Supplementary Material for *"A Nonlinear Support Vector Machine based Feature Selection Approach for Fault Detection and Diagnosis: Application to the Tennessee Eastman Process"*

Melis Onel[1,2], Chris A. Kieslich[1,2,3], Efstratios N. Pistikopoulos[1,2,*]

1. Artie McFerrin Department of Chemical Engineering,

Texas A&M University, College Station, TX 77843, USA

2. Texas A&M Energy Institute,

Texas A&M University, College Station, TX 77843, USA

3. Coulter Department of Biomedical Engineering,

Georgia Institute of Technology, Atlanta, GA

* *To whom correspondence should be addressed.* Email:stratos@tamu.edu

1 Tennessee Eastman Process Reactions & Process Variables

The reactions occurring in the reactor are as follows:

$$\begin{aligned} A_{(g)} + C_{(g)} + D_{(g)} &\rightarrow G_{(l)} \\ A_{(g)} + C_{(g)} + E_{(g)} &\rightarrow H_{(l)} \\ A_{(g)} + E_{(g)} &\rightarrow F_{(l)} \\ 3D_{(g)} &\rightarrow 2F_{(l)} \end{aligned} \tag{1}$$

The process contains 11 manipulated (Table S2) and 41 measured (Table S1) variables.

Table S1: Measured variables in the Tennessee Eastman process.

Variable No	Description	Measurement Type
1	Feed A (Stream 1)	Process
2	Feed D (Stream 2)	Process
3	Feed E (Stream 3)	Process
4	Total Feed (Stream 4)	Process
5	Recycle Flow (Stream 8)	Process
6	Reactor Feed Rate (Stream 6)	Process
7	Reactor Pressure	Process
8	Reactor Level	Process
9	Reactor Temperature	Process
10	Purge Rate (Stream 9)	Process
11	Product Separator Temperature	Process
12	Product Separator Level	Process
13	Product Separator Pressure	Process
14	Product Separator Underflow	Process
15	Stripper Level	Process
16	Stripper Pressure	Process
17	Stripper Underflow (Stream 11)	Process
18	Stripper Temperature	Process
19	Stripper Steam Flow	Process
20	Compressor Work	Process
21	Reactor Cooling Water Outlet Temperature	Process
22	Separator Cooling Water Outlet Temperature	Process
23	Component A (Stream 6)	Composition
24	Component B (Stream 6)	Composition
25	Component C (Stream 6)	Composition
26	Component D (Stream 6)	Composition
27	Component E (Stream 6)	Composition
28	Component F (Stream 6)	Composition
29	Component A (Stream 9)	Composition
30	Component B (Stream 9)	Composition
31	Component C (Stream 9)	Composition
32	Component D (Stream 9)	Composition
33	Component E (Stream 9)	Composition
34	Component F (Stream 9)	Composition
35	Component G (Stream 9)	Composition
36	Component H (Stream 9)	Composition
37	Component D (Stream 11)	Composition
38	Component E (Stream 11)	Composition
39	Component F (Stream 11)	Composition
40	Component G (Stream 11)	Composition
41	Component H (Stream 11)	Composition

Table S2: Manipulated variables in the Tennessee Eastman process.

Variable No	Description
42	D Feed Flow (Stream 2)
43	E Feed Flow (Stream 3)
44	A Feed Flow (Stream 1)
45	Total Feed Flow (Stream 4)
46	Compressor Recycle Valve
47	Purge Valve (Stream 9)
48	Separator Pot Liquid Flow (Stream 10)
49	Stripper Liquid Product Flow
50	Stripper Steam Valve
51	Reactor Cooling Water Flow
52	Condenser Cooling Water Flow

2 Performance Metric Terminology & Formulations

In this study, the model performances are assessed with 5 metrics: (i) area under the curve (AUC), (ii) fault detection rate, or recall, (iii) accuracy, (iv) false alarm rate, and (v) false negative rate. These metrics are derivations achieved from the confusion (a.k.a error) matrix, which is a two-dimensional contingency table used to evaluate performance of a classifier model in statistics and machine learning. Below, we provide terminology and formulation of the 5 metrics adopted in this study.

Confusion matrix is a two-dimensional matrix containing the number of correct and false classification of instances for a binary classification problem. The elements of the matrix are True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In this study, these numbers indicate:

True Positives (TP): Predicting faulty operation as faulty.

True Negatives (TN): Predicting fault-free (normal) operation as normal.

False Positives (FP): Predicting normal operation as faulty.

False Negatives (FN): Predicting faulty operation as normal.

Accordingly,

Positive (P): $TP+FN$

Negative (N): TN+FP

Accuracy is the percentage of correctly classified instances among all instances, which is calculated as follows:

$$Accuracy = \frac{TP+TN}{P+N}.$$

Accuracy is the most effective metric in the cases where class distribution is somewhat balanced. Recall (also referred as detection rate) is a measure of completeness.

$$Recall = \frac{TP}{TP+FN}.$$

In this work, misclassification of the instances are assessed via false alarm rate (FAR) and false negative rate (FNR):

$$FAR = \frac{FP}{FP+TN}.$$

$$FNR = \frac{FN}{FN+TP}.$$

In the case of imbalanced data sets, collective evaluation of two metrics, namely recall and specificity, gains importance where

$$Specificity = \frac{TN}{TN+FP}.$$

Receiver Operating Characteristics (ROC) curve is the plot illustrating the classifier performance based on recall and specificity at varying classification thresholds. Particularly, the curve demonstrates true positive rate (recall) versus false positive rate (1-$Specificity$) for all possible classification thresholds and area under this curve (a.k.a Area Under the Curve (AUC)) is a single metric derived from this advanced performance assessment.

3 *C*-SVM Model Parameters

Table S3: *C*-SVM hyperparameters for the models reported in Table 2.

Fault	**Optimal Feature Subset Size**	$\hat{C}$	$\hat{\gamma}$
1	2	6	0
2	5	2	-4
3	13	0	3
4	1	-8	-1
5	4	10	-6
6	2	-9	0
7	3	-8	0
8	7	3	4
9	17	2	2
10	14	6	4
11	29	5	3
12	9	10	7
13	7	10	3
14	3	-8	3
15	27	1	2
16	5	4	2
17	32	10	-5
18	34	10	-4
19	2	9	10
20	14	8	3
21	1	0	9

Table S4: *C*-SVM hyperparameters for the models reported in Table 3.

Fault	**Optimal Feature Subset Size**	$\hat{C}$	$\hat{\gamma}$
8	4	5	4
17	27	10	-5

Table S5: C-SVM hyperparameters for the models reported in Table 4.

Fault	**Optimal Feature Subset Size**	$\hat{C}$	$\hat{\gamma}$
1	18	3	0
2	10	9	-3
3	10	9	7
4	1	-8	1
5	14	0	-1
6	2	-10	0
7	4	-1	5
8	12	4	1
9	2	1	4
10	15	10	-2
11	2	9	-2
12	5	4	4
13	8	5	1
14	2	-5	8
15	16	0	5
16	2	2	3
17	28	10	-5
18	5	9	0
19	10	10	-3
20	14	10	-3

Table S6: C-SVM hyperparameters for the models reported in Table 5.

Fault	**Optimal Feature Subset Size**	$\hat{C}$	$\hat{\gamma}$
5	3	9	-7
18	2	9	1
19	3	10	-2
20	13	10	-3

4 Diagnosis of All Faults

Table S7: Diagnosis from the Table 2 end-models developed with *Chiang et. al* dataset. Faults 3, 9, and 15 are excluded due to poor model performance.

Fault	Optimal Feature Subset Size	Selected Process Variables
1	2	16, 44
2	5	7, 16, 10, 47, 13
3	13	24, 28, 29, 26, 33, 31, 50, 25, 30, 27, 35, 18, 1
4	1	51
5	4	4, 11, 52, 17
6	2	44, 1
7	3	45, 7, 13
8	7	39, 44, 16, 20, 7, 23, 46
9	17	39, 41, 38, 40, 37, 50, 18, 19, 7, 13, 16, 20, 31, 33, 29, 35, 28
10	14	41, 39, 38, 37, 40, 50, 19, 18, 20, 7, 13, 16, 31, 29
11	29	24, 29, 26, 30, 31, 32, 35, 50, 25, 33, 34, 23, 27, 18, 7, 16, 20, 38, 10, 19, 13, 37, 39, 44, 1, 41, 51, 47, 9
12	9	7, 16, 50, 18, 13, 19, 20, 38, 33
13	7	39, 40, 18, 7, 38, 23, 3
14	3	9, 51, 21
15	27	39, 41, 37, 40, 38, 50, 18, 19, 20, 7, 13, 16, 1, 44, 25, 31, 23, 33, 29, 36, 35, 34, 24, 30, 27, 47, 10
16	5	50, 19, 18, 20, 13
17	32	38, 39, 40, 41, 21, 37, 19, 20, 33, 27, 34, 30, 1, 11, 25, 28, 24, 23, 35, 36, 26, 10, 3, 2, 22, 14, 48, 47, 32, 42, 8, 49
18	34	39, 40, 37, 41, 14, 49, 17, 48, 5, 52, 15, 12, 3, 9, 2, 32, 10, 26, 28, 24, 6, 11, 36, 20, 50, 22, 44, 13, 34, 1, 7, 8, 25, 18
19	2	32, 31
20	14	41, 39, 40, 37, 50, 18, 46, 13, 19, 7, 16, 11, 33, 27
21	1	45

Table S8: Diagnosis from the Table 4 end-models developed with *Rieth et. al* dataset. Faults 3, 9, and 15 are excluded due to poor model performance.

Fault	**Optimal Feature Subset Size**	**Selected Process Variables**
1	18	44, 16, 41, 4, 1, 11, 18, 21, 22, 20, 7, 51, 46, 38, 33, 13, 23, 24
2	10	38, 41, 10, 25, 16, 21, 7, 20, 47, 30
3	10	47, 51, 38, 16, 50, 18, 19, 21, 20, 13
4	1	51
5	14	52, 11, 17, 4, 18, 19, 46, 50, 20, 16, 44, 38, 29, 22
6	2	44, 1
7	4	19, 18, 50, 45
8	12	39, 41, 37, 16, 20, 44, 7, 46, 1, 27, 29, 40
9	2	51, 13
10	15	41, 38, 39, 37, 18, 19, 40, 25, 31, 29, 26, 23, 1, 50, 16
11	2	9, 51
12	5	16, 38, 35, 25, 11
13	8	41, 39, 7, 37, 40, 16, 32, 21
14	2	51, 9
15	16	22, 7, 13, 18, 50, 19, 11, 16, 38, 35, 20, 9, 21, 46, 4, 29
16	2	19, 50
17	28	35, 24, 38, 28, 18, 20, 19, 21, 46, 26, 36, 42, 37, 25, 29, 30, 39, 41, 40, 44, 32, 34, 22, 8, 10, 27, 31, 23
18	5	22, 8, 20, 11, 31
19	10	13, 16, 46, 50, 19, 5, 7, 20, 38, 6
20	14	38, 39, 41, 16, 52, 17, 18, 30, 35, 29, 40, 13, 7, 47