Machine Learning pour la cybersécurité

Leclerc_SXXI

Vlad Argatu

Raphaël Mourot-Pelade

Maxence Oden

Gilles Recouvreux

Bastien Pouëssel

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Table des matières

Présentation du jeu de données

- Hardware in the loop (HIL)
  - Network Dataset
  - Physical Dataset

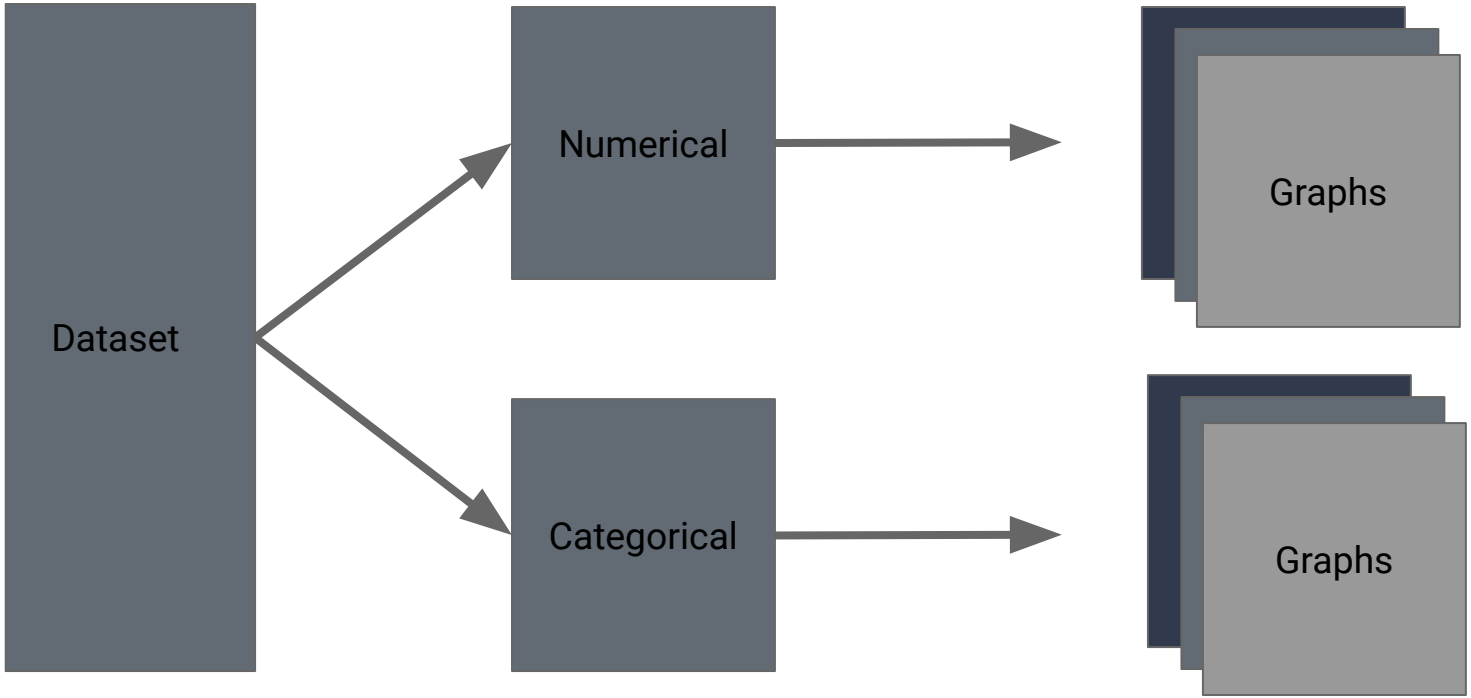
№	*Features*	*Description*
1	Time	Date of acquisition
2	Src IP address	Source IP address
3	Dst IP address	Destination IP address
4	Src MAC address	Source MAC address
5	Dst MAC address	Destination MAC address
6	Src Port	Source port
7	Dst port	Destination port
8	Proto	Protocol
9	TCP flags	CWR | ECN | URG | ACK | PSH | RST | SYN | FIN flags
10	Payload size	Size of packet payload
11	MODBUS code	MODBUS function code
12	MODBUS value	MODBUS response value
13	num_pkts_src	Number of packets of the same source address in the last 2 seconds
14	num_pkts_dst	Number of packets of the same destination address in the last 2 seconds

Network Dataset

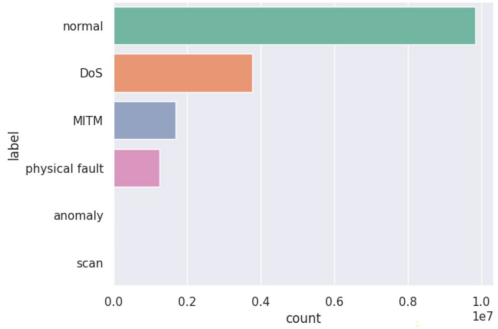
№	*Features*	*Description*	№	*Features*	*Description*
1	Time	Datetime of acquisition	*22*	Valv_3	State of solenoid valve 3
2	Tank_1	Pressure sensor value of tank 1	*23*	Valv_4	State of solenoid valve 4
3	Tank_2	Pressure sensor value of tank 2	*24*	Valv_5	State of solenoid valve 5
4	Tank_3	Pressure sensor value of tank 3	*25*	Valv_6	State of solenoid valve 6
5	Tank_4	Pressure sensor value of tank 4	*26*	Valv_7	State of solenoid valve 7
6	Tank_5	Pressure sensor value of tank 5	*27*	Valv_8	State of solenoid valve 8
7	Tank_6	Pressure sensor value of tank 6	*28*	Valv_9	State of solenoid valve 9
8	Tank_7	Pressure sensor value of tank 7	*29*	Valv_10	State of solenoid valve 10
9	Tank_8	Pressure sensor value of tank 8	*30*	Valv_11	State of solenoid valve 11
10	Pump_1	State of pump 1	*31*	Valv_12	State of solenoid valve 12
11	Pump_2	State of pump 2	*32*	Valv_13	State of solenoid valve 13
12	Pump_3	State of pump 3	*33*	Valv_14	State of solenoid valve 14
13	Pump_4	State of pump 4	*34*	Valv_15	State of solenoid valve 15
14	Pump_5	State of pump 5	*35*	Valv_16	State of solenoid valve 16
15	Pump_6	State of pump 6	*36*	Valv_17	State of solenoid valve 17
16	Flow_sensor_1	Flow sensor value 1	*37*	Valv_18	State of solenoid valve 18
17	Flow_sensor_2	Flow sensor value 2	*38*	Valv_19	State of solenoid valve 19
18	Flow_sensor_3	Flow sensor value 3	*39*	Valv_20	State of solenoid valve 20
19	Flow_sensor_4	Flow sensor value 4	*40*	Valv_21	State of solenoid valve 21
20	Valv_1	State of solenoid valve 1	*41*	Valv_22	State of solenoid valve 22
21	Valv_2	State of solenoid valve 2			

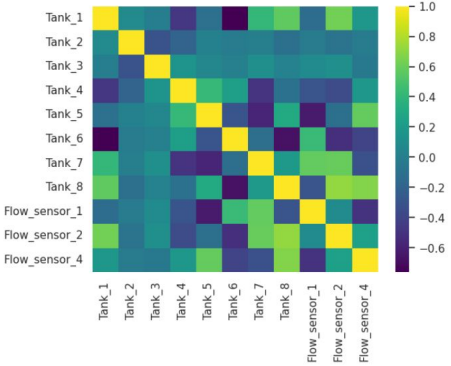
Physical Dataset Features

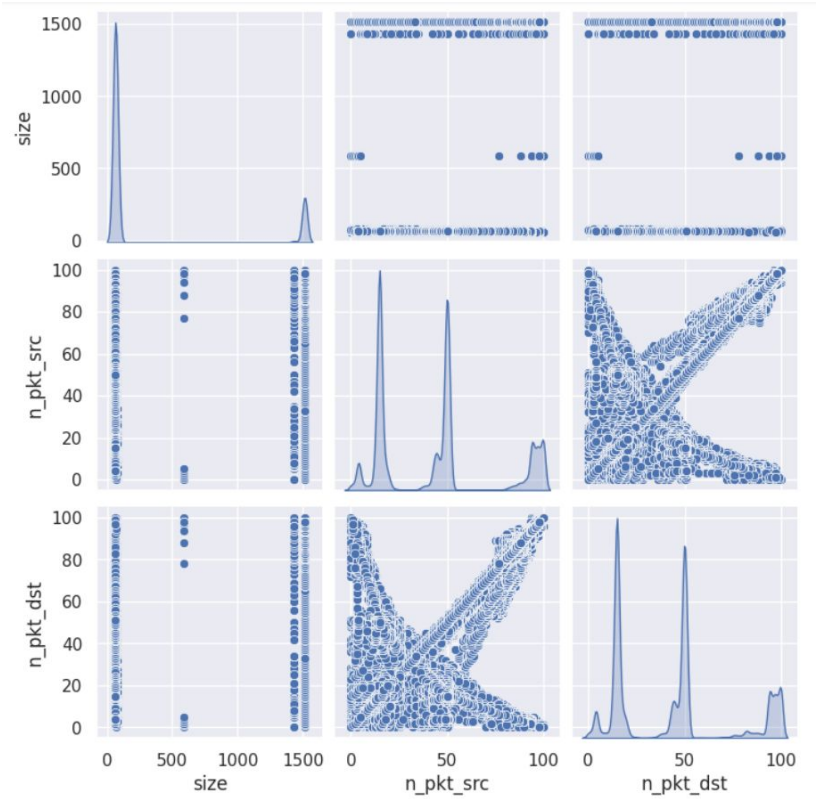
Analyse exploratoire des données

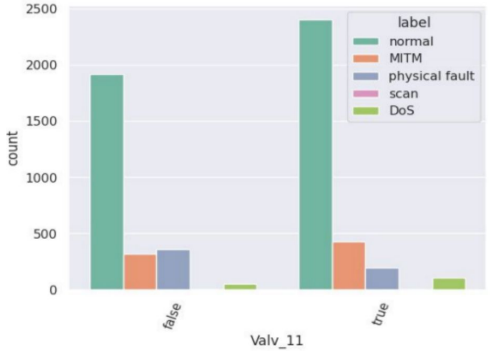


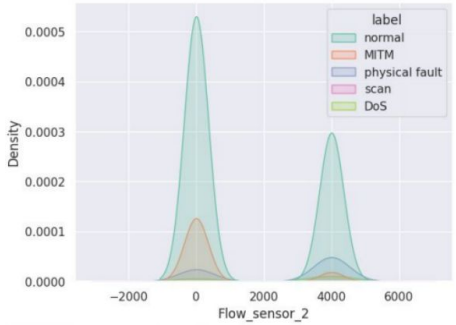
Analyse exploratoire des données





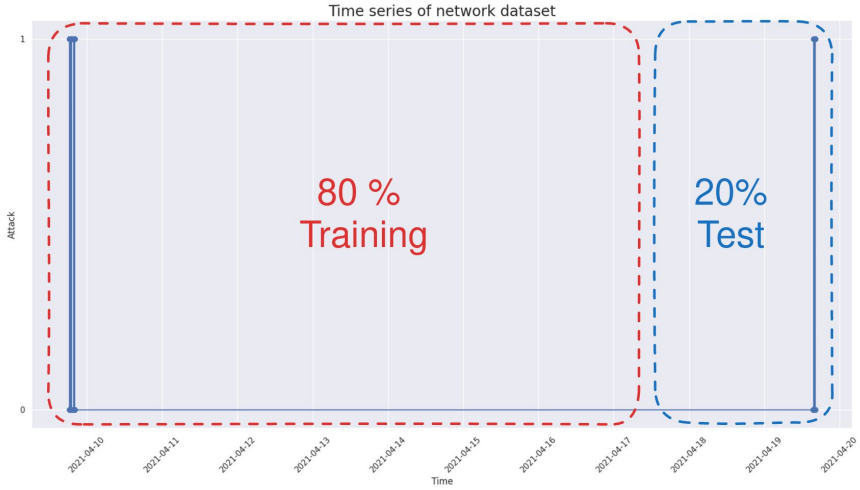






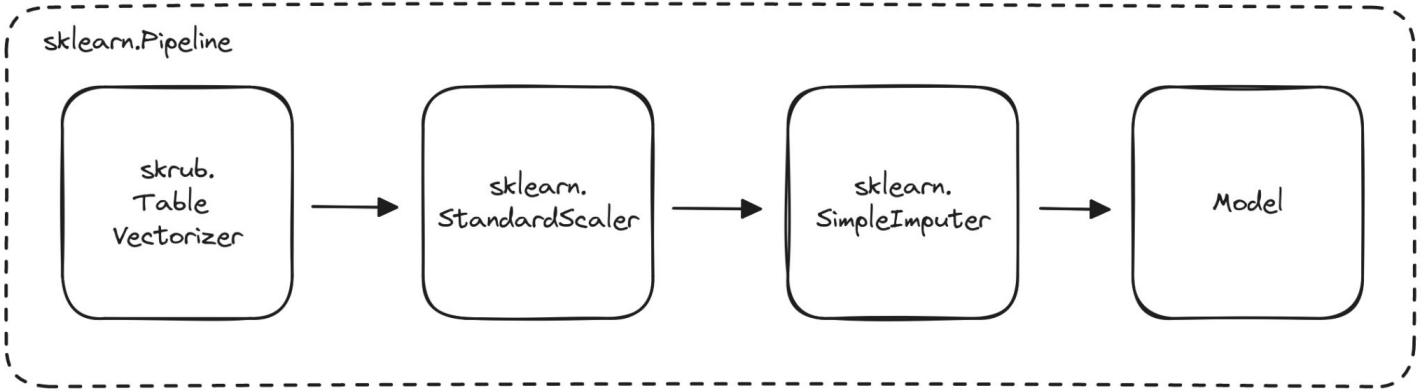
Traitement des données

- Apprentissage déséquilibré
  - Undersampling
  - Balanced RandomForestClassifier
  - Oversampling (SMOTE)
- Séparation temporelle
  - Éviter les biais sur les prédictions temporelles

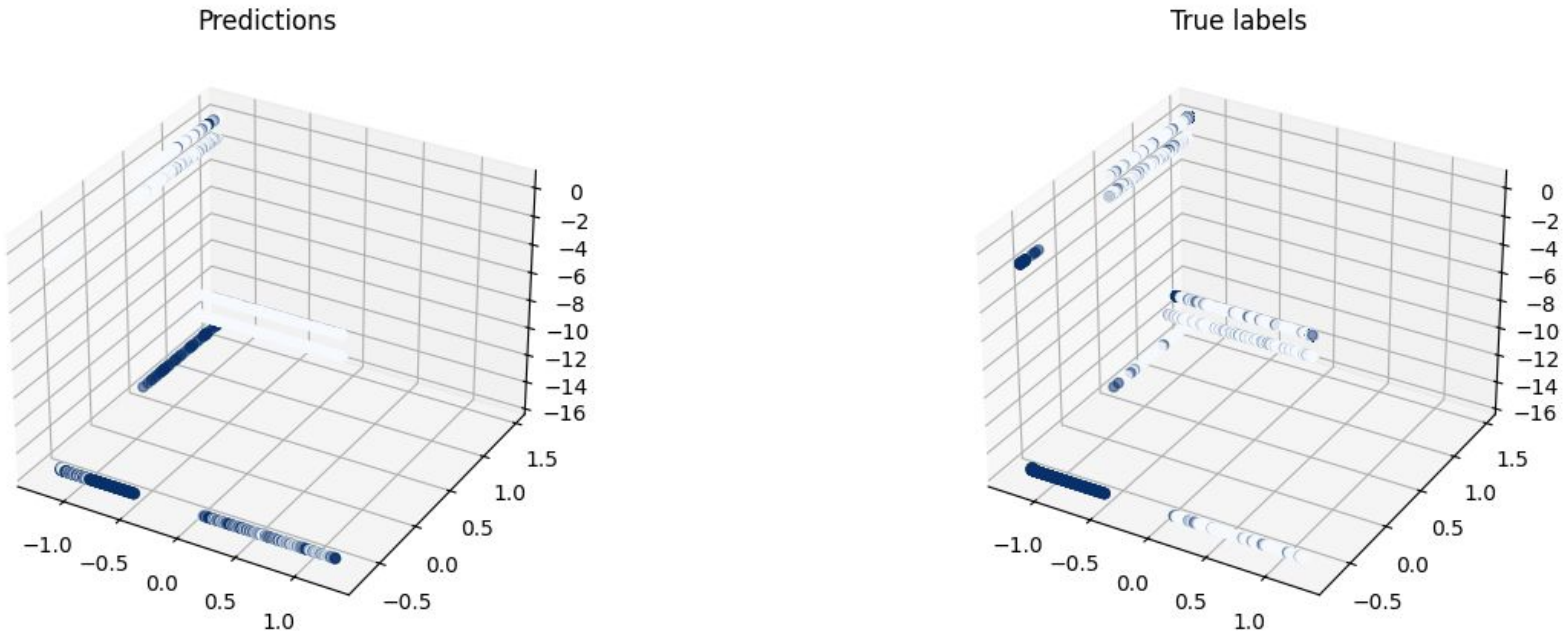


Prétraitement et transformation

- Échantillonnage du jeu de données network
- TableVectorizer → OneHotEncoder
- Imputation des valeurs manquantes
- Scaling des features numériques

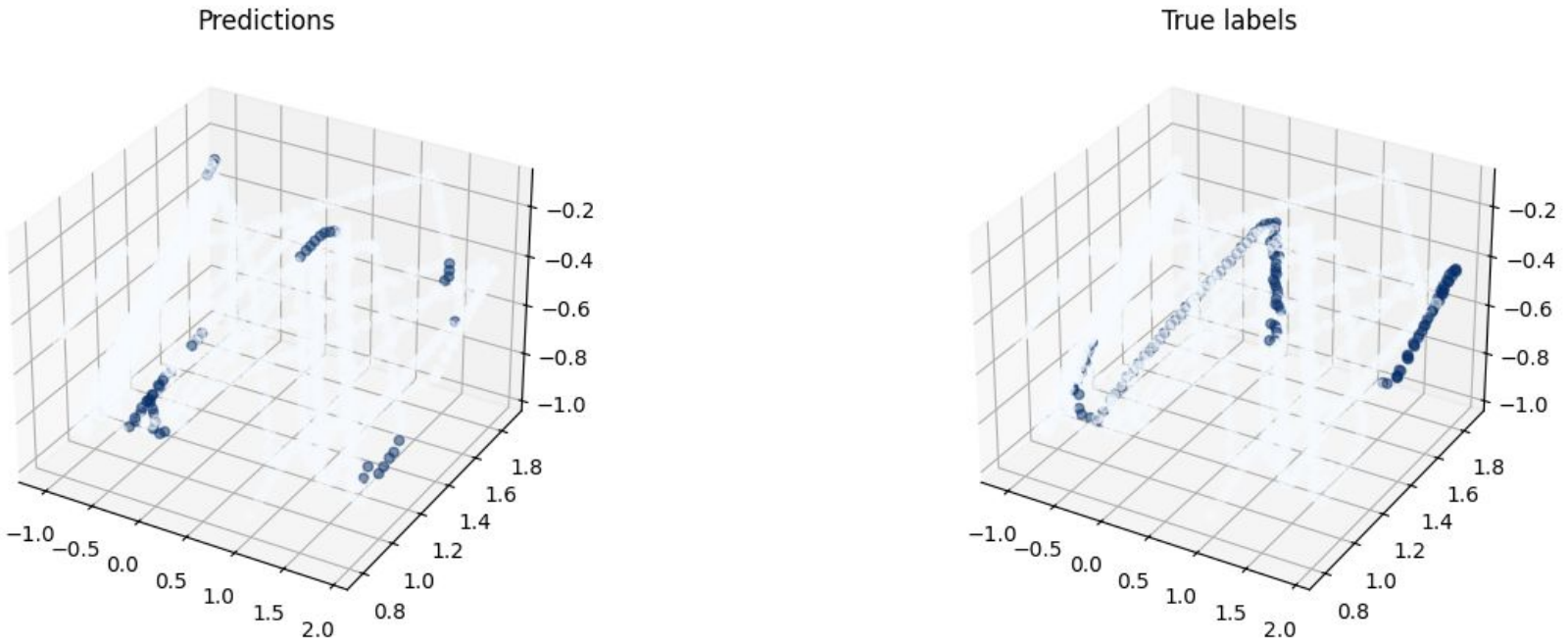


LocaloutlierFactor et IsolationForest



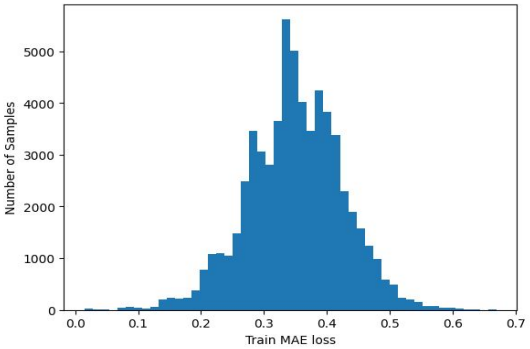
LocalOutlierFactor - network Dataset - contamination 0.01

LocaloutlierFactor et IsolationForest

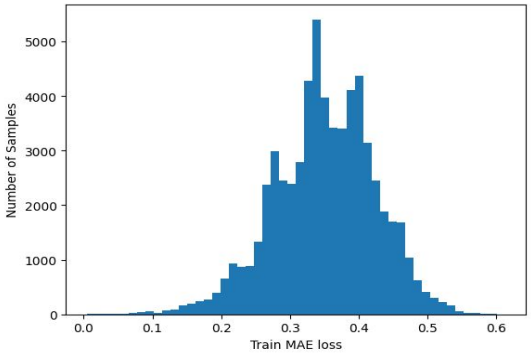


IsolationForest - physical Dataset - contamination 0.05

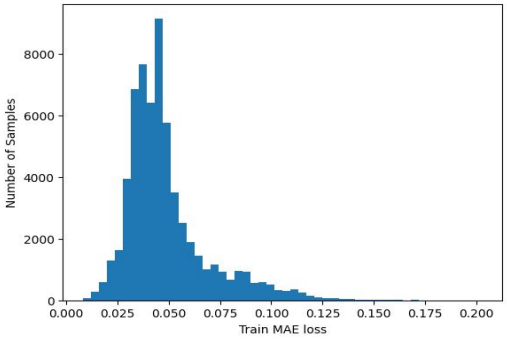
LSTM



(a) Histogram of train MAE loss for n pkt dst

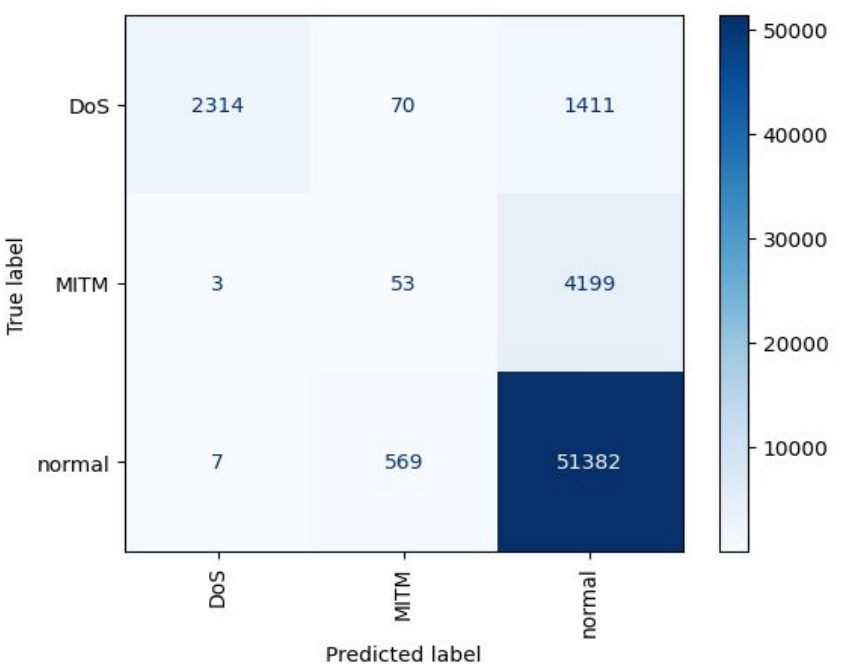


(b) Histogram of train MAE loss for n pkt src

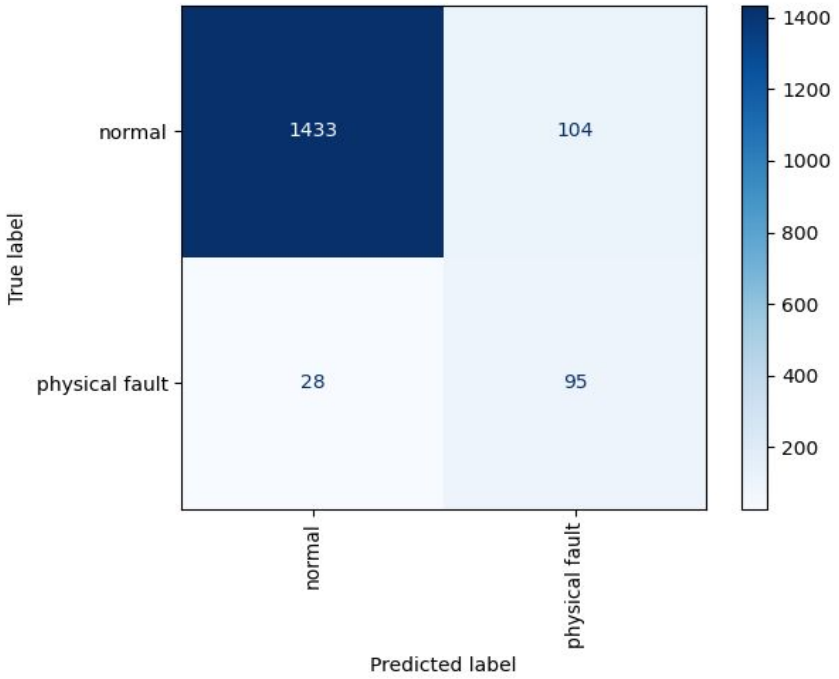


(c) Histogram of train MAE loss for size

DecisionTree

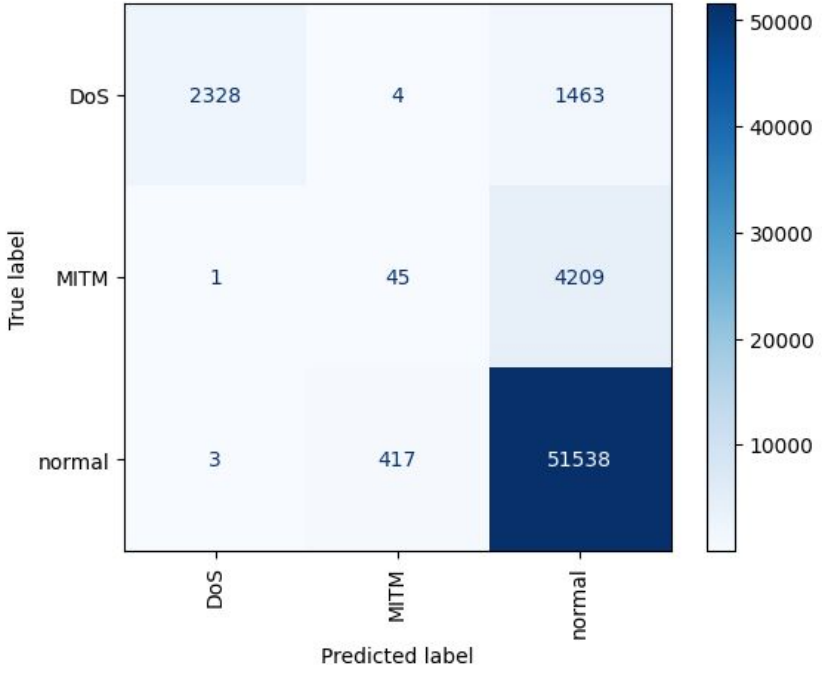


network dataset

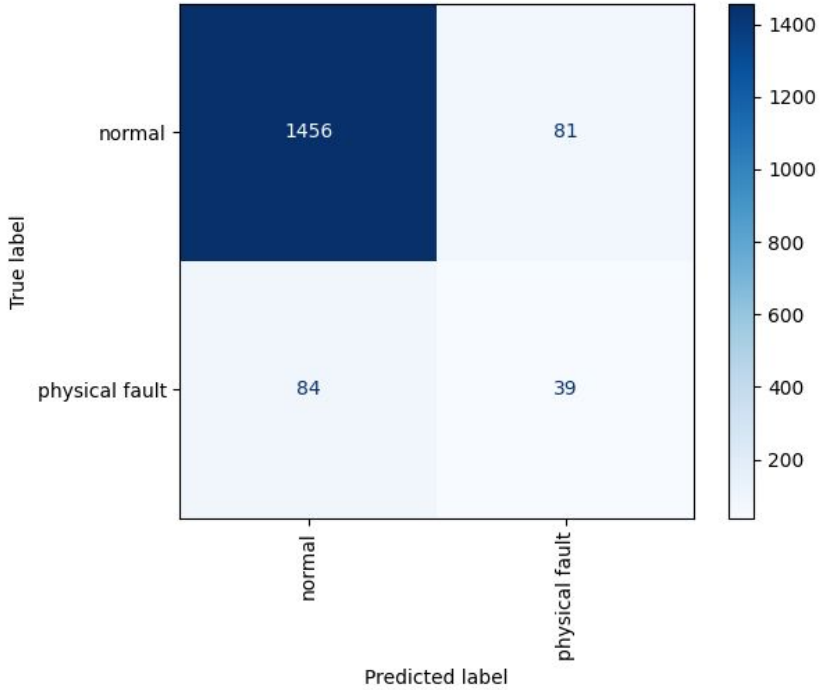


physical dataset

RandomForest

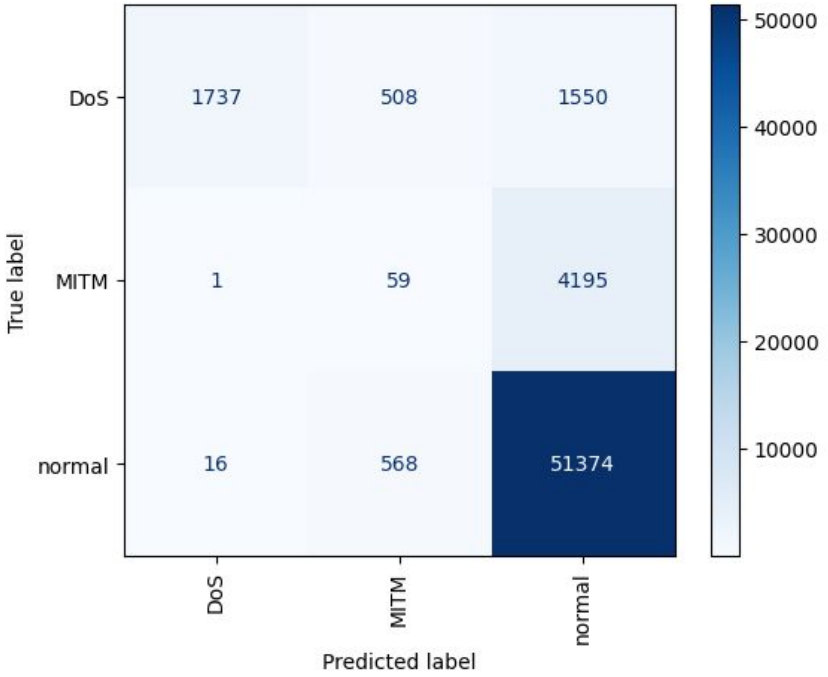


network dataset

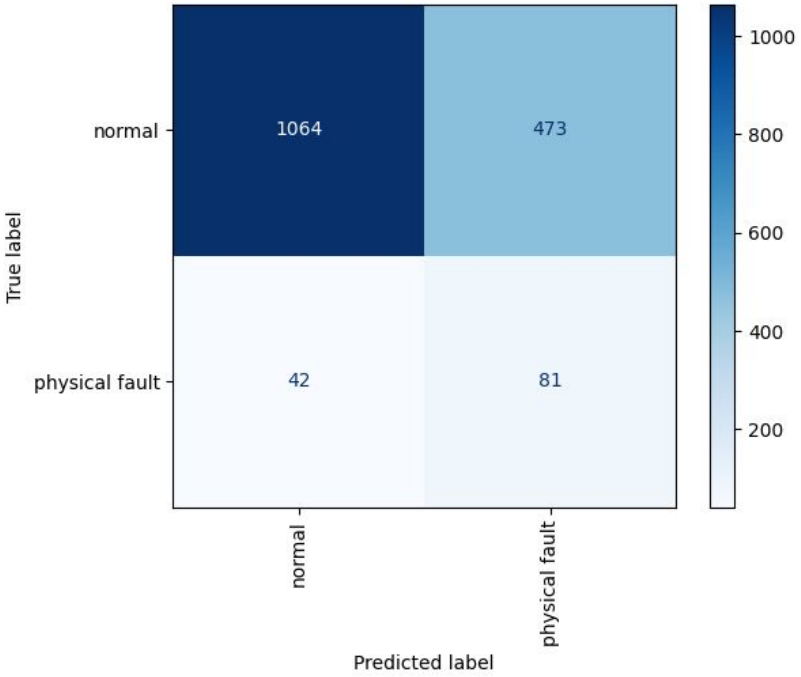


physical dataset

LinearSVM

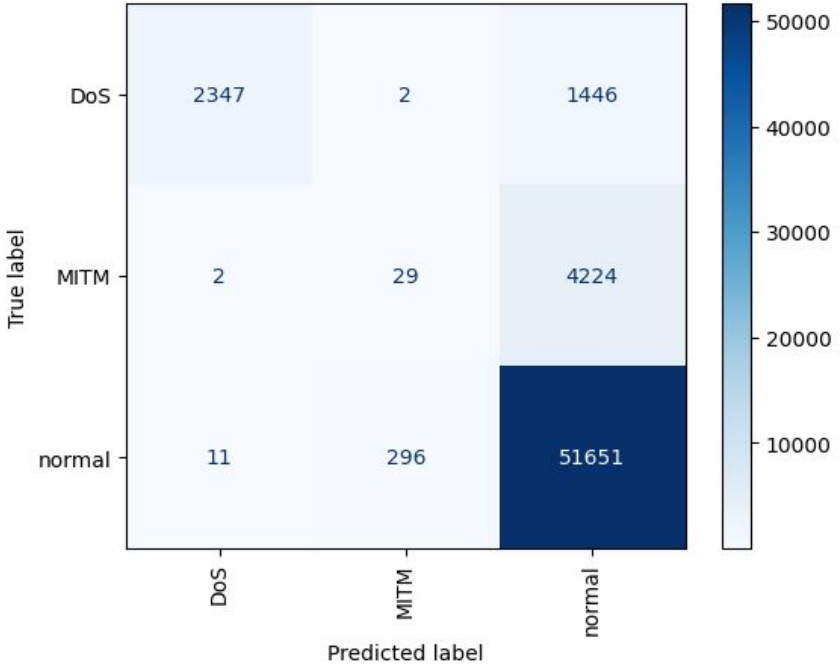


network dataset

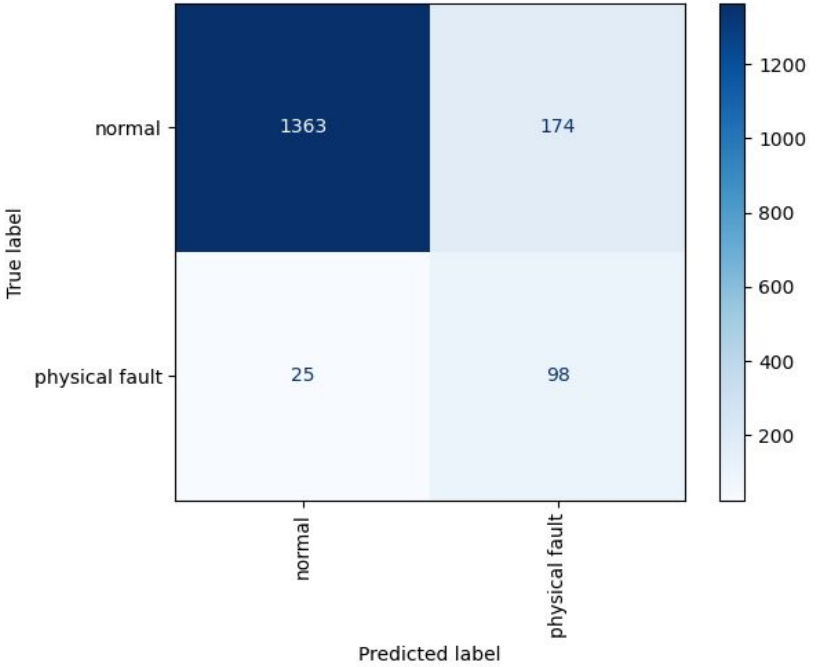


physical dataset

XGBoost

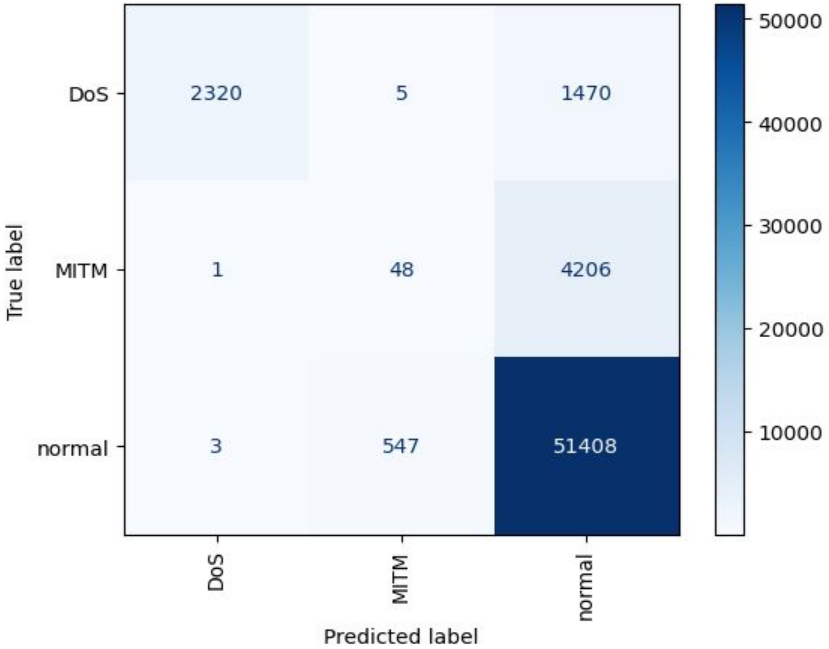


network dataset

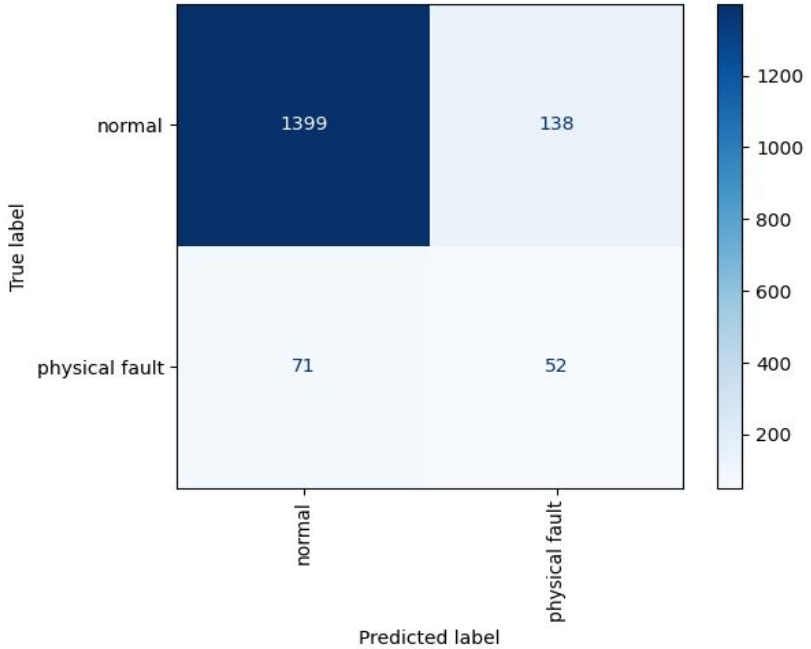


physical dataset

BalancedRandomForest



network dataset



physical dataset

Résultat des modèles – Network Dataset

	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	MCC
Decision Tree	0.90	0.55	0.85	0.90	0.87	0.49
Random Forest	0.90	0.54	0.85	0.90	0.87	0.47
Balanced Random Forest	0.90	0.54	0.85	0.90	0.87	0.46
LinearSVC	0.89	0.49	0.84	0.89	0.86	0.39
XGBoost	0.90	0.54	0.85	0.90	0.87	0.48
Local Outlier Factor	0.86	0.56	0.82	0.86	0.82	0.19
Isolation Forest	0.86	0.65	0.85	0.87	0.86	0.35

Résultat des modèles – Physical Dataset

	Accuracy	Balanced Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	MCC
Decision Tree	0.92	0.85	0.94	0.92	0.93	0.56
Random Forest	0.90	0.63	0.90	0.90	0.90	0.26
Balanced Random Forest	0.87	0.66	0.90	0.87	0.89	0.27
LinearSVC	0.69	0.67	0.90	0.69	0.76	0.19
XGBoost	0.88	0.84	0.94	0.88	0.90	0.48
Local Outlier Factor	0.92	0.52	0.89	0.93	0.89	0.15
Isolation Forest	0.92	0.50	0.87	0.89	0.86	0.03

Conclusion

- Taux de détection élevé pour les attaques DoS avec un taux minimal de **faux positifs**.
- **Détection des attaques MITM** peut nécessiter des caractéristiques supplémentaires.
- Le **prétraitement** à l'impact le plus important sur le modèle.
- Le traitement des **déséquilibres** des classes a un impact considérable sur les résultats du modèle
- Un modèle simple, comme le classificateur d'arbre de décision, présente de bonnes performances. **Recommandation XGBoost pour l'overfitting**.

Demo