Fachhochschule Südwestfalen
University of Applied Sciences

Systems Engineering

Comparison of Data-driven Fault Detection and Identification Methods for the Tennessee Eastman Process

5 February 2018

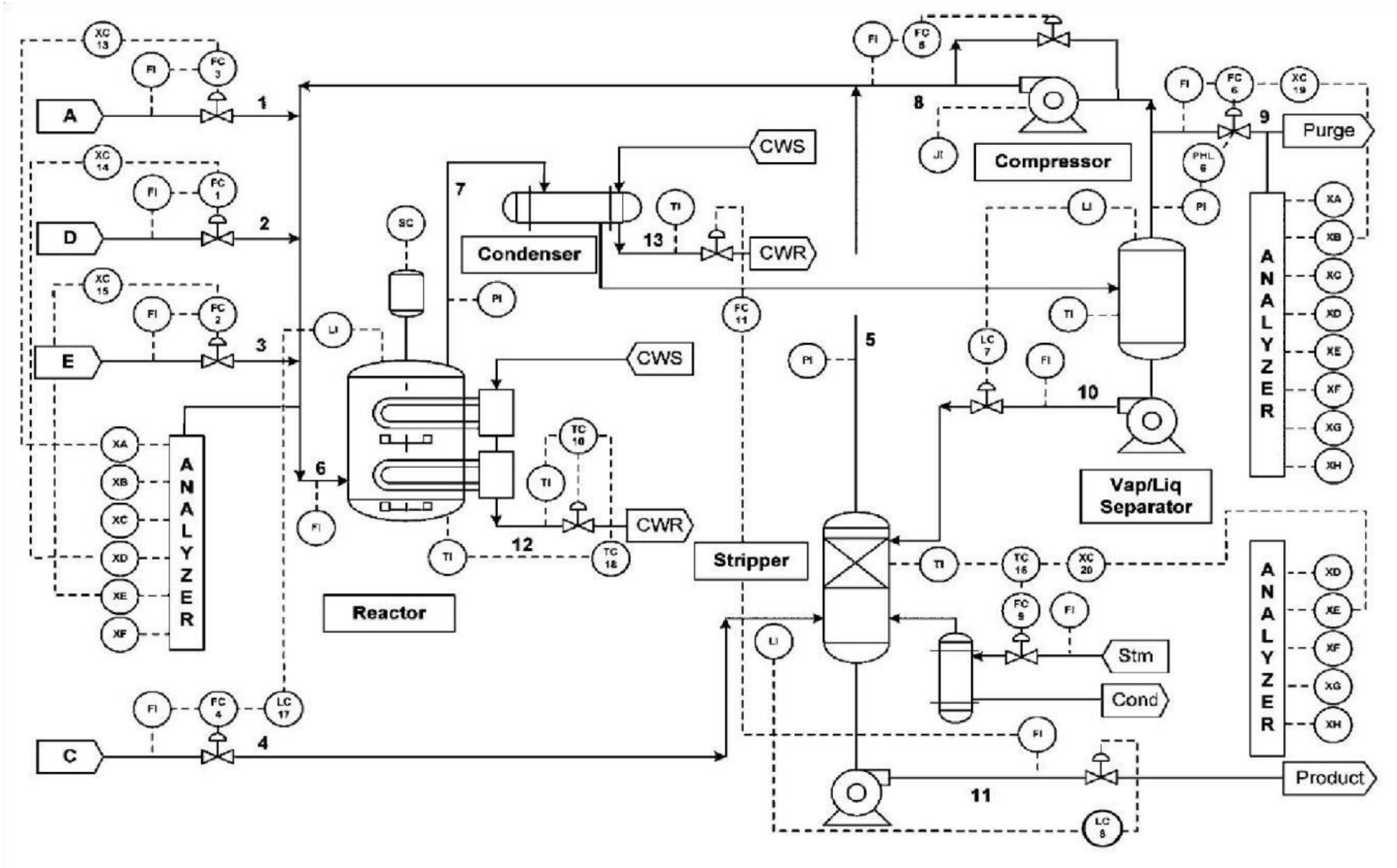
Fault Detection & Identification

- Faults within process lead to deviation from normal behavior of plant and lower the performance of overall system
- Fault Detection:
  - Discovers abnormalities in process
- Data Driven FD:
  - Use of Data analysis tools on evaluation of production data for process monitoring

Fault Detection & Identification

- Benefits of Fault Detection and Identification
    - Prediction about expected maintenance
    - Improves plant efficiency
    - Avoids unnecessary shutdowns
    - Ensures safety and reliability of process

Tennessee Eastman Process (TEP)



Tennessee Eastman Process (TEP)

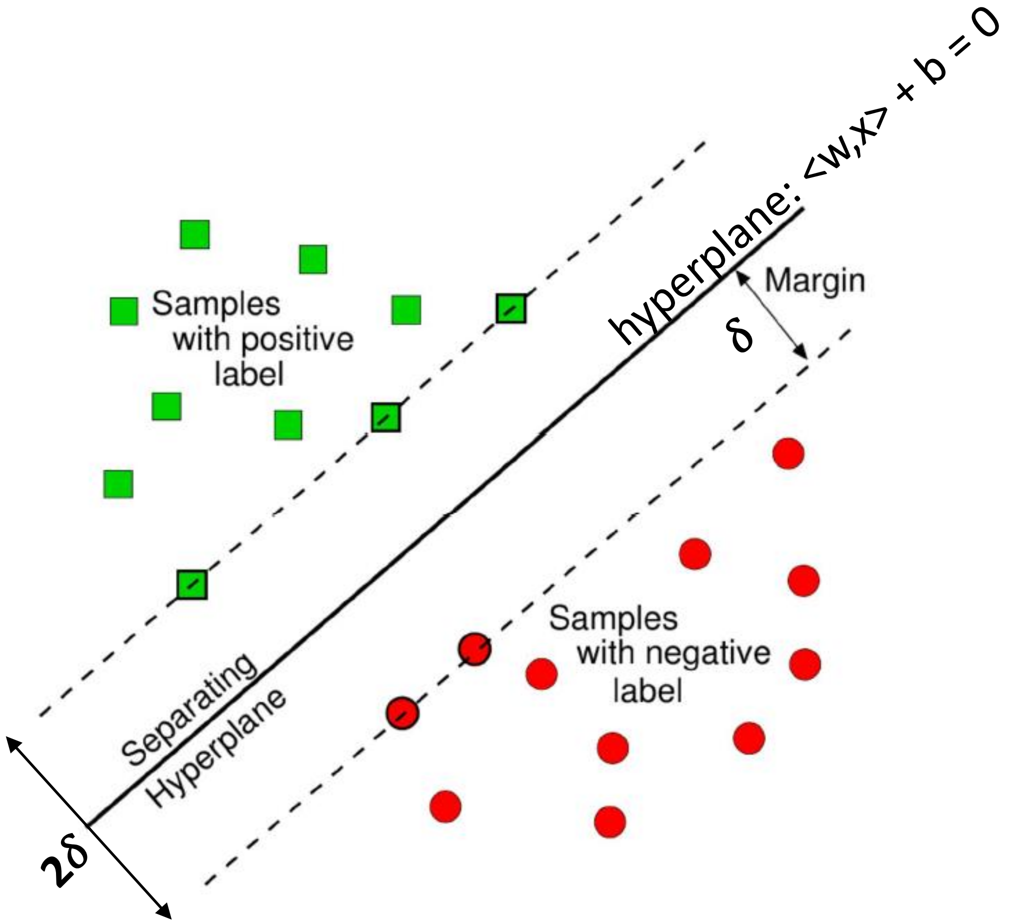
- Major Unit Operations
  1. Reactor
  2. Product Condenser
  3. Vapor Liquid Separator
  4. Recycle Compressor
  5. Product Stripper

Tennessee Eastman Process (TEP)

- 21 Pre-programmed Faults
- 52 Variables
  - 22 Process Measurement Variables, 19 Composition Measurement Variables, 11 Manipulated Variables
- Generate simulated data at an interval of 3 minutes
- Normal-Train Data: 500 Samples
- Normal-Test Data: 960 Samples
- Fault-Train Data: 480 Samples
- Fault-Test Data: 960 Samples
  - The first 160 samples of fault-test data are normal conditions

SVM Classifier

- Finding the best decision boundary that will help separate the classes
- Optimum decision boundary has maximum distance from closest points in each class i.e highest 2δ
- The closest points are known as support vectors



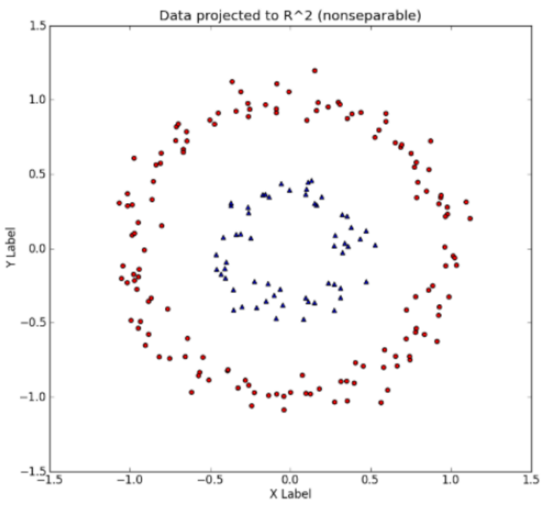
SVM – Kernel

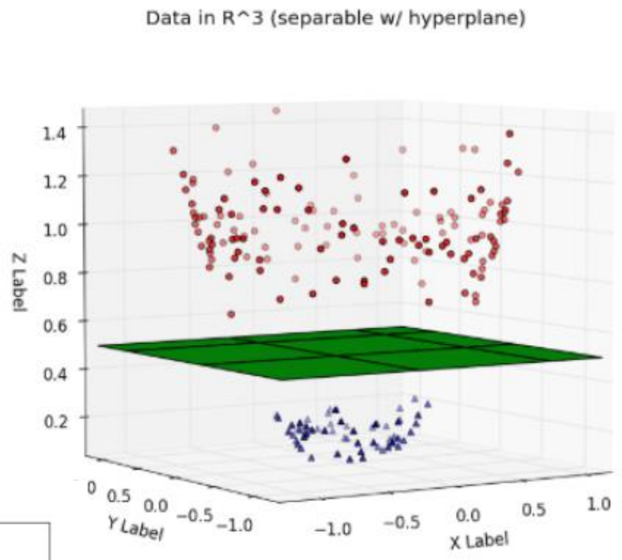
- If data is not linearly separable then linear classifier is not suitable

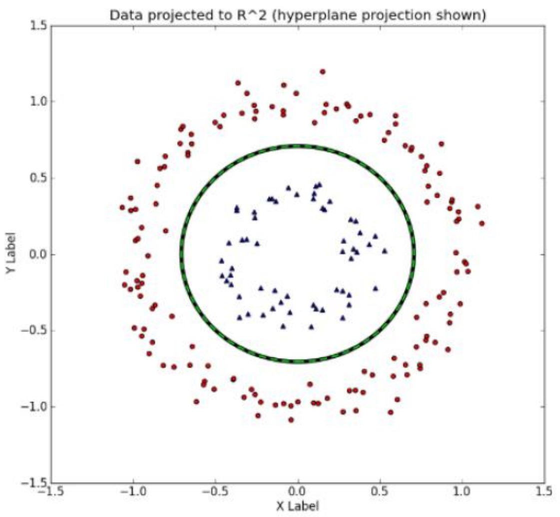
- Project the data to higher dimension where it can be linearly classified

- Kernel function allows computation in more efficient way

Linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
Rbf	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2),\ \gamma > 0.$
Polynomial	$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d,\ \gamma > 0$







SVM Results for Fault Cases

Fault 4: Reactor cooling water inlet temperature

	0	1
0	1054	66
1	21	779

C = 1
Kernel = rbf
Gamma = 0.05
Accuracy on test set = 95.46%
k-fold Mean accuracy = 94.89%

Fault 12: Condenser cooling water inlet temperature

	0	1
0	1017	103
1	302	498

C = 10
Kernel = linear
Accuracy on test set = 78.90%
k-fold Mean accuracy = 79.08%

Faults : 1,4,5,7,11

	0	1	2	3	4	5
0	1596	0	0	0	0	164
1	7	793	0	0	0	0
2	0	0	798	0	0	2
3	1	0	0	798	0	1
4	0	0	0	0	800	0
5	270	0	184	0	0	346

C = 1

Kernel = rbf

Gamma = 0.02

Accuracy on test set = 89.07%

k-fold Mean accuracy = 85.72%

Fault	C	Kernel	Degree	Gamma	Accuracy
IDV(1)	1	Linear	-	-	97.55
	1	RBF	-	0.05	99.08
	1	Poly	2	0.03	93.77
IDV(2)	1	Linear	-	-	97.14
	1	RBF	-	0.009	97.44
	10	Poly	3	0.07	94.79
IDV(3)	1	Linear	-	-	49.79
	1	RBF	-	0.05	49.38
	1	Poly	3	0.05	53.16
IDV(4)	10	Linear	-	-	100
	1	RBF	-	0.05	94.89
	10	Poly	4	0.05	42.24
IDV(5)	10	Linear	-	-	97.44
	10	RBF	-	0.05	98.87
	100	Poly	4	0.05	63.46
IDV(6)	10	Linear	-	-	97.44

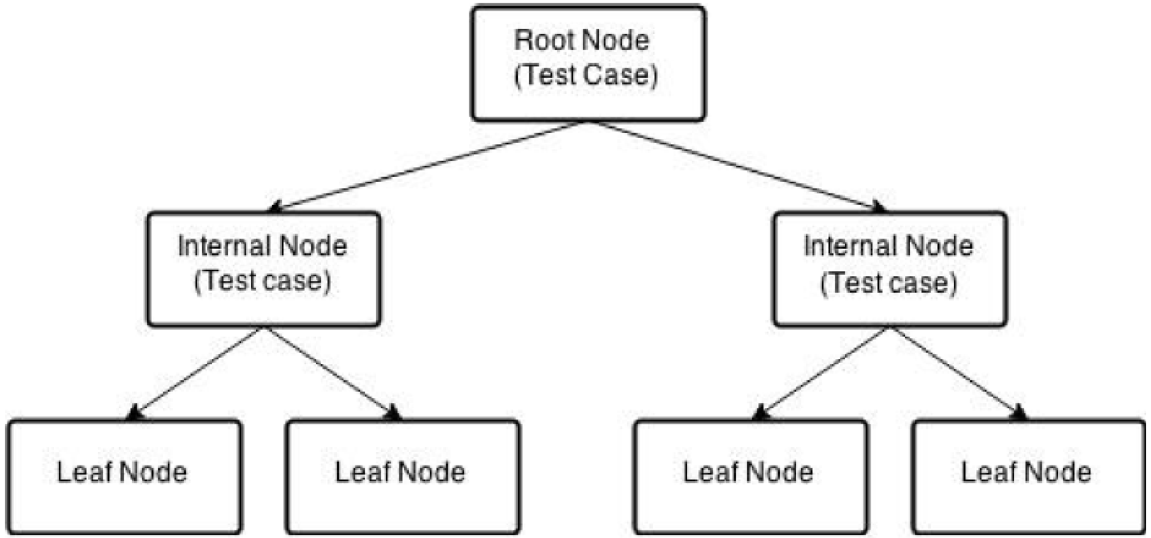
	1	RBF	-	0.08	94.69
	1	Poly	3	0.03	98.46
IDV(7)	1	Linear	-	-	100
	1	RBF	-	0.009	100
	10	Poly	3	0.02	93.67
IDV(8)	10	Linear	-	-	72.65
	1	RBF	-	0.06	96.83
	1	Poly	4	0.07	85.30
IDV(9)	10	Linear	-	-	50.61
	10	RBF	-	0.08	60.10
	1	Poly	4	0.06	60.40
IDV(10)	1	Linear	-	-	53.16
	1	RBF	-	0.04	77.50
	10	Poly	3	0.05	55.71
IDV(11)	10	Linear	-	-	63.06
	1	RBF	-	0.08	66.22
	1	Poly	4	0.01	52.34
IDV(12)	10	Linear	-	-	79.08
	1	RBF	-	0.02	95.91
	100	Poly	4	0.05	88.06

IDV(13)	1	Linear	-	-	86.83
	1	RBF	-	0.08	97.04
	1	Poly	3	0.05	71.83
IDV(14)	100	Linear	-	-	92.24
	10	RBF	-	0.01	89.48
	10	Poly	3	0.04	66.42
IDV(15)	10	Linear	-	-	58.87
	100	RBF	-	0.05	57.24
	10	Poly	4	0.06	49.59
IDV(16)	1	Linear	-	-	61.73
	1	RBF	-	0.02	69.89
	1	Poly	4	0.09	62.75
IDV(17)	10	Linear	-	-	91.22
	1	RBF	-	0.04	73.87
	1	Poly	3	0.01	65.91
IDV(18)	1	Linear	-	-	90.00
	10	RBF	-	0.02	88.57
	10	Poly	3	0.03	80.40
IDV(19)	10	Linear	-	-	40.61

	100	RBF	-	0.01	64.69
	10	Poly	3	0.08	63.57
IDV(20)	100	Linear	-	-	87.24
	10	RBF	-	0.02	82.55
	100	Poly	4	0.09	69.79
IDV(21)	10	Linear	-	-	69.18
	10	RBF	-	0.009	72.04
	1	Poly	3	0.06	63.87
IDV(1), IDV(2), IDV(4)	1	Linear	-	-	97.37
	1	RBF	-	0.02	98.24
	1	Poly	3	0.02	95.00
IDV(1), IDV(4), IDV(5), IDV(7), IDV(11)	1	Linear	-	-	86.82
	1	RBF	-	0.02	85.72
	1	Poly	3	0.02	64.11

Decision Trees Classifier

- The model uses if-then-else rules to interpret the value of target variable
- Most important attribute is placed at root node
- Data is split into subsets to get homogeneous sample
- Final data attribute which cannot be classified further is known as leaf



Decision Trees – Splitting, Pruning

- Splitting is done based on-
  - Information Gain: Difference in entropies
  - Gini Index: Quantifying amount of uncertainty (impurity) in data
- Prone to overfitting due to excessive splitting of data
- Pruning makes model simpler, generalized and avoids overfitting
- Pruning done by reducing depth of tree by replacing bad branches (subtrees) by leaves

Decision Trees Results for Fault Cases

Fault 5: Condenser cooling water inlet temperature

	0	1
0	760	360
1	31	769

Criterion = entropy
Max_depth = 3
Min_samples_leaf = 110
Accuracy on test set = 79.63%
k-fold Mean accuracy = 81.42%

Fault 19: Unknown

	0	1
0	794	326
1	104	696

Criterion = gini
Max_depth = 4
Min_samples_leaf = 70
Accuracy on test set = 77.60%
k-fold Mean accuracy = 80.20%

Faults : 0,1,4,5,7,11

	0	1	2	3	4	5
0	1368	0	0	221	1	170
1	10	784	0	0	0	6
2	1	0	798	1	0	0
3	176	0	0	597	3	24
4	0	0	0	0	800	0
5	194	0	250	69	0	287

Criterion = gini
Max_depth = 8
Min_samples_leaf = 280
Accuracy on test set = 80.45%
k-fold Mean accuracy = 79.10%

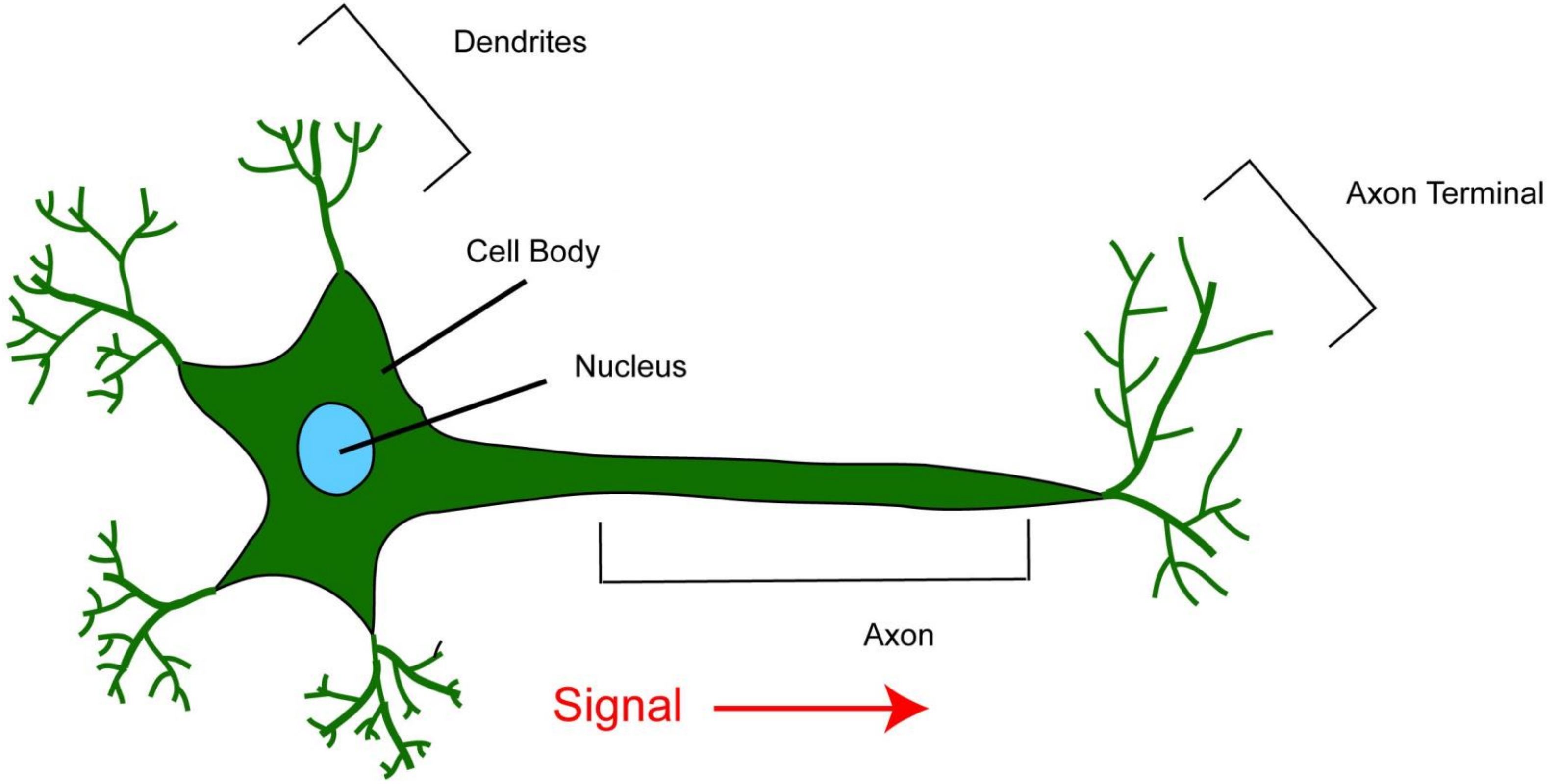
Fault	criterion	max_depth	min_samples_leaf	Accuracy
IDV(1)	gini	2	70	95.10
	entropy	10	100	95.10
IDV(2)	gini	2	70	96.02
	entropy	5	90	96.02
IDV(3)	gini	5	100	42.65
	entropy	6	90	43.16
IDV(4)	gini	2	70	100
	entropy	4	80	100
IDV(5)	gini	4	70	94.89
	entropy	3	110	81.42
IDV(6)	gini	2	70	95.61
	entropy	7	90	95.61
IDV(7)	gini	2	70	99.79
	entropy	2	90	99.79
IDV(8)	gini	3	90	85.30
	entropy	4	70	86.22
IDV(9)	gini	7	100	52.04
	entropy	2	70	57.44

IDV(10)	gini	6	80	72.55
	entropy	2	70	77.95
IDV(11)	gini	5	90	88.87
	entropy	2	70	89.28
IDV(12)	gini	2	110	87.85
	entropy	7	110	87.85
IDV(13)	gini	2	130	89.08
	entropy	4	100	85.81
IDV(14)	gini	4	80	94.69
	entropy	2	70	94.89
IDV(15)	gini	3	70	46.12
	entropy	4	120	55.61
IDV(16)	gini	4	70	80.00
	entropy	5	110	64.28
IDV(17)	gini	2	70	92.44
	entropy	3	90	91.53

IDV(18)	gini	2	110	86.42
	entropy	6	100	76.53
IDV(19)	gini	4	70	80.20
	entropy	4	100	74.69
IDV(20)	gini	3	80	75.30
	entropy	2	110	84.38
IDV(21)	gini	2	70	98.97
	entropy	5	90	98.97
IDV(1), IDV(2), IDV(4)	gini	6	190	95.25
	entropy	5	190	95.25
IDV(1), IDV(4), IDV(5), IDV(7), IDV(11)	gini	8	280	79.10
	entropy	7	280	79.13

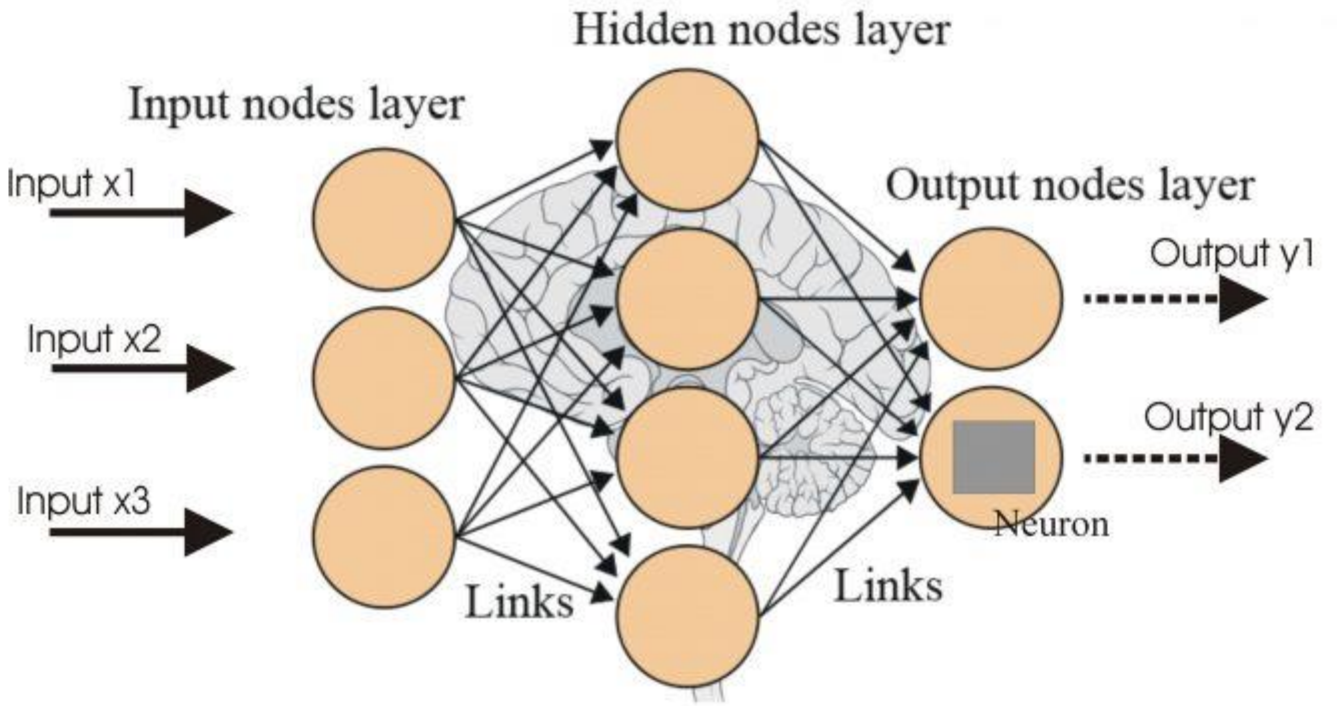
Artificial Neural Networks Classifier

- Human Brain



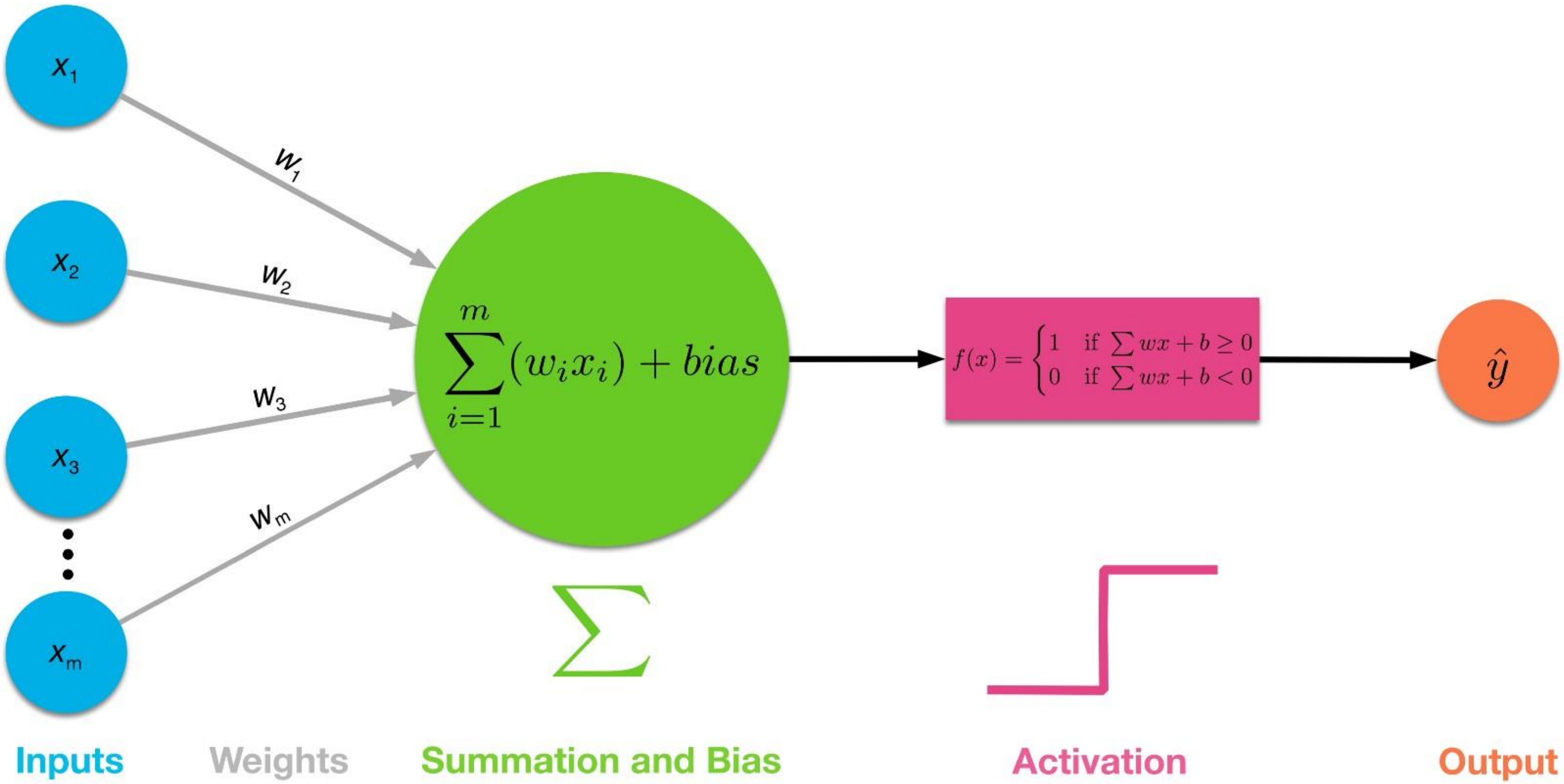
Neural Networks

- Artificial Neural Network



Neural Networks

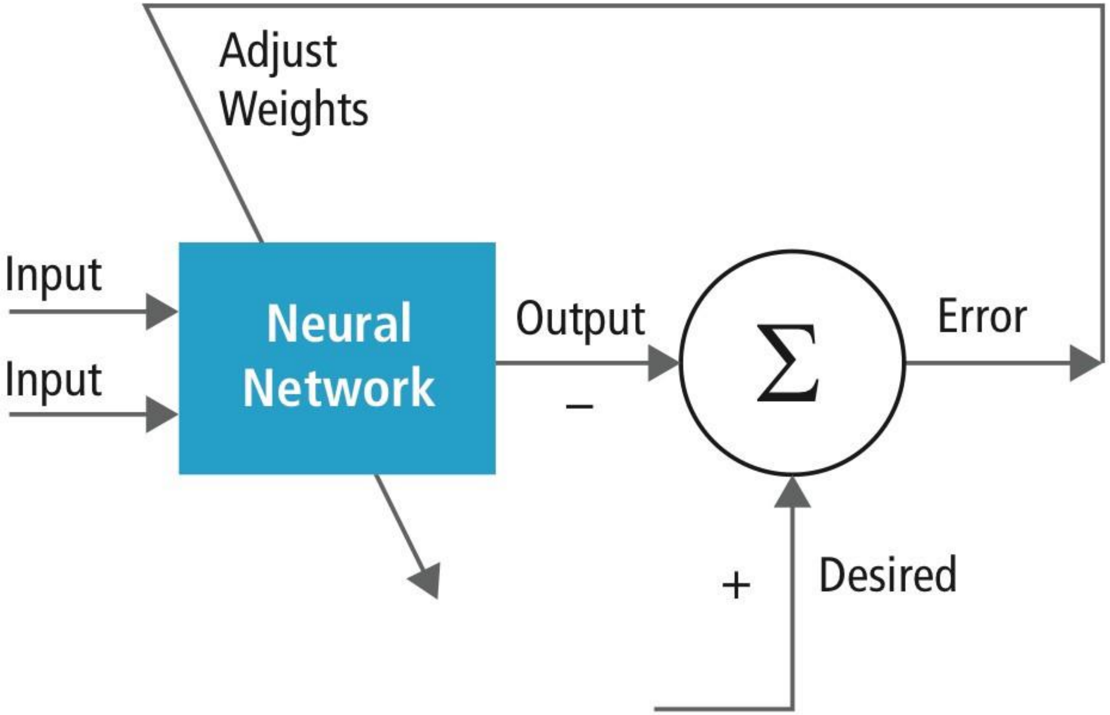
- Basic Working



Neural Networks

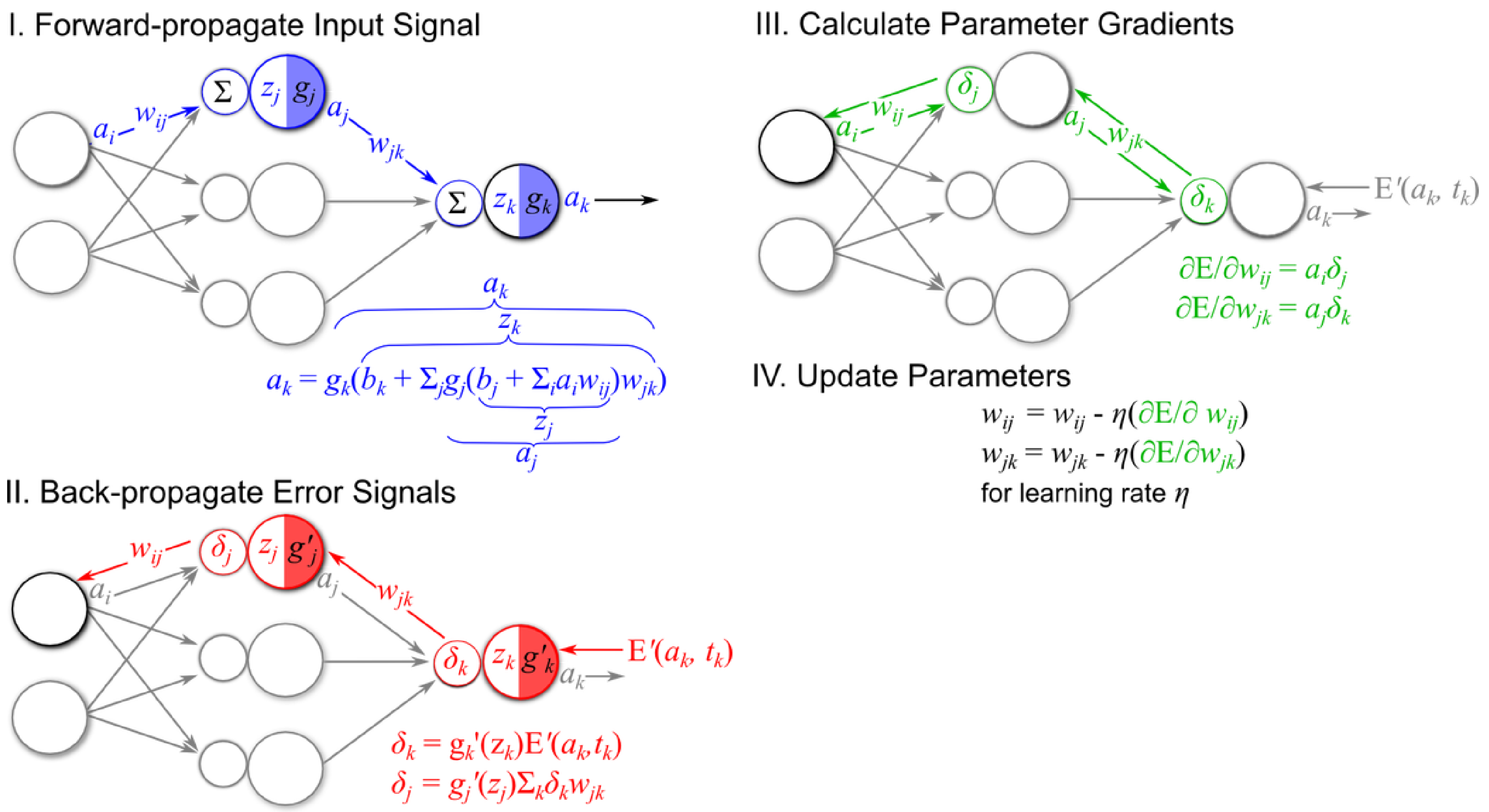
- Learning: W and θ are adjusted, to minimize the error

$$\varepsilon = \sum_{i=1}^{N} (d_i - f_{\mathbf{w}}(\mathbf{x}_i))^2 = \sum_{i=1}^{N} (d_i - f_i)^2 \overset{!}{=} min$$



Neural Networks

- Backpropagation



ANN Results for Fault Cases

Fault 1,4,5,7,11 (Combined) (for Batch Size 10)

1500	6	8	2	0	244
2	796	0	0	0	2
0	0	715	0	0	85
0	1	0	799	0	0
0	0	0	0	799	1
179	1	144	1	0	475

accuracy	float64	1	0.8895833333333328

ANN Results for Fault Cases

Fault 1, 2, 4 (Combined) (for Batch Size 10)

1420	4	15	1
2	798	0	0
14	0	786	0
0	0	0	800

accuracy	float64	1	0.9906249999999998

ANN Results for Fault Cases

Fault No.	Accuracy
IDV 1	99.89
IDV 2	96.97
IDV 3	57.08
IDV 4	100
IDV 5	99.94
IDV 6	99.79
IDV 7	100
IDV 8	88.12
IDV 9	52.39
IDV 10	68.64
IDV 11	71.35

Fault No.	Accuracy
IDV 12	93.75
IDV 13	79.58
IDV 14	99.89
IDV 15	48.38
IDV 16	76.87
IDV 17	88.12
IDV 18	92.50
IDV 19	90.20
IDV 20	83.07
IDV 21	64.58
IDV 1, 4, 5, 7, 11	88.26
IDV 1, 2, 4	99.06

Detailed Comparison

Faults	Accuracy			Best Classifier
	SVM	Decision Trees	DNN	
IDV 16	69.89	80.00	76.87	DT
IDV 17	91.22	92.44	88.12	DT
IDV 18	90.00	86.42	93.48	DNN
IDV 19	64.69	80.20	90.20	DNN
IDV 20	87.24	84.38	83.07	SVM
IDV 21	72.04	98.97	64.58	DT
IDV 1, 2, 4	98.24	95.25	99.29	DNN
IDV 1, 4, 5, 7, 11	86.82	79.13	91.97	DNN
IDV 10	77.5	77.95	69.73	DT
IDV 11	66.22	89.28	71.35	DT
IDV 12	95.91	87.85	94.58	SVM
IDV 13	97.04	89.08	79.58	SVM
IDV 14	92.24	94.89	99.89	DNN
IDV 15	58.87	55.61	48.95	SVM

In General:

SVM	Decision Trees	Deep Neural Network
• Not biased by outliers • Kernel SVM has high performance on non-linear problem	• No need of feature scaling • Works on both - linear and non-linear cases	• Different algorithms can be used for training
• Not the best choice for large number of features	• Poor results on too small dataset • Prone to Overfitting	• Require large dataset and computational power

Conclusion

- The performance of all three classifiers is above the baseline accuracy in all fault cases except 3, 9, 11 and 15.
- The accuracy of classifiers decreases with increase in number of class labels.
- On comparing the performance, it can be concluded that Deep Neural Network model gives the best accuracy in most cases followed by the SVM and Decision Tree respectively.