Reproducibility Report:
"Wind Turbine Gearbox Failure Detection Through Cumulative Sum of Multivariate Time Series Data"

1. Computer and software environment

Hardware and Operating System:
Any operating system capable of installing R software (Windows/MacOS/Ubuntu etc.) with a recommended RAM of 16 GB or above.

Software used:
R (https://www.R-project.org/)

2. Downloading and creating the dataset files

The following wind farm data are granted free use by CC-BY-SA license. **Note that our reproducibility package does not include the datasets**, as we do not own the data. But the instructions below describe precisely where to download and how to create the dataset files that will be the same as we used for producing the results in the paper.

Dataset	File names	Description
#1	wind-farm-1-signals-training.csv	(1) Download the wind-farm-1-signals-2016.csv and wind-farm-1-signals-2017.csv from EDP Open Data website (the Open Wind Farm portion, and registration is free). (2) Combine the data from the entire wind-farm-1-signals-2016.csv and that from Jan 1 through August 31 of the wind-farm-1-signals-2017.csv, and save the new data file with the name "wind-farm-1-signals-training.csv" (3) This is the recorded SCADA data on wind turbine T01, T06, T07, T09, T11 from January 1, 2016 to August 31, 2017, and will be used as the training set.
#2	wind-farm-1-signals-testing.csv	(1) Download the wind-farm-1-signals-2017.csv from EDP Open Data website. (2) Take the data from Sept 1 through Dec 31 of the wind-farm-1-signals-2017.csv, and save the new data file with the name "wind-farm-1-signals-testing.csv" (3) This is the recorded SCADA data on wind turbine T01, T06, T07, T09, T11 from September 1, 2017 to December 31, 2017, and will be used as the testing set.
#3	htw-failures-2016.csv	Download the htw-failures-2016.csv directly from EDP Open Data website. This is the wind turbine failures recorded in the year of 2016.
#4	htw-failures-2017.csv	Download the htw-failures-2017.csv directly from EDP Open Data website. This is the wind turbine failures recorded in the year of 2017.

3. Explanation of the header of the data files

The original SCADA data have 83 columns (headers). In the following table, only the variables used in the analysis are listed.

Header name	Dataset	Description
Turbine_ID	#1 – #4	Turbine ID
Timestamp	#1 – #4	Time stamp of the data collection (*yyyy-mm-ddThh:mm:ss)*
Gear_Oil_Temp_Avg	#1 & #2	10-minutes average temperature of oil in gearbox (degree Celsius)
Gear_Bear_Temp_Avg	#1 & #2	10-minutes average temperature in gearbox bearing on high speed shaft (degree Celsius)
Nac_Temp_Avg	#1 & #2	10-minutes average temperature of nacelle (degree Celsius)
Rtr_RPM_Avg	#1 & #2	10-minutes average rotor speed (RPM)
Amb_WindDir_Relative_Avg	#1 & #2	10-minutes average wind relative direction (degree)
Grd_Prod_Pwr_Avg	#1 & #2	10-minutes average produced power (kW)
Component	#3 & #4	The component in which a failure is detected
Remarks	#3 & #4	Information about the failure or action taken as a failure response

4. Reproducing the results in the paper

Code File	What does it do?	Required Data or Data File	Output
LoMSTEDP.R	Provides function to perform LoMST	n/a	n/a
CUSUM_LoMST.R	● Performs LoMST. ● Filters the anomaly scores based on the offset. ● Clusters the filtered anomaly scores based on the accumulation window. ● Obtains cumulative scores and save them in *.csv* files.	*LoMSTEDP.R* *Dataset #1 – 4*	Anomaly scores in *LoMST_Gbx.csv* and cumulative scores in: *cusum_T01.csv* *cusum_T06.csv* *cusum_T07.csv* *cusum_T09.csv* *cusum_T11.csv*
Figure2.R	Plots the probability density function (pdf) of the anomaly scores.	*LoMST_Gbx.csv*	Figure 2 in *.pdf*
Figure3.R	Plots the illustration of the method using T09 as an example.	*LoMST_Gbx.csv* *cusum_T09.csv*	Figure 3 in *.pdf*
Figure4.R	Plots the results of the method on the five wind turbines.	*cusum_T01.csv* *cusum_T06.csv* *cusum_T07.csv* *cusum_T09.csv* *cusum_T11.csv*	Figure 4 in *.pdf*
Comparison.xlsx	Provides a performance summary of other methods in the comparison study.	*n/a*	Table 3 and Table 4.
Figure5.R	Plots a small section of CUSUM method from (Dao, 2021).	*Dataset #1*	Figure 5 in *.pdf*
Parameter_setting.xlsx	This is an intermediate tool for determining the threshold to be used in the CUSUM-LoMST method. One can produce the plots of cumulative scores of the five wind turbines with a choice of threshold values and tune it based on the resulting savings/cost.	*n/a*	The estimated total saving/cost from the chosen parameter setting.