runs contiguous 5-fold validation on WT2, and leads testing/reporting; *Arman Golbidi* calibrates Robust-PCA (MCD/ROBPCA) and k-PCA (tuning σ via blocked CV) and compares against PCA; *All members* produce T^2/Q control charts for WT14/WT39, top-Q contribution diagnostics, and consolidate figures/results for the report.