École d'ingénieurs
Télécom Physique
Université de Strasbourg

Partenaire stratégique
Institut Mines-Télécom

Rapport Du Projet Analyse de la Sécurité des Systèmes Cyber-Physiques

FAID Mohamed - ELMAHI Aymane

Class of 2024

Filière Informatique et réseau
Science des données et systèmes complexes

2023/2024

22- 11 - 2023

Contents

Contexte et Objectifs du Projet

Dans un contexte auquel la cybersécurité des systèmes industriels s'avère cruciale, ce projet se focalise sur la validation des solutions de détection d'intrusion, notamment les Systèmes de Détection d'Intrusion (IDS), à travers l'analyse d'un dataset novateur. Issu du Water Distribution Testbed (WDT), cet ensemble de données combine des informations physiques et réseau, permettant ainsi une évaluation holistique des attaques dans les Systèmes Cyber-Physiques (CPS). Face aux lacunes des datasets existants, cette recherche propose une approche plus réaliste, mettant l'accent sur les conséquences des attaques à la fois sur les processus physiques et le comportement du réseau. Ce paragraphe expose la nécessité d'évaluer les performances des algorithmes d'IDS dans des conditions réalistes, relevant ainsi les enjeux cruciaux de la sécurité dans les environnements industriels modernes.

Analyse du Dataset

I. Description du Dataset

Les données utilisées dans cette étude proviennent du Water Distribution Testbed (WDT) et se composent de deux sous-systèmes distincts : un réel et un simulé. Le sous-système réel comprend cinq réservoirs, vingt vannes électromagnétiques, quatre pompes et cinq capteurs de pression, tandis que le sous-système simulé, créé à l'aide de l'outil minicps, ajoute trois réservoirs, deux pompes, quatre capteurs de débit, deux vannes électromagnétiques et trois capteurs de pression pour chaque réservoir. Cette configuration crée un banc d'essai en boucle matériel-logiciel, où l'eau circule entre les deux sous-systèmes, offrant ainsi une représentation complète et réaliste des systèmes cyber-physiques.

Le trafic réseau de tous les segments du réseau a été capturé à l'aide du logiciel Wireshark, et les caractéristiques ont été extraites des fichiers pcap résultants à l'aide de Python. La sélection des caractéristiques a été réalisée en considérant la nature déterministe et statique des réseaux de systèmes de contrôle industriels (ICS). Ces caractéristiques incluent des éléments tels que les adresses IP source et de destination, les adresses MAC source et de destination, les ports source et de destination, le protocole, les indicateurs TCP, la taille de la charge utile, le code et la valeur MODBUS, ainsi que des métriques liées au nombre de paquets source et destination dans les deux dernières secondes.

II. Prétraitement de Données

Le prétraitement des données constitue une étape fondamentale pour garantir la qualité et la cohérence des ensembles de données, et cette procédure demeure iden-

tique pour l'ensemble des datasets utilisés dans cette étude. Dans cette optique, plusieurs opérations sont effectuées de manière uniforme. Tout d'abord, les valeurs manquantes (NaN) sont remplacées soit par la moyenne (pour les variables numériques), soit par le mode (pour les variables catégorielles). Ceci vise à conserver l'intégrité des données en fournissant des approximations pertinentes. Ensuite, une vérification et rectification des noms de colonnes sont entreprises pour assurer une uniformité et faciliter la manipulation ultérieure.

L'élimination des caractéristiques dépourvues de pertinence constitue une étape cruciale, éliminant ainsi les variables qui ne présentent qu'une seule valeur, n'apportant aucune information significative à l'analyse. Cette action contribue à réduire la dimensionnalité des données tout en conservant uniquement les attributs ayant un impact substantiel sur les résultats.

La normalisation des variables numériques est effectuée pour standardiser l'échelle des différentes caractéristiques, garantissant ainsi une comparaison équitable et éliminant tout biais dû aux unités de mesure différentes. Enfin, en ce qui concerne les variables catégorielles, une option est offerte pour le codage one-hot (one hot encoding), permettant de convertir ces attributs en une forme numérique sans introduire d'ordre artificiel entre les catégories. Cette approche flexible dans le choix du codage vise à répondre aux besoins spécifiques des algorithmes d'apprentissage automatique tout en préservant la représentation adéquate des caractéristiques catégorielles.

III. Analyse Exploiratoire

L'analyse exploratoire des données (AED) est une étape importante qui permet de comprendre la structure des données, d'identifier des tendances, des schémas, et de mettre en évidence des informations importantes. Dans le cadre de ce projet, l'AED sera menée de manière approfondie pour chaque sous-ensemble de données, à savoir le réseau, le physique et la combinaison des deux.

III.1. Données Physique

Les données physiques se distinguent par leur clarté, rendant toute anomalie particulièrement prononcée dans le fonctionnement des composants tels que les pompes et les réservoirs. Les dysfonctionnements, qu'il s'agisse de retards de remplissage ou de fuites, sont rapidement identifiables, simplifiant la détection des problèmes. La surveillance des données physiques offre une visibilité directe sur l'état opérationnel des composants, facilitant ainsi une prise de décision réactive pour assurer le bon fonctionnement du système cyber-physique.

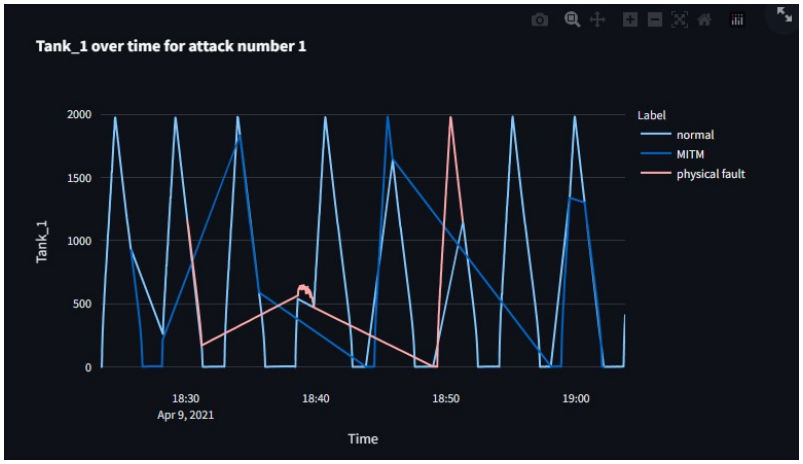

Figure 2: Variable Tank1 et type d'attaque dans le temps

III.2. Données Réseaux

Dans le contexte du jeu de données réseau, une stabilité notable est observée dans la correspondance entre les adresses MAC et les adresses IP dans les scénarios normaux. Toute altération de cette correspondance pourrait indiquer une attaque de type Man-in-the-Middle (MITM). Ces changements dans le mappage des adresses MAC vers les adresses IP sont cruciaux à surveiller, car ils peuvent compromettre la sécurité des échanges.

Il est important de souligner que les attaques au niveau du réseau ont des répercussions directes sur le système physique, influant sur des aspects tels que le remplissage des réservoirs et les fuites. Ainsi, les anomalies dans le comportement réseau servent d'indicateurs clés pour évaluer l'impact sur les composants physiques du système.

En outre, un volume élevé de paquets peut signaler une attaque de déni de service (DoS), affectant la disponibilité du système. Les modifications du mappage des adresses MAC vers les adresses IP peuvent également être des signaux d'alerte pour des attaques MITM et de scanning.

III.3. Données Réseaux et Physique

Pour le jeu de données réseau, nous avons opté pour une stratégie de sous-échantillonnage, préservant la proportion relative des différents labels (normal, DoS, MITM, scanning, etc.). Bien que cela implique la perte d'un pourcentage significatif de données, cette démarche est justifiée par le fait que les attaques se manifestent à l'échelle des secondes plutôt que des millisecondes, minimisant ainsi la perte d'informations cruciales. De plus, le regroupement (group by) est effectué en fonction de la colonne temps, assurant une cohérence temporelle dans l'analyse.

IV. Analyse Statistique

Dans le cadre de l'analyse statistique, des tests significatifs ont été appliqués pour évaluer les relations entre les variables. La corrélation de Pearson a été utilisée pour mesurer la force et la direction des liens linéaires entre les caractéristiques du réseau et les mesures physiques. Cette analyse a permis d'identifier les associations potentielles entre les comportements du réseau et les réponses physiques du système.

Parallèlement, le test ANOVA (Analyse de la variance) a été employé pour évaluer les différences significatives entre les moyennes de plusieurs groupes. C'était particulièrement utile pour comprendre les variations statistiques entre les différentes classes d'attaques et les scénarios normaux, tant au niveau du réseau que des mesures physiques.

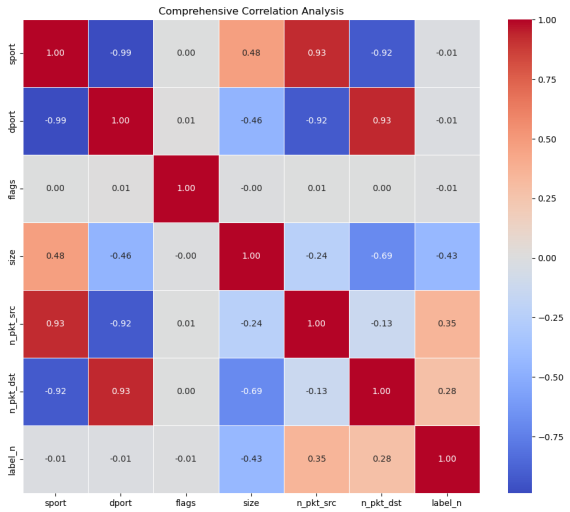

Figure 3: Matrices de corrélations entre les variables numériques des données réseaux

Expérimentations et Apprentissage Automatique

Nous avons conduit des expérimentations en utilisant différentes configurations de données pour évaluer la performance de divers algorithmes d'apprentissage automatique dans la détection d'anomalies. Trois ensembles de données distincts ont été considérés : les données physiques seules, les données réseau seules, et une combinaison des deux.

Les algorithmes, incluant k-NN, CART, Random Forest, SVM, XGBoost, MLP, et Naive Bayes, ont été appliqués à chaque ensemble de données. Nous avons varié

la taille des ensembles (de 1% à 100%) pour les données réseaux, afin d'évaluer la robustesse des modèles dans différentes conditions d'apprentissage.

Les résultats, mesurés à l'aide de métriques telles que la précision, le rappel, le F-score, et les matrices de confusion, offrent des indications cruciales sur la capacité de chaque algorithme à détecter efficacement les anomalies, guidant ainsi le choix optimal pour la conception d'un système de détection d'intrusion pour le système cyber-physique examiné.

I. Évaluation et Comparaison

Pour les données réseaux on reduit la taille à 1% de la original du dataset

Pour les ajustements fins des modèles, veuillez vous référer au notebook correspondant pour obtenir des détails spécifiques.

On a pu extraire les features important pour chaque algorithme, exemple figure 4 ci-dessous :

Algorithm	Performance Metric	Physical Dataset	Network Dataset	Mixed Dataset
KNN	Accuracy	0.91	0.77	0.92
	Recall	0.89	0.44	0.77
	Precision	0.92	0.68	0.91
	F1 score	0.85	0.53	0.83
Random Forest	Accuracy	0.96	0.75	0.93
	Recall	0.89	0.52	0.87
	Precision	0.91	0.50	0.89
	F1 score	0.92	0.61	0.88
SVM	Accuracy	0.90	0.61	0.82
	Recall	0.91	0.20	0.94
	Precision	0.64	0.50	0.90
	F1 score	0.73	0.20	0.84
Cart	Accuracy	0.91	0.45	0.90
	Recall	0.90	0.15	0.87
	Precision	0.72	0.50	0.69
	F1 score	0.77	0.27	0.65
SVM	Accuracy	0.90	0.75	0.94
	Recall	0.91	0.15	0.96
	Precision	0.64	0.95	0.80
	F1 score	0.71	0.27	0.83
Xgboost	Accuracy	0.91	0.73	0.90
	Recall	0.91	0.15	0.86
	Precision	0.66	0.80	0.80
	F1 score	0.72	0.19	0.84
Mlp	Accuracy	0.93	0.75	0.94
	Recall	0.82	0.15	0.70
	Precision	0.66	0.70	0.88
	F1 score	0.64	0.17	0.79

Table 1: Performances des algorithmes sur les jeux de données physiques et réseau et combinés.

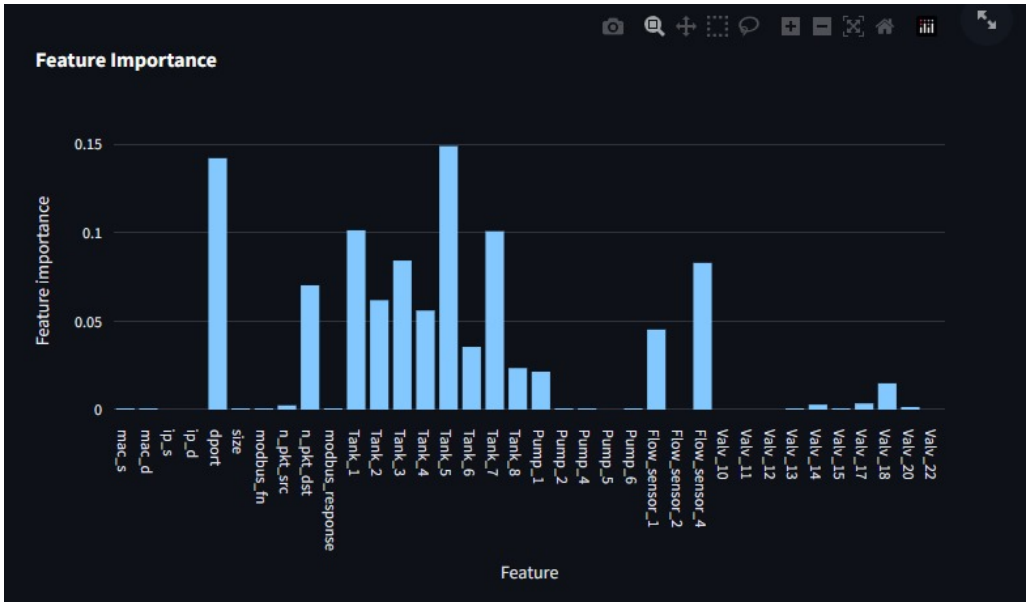

Figure 4: features d'importance pour le decision tree algoithm sur les données mixtes

II. Impact de la Taille de l'Ensemble d'Apprentissage

Lorsqu'il s'agit de construire des modèles d'apprentissage automatique pour les données réseau, la taille de l'ensemble d'apprentissage joue un rôle crucial dans la performance du modèle. Utiliser l'intégralité de l'ensemble de données peut sembler être une approche intuitive, mais cela peut souvent conduire à des problèmes dus à un excès de bruit qui perturbe la capacité du modèle à généraliser.

D'un autre côté, avoir un ensemble d'apprentissage trop petit peut ne pas fournir suffisamment d'informations au modèle pour apprendre efficacement les motifs sous-jacents. Une taille d'ensemble d'apprentissage insuffisante peut conduire à une sous-représentation des cas et à une mauvaise généralisation.

Des expériences menées avec différents algorithmes de classification sur des données réseau ont conduit à la conclusion qu'opter pour 0.5% de la taille originale de l'ensemble de données constitue un compromis judicieux entre les performances du modèle et la rapidité d'apprentissage.

III. Consommation de Ressources

Pour les données réseau plus volumineuses, le temps d'exécution et la consommation de RAM dépendent du downsampling. Cependant, en utilisant seulement 0,5% des données, on parvient à des temps d'exécution raisonnables, réduisant le facteur à environ un quart par rapport aux données physiques.

Pour plus de données sur le temps d'exécution et la RAM consommée, consulter

Algorithme	Temps d'Exécution (s)	Consommation RAM (Mo)
KNN	0.2	15
RF	1.3	104
SVM	34	78
NB	13.2	56

Table 2: Performances de quelques algorithmes en termes de temps d'exécution et de consommation RAM pour les données physiques et mixtes.

les notebooks.

Comparaison avec les Résultats Publiés

Nos résultats, en particulier l'application de Random Forest (RF) et de k-NN sur les données physiques, montrent des performances similaires à celles rapportées dans la littérature. De manière notable, notre approche hybride, combinant les données réseau et physiques, démontre une efficacité marquée, alignant sur les avancées antérieures dans la détection d'anomalies cyber-physiques.

Application Web

Nous avons utilisé le framework Streamlit pour créer une application web qui permet de visualiser les données de notre dataset.

Nous avons introduit le concept du multipage dans notre application, ce qui permet de naviguer entre les différentes pages de l'application. Cela a permis de rendre l'application plus ergonomique et plus facile à utiliser.

Les trois pages de l'application sont les suivantes: - La page d'accueil - La page de visualisation des données physiques - La page de visualisation des données network - La page de visualisation des données mixtes

Dans chaque page, nous avons utilisé des widgets pour permettre à l'utilisateur de choisir les données qu'il souhaite visualiser.

Nous avons divisé chaque page en plusieurs parties: les parties d'analyse descriptive et des parties d'entrainement et de test des modèles de machine learning. Nous avons permis à l'utilisateur de choisir les données qu'il souhaite visualiser et les modèles qu'il souhaite entrainer et tester.

Veuillez vous rendre sur le README.md pour plus d'informations sur l'application web.

Conclusion

En résumé, l'approche hybride, combinant données physiques et réseau, surmonte les limitations des jeux de données existants pour la détection d'anomalies dans les systèmes cyber-physiques. Les résultats des expérimentations avec quatre algorithmes d'apprentissage automatique soulignent l'importance cruciale de cette intégration pour une détection plus complète des attaques.

Contribution

Mohamed Faid:

- Analyse du dataset réseau
- Exploration d'algorithmes de classification
- Rédaction du rapport

Aymane Elmahi:

- Analyse du dataset physique et mixte
- Développement de l'application Streamlit
- Organisation du dépôt Github