ADAML 2025 Project Work – Wind Turbines Level B2

Intermediary Submission 1

Syeda Shaher Bano Bukhari

Lauri Heininen

14 September 2025

Communication

We established a GitHub repository for code sharing and version control. WhatsApp and online meetings are used for communication, ensuring both quick feedback and long-term traceability of scripts.

Data Format

The dataset is an Excel file with four sheets, each containing time series from a single turbine. The sheet for the healthy turbine, WT2, contains 1570 rows and 28 variables. WT14 has 686 rows and 27 variables, and WT39 has 1405 rows and 27 variables. WT3 was excluded because of an inconsistent structure.

All data were imported into MATLAB using *readmatrix*, resulting in matrices where rows correspond to observations and columns correspond to variables.

Challenges of the Data

Several challenges were identified. The number of variables differs across turbines (28 in WT2 versus 27 in WT14 and WT39). The time series are not synchronized, as each turbine has a different number of rows. WT14 contains one missing entry. Both WT14 and WT39 contain faulty operation segments that deviate strongly from the normal behavior.

The dataset lacks descriptive headers, which makes interpretation of the physical meaning of each variable difficult. More critically, some variables appear to be measured in different unit systems across turbines. For example, a variable with values around 1000–1500 in WT2 and WT39 reaches 17,000 in WT14, which suggests a difference between kW and W. Another variable shows ranges consistent with a °C vs °F mismatch. These inconsistencies need to be resolved before reliable interpretation.

Data Pretreatment

Before PCA, WT3 was excluded, and only the 27 variables common to WT2, WT14, and WT39 were kept. Potential timestamp or ID columns were removed. The missing value in WT14 was interpolated. WT14 and WT39 were split into faulty and healthy segments.

Z-score normalization was applied so that differences in scaling and possible unit mismatches would not bias PCA. This ensures comparability across turbines, although it limits the physical interpretation of the loadings. Figure (1) shows the variance explained by the principal components, which guided our choice of six components.

Exploratory Data Analysis with PCA

PCA was computed on the combined healthy data (WT2, WT14 good, WT39 good). The variance explained curve in Figure (1) shows that the first few PCs capture most of the variability.

The 2D and 3D PCA projections in Figure (2) show that the healthy turbines cluster closely together, while the faulty segments project far outside these regions. This indicates that PCA is able to separate faulty from normal operating states.

Figure (3) shows the biplots for the three turbines. WT2 forms a smooth, consistent pattern, while the faulty sections of WT14 and WT39 scatter more widely, reflecting unstable behavior. Both turbines, however, also contain clusters of points resembling healthy operation, suggesting periods of partial recovery.

Finally, Figure (4) displays the sensor loadings for the first three components. These loadings indicate which sensors drive the largest variance, though the absence of variable names and confirmed units makes detailed interpretation difficult.

Figures

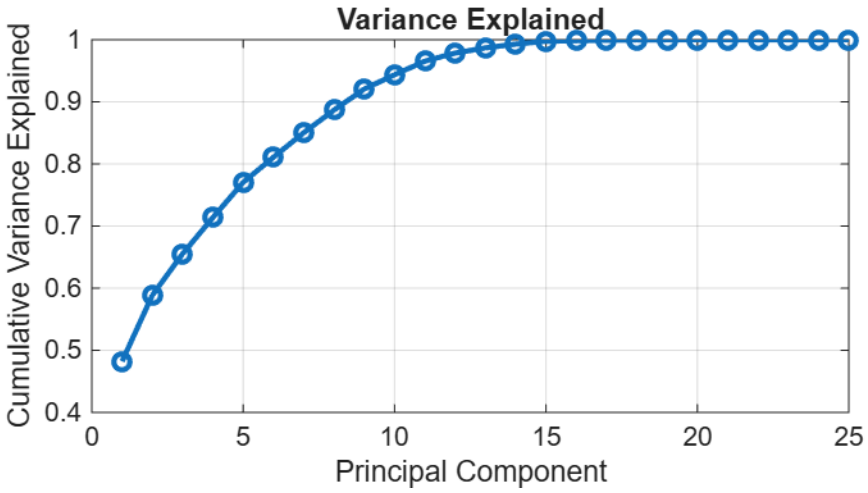

Figure 1. Cumulative variance explained by principal components.

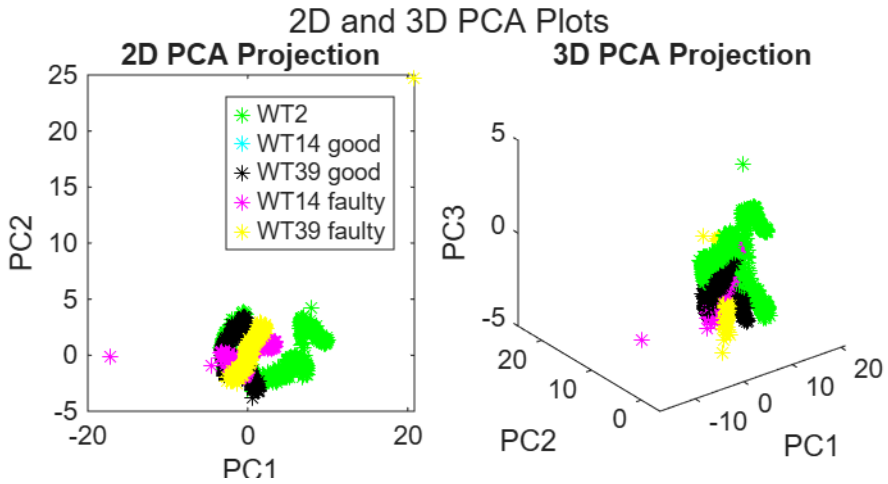

Figure 2. 2D and 3D PCA projections of healthy and faulty turbines.

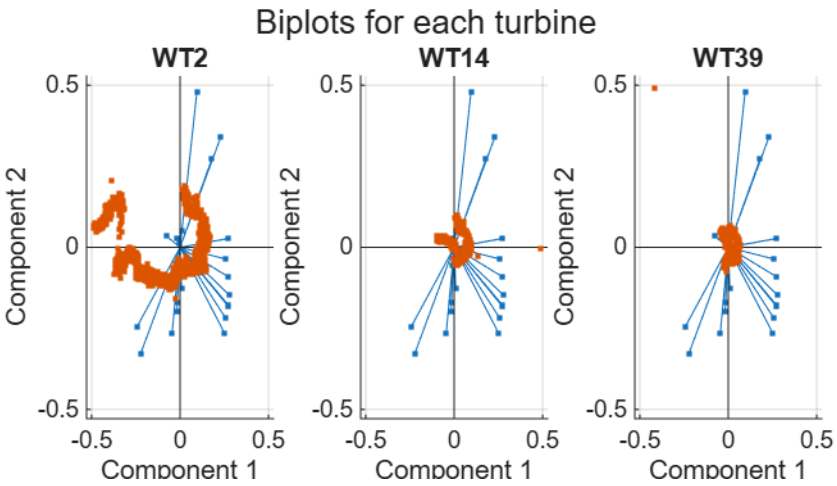

Figure 3. Biplots for turbines WT2, WT14, and WT39.

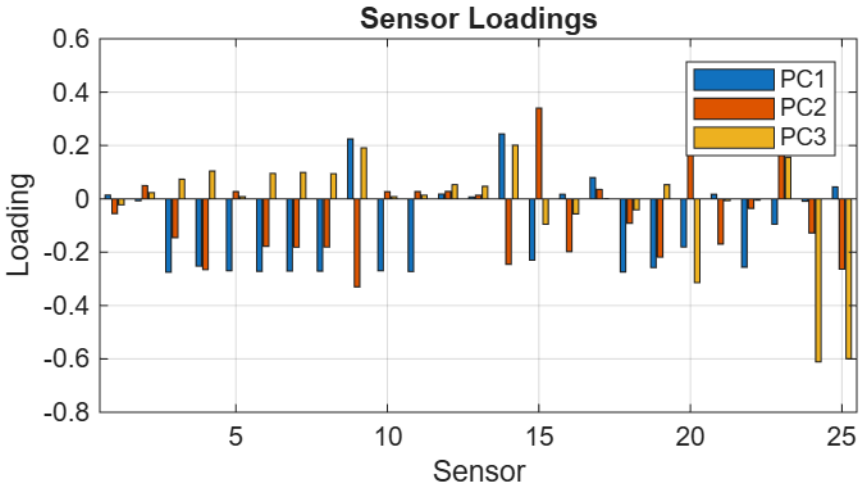

Figure 4. Sensor loadings for the first three components.