**Explicabilité des comportements malveillants dans les applications Windows via des graphes de flux de contrôle (CFG) et des méthodes d'explicabilité avancées**

**Résumé**

Cet article présente une approche d’explicabilité appliquée à la détection et l’analyse des comportements malveillants dans les applications Windows en utilisant des graphes de flux de contrôle (CFG) et des techniques d’explicabilité avancées. Nous combinons des méthodes d’explicabilité locales (LIME, SHAP) et globales (TreeExplainer, Grad-CAM) pour mieux comprendre les prédictions des modèles de classification de graphes. Les résultats montrent une amélioration significative de la compréhension des comportements malveillants.

**Introduction**

L’analyse des applications Windows pour détecter les comportements malveillants est un défi majeur en cybersécurité. Les modèles de classification de graphes permettent d’extraire des caractéristiques importantes des flux de contrôle, mais leur nature complexe rend difficile l’interprétation des décisions. Cet article propose une méthodologie d’explicabilité combinant des méthodes locales et globales pour fournir une meilleure interprétation des modèles.

**Méthodologie**

Notre méthodologie repose sur l'utilisation de modèles de classification de graphes appliqués aux CFGs extraits des applications Windows. Elle repose sur une représentation des applications Windows sous forme de graphes de flux de contrôle (CFG). Les nœuds de ces graphes représentent les instructions assembleur, tandis que les arêtes indiquent les transitions possibles entre ces instructions. L'approche comprend les étapes suivantes :

1. **Extraction des CFGs** : Les applications sont analysées pour extraire leurs CFGs, en identifiant les fonctions et les blocs de base constitués d'instructions assembleur.
2. **Classification des CFGs** : Un modèle d'apprentissage est entraîné sur ces graphes pour détecter des comportements potentiellement malveillants.
3. Protocole d’évaluation
4. **Explicabilité** : Les décisions du modèle sont expliquées en identifiant les parties du graphe qui contribuent le plus à la classification.

*Protocole d’évaluation* :

* **Matrice de confusion multi label, avec performance (Accuracy, Recall, Precision, F1-Score par label)**

A screenshot of a computer

AI-generated content may be incorrect.

TP: True Positive (threat correctly detected), TN: True Negative (no threat correctly detected), FP: False Positive (threat missed), FN: False Negative (no-threat graph classified as threat)

La performance par label (ou groupe de labels) permet de vérifier la performance du label sur certains labels et valider le design et l’entrainement . Pour les labels, ayant des scores insuffisants, il permet de revoir la conception du modèle et de l’entrainement en considérant les features manquantes

* **Le F1-score macro (sur l’ensemble des labels) permet de « benchmarker » le modèle par rapport à d’autres solutions (modèles proposés)**

Pour améliorer l’interprétabilité de ces modèles, nous appliquons différentes méthodes d’explicabilité :

**Méthodes locales**

* **LIME (Local Interpretable Model-agnostic Explanations)** : Perturbe les données d’entrée et mesure l’impact sur la sortie pour construire un modèle interprétable localement.
* **SHAP (SHapley Additive exPlanations)** : Basé sur la théorie des valeurs de Shapley, il calcule l’importance de chaque feature en attribuant un score d’explication.

**Méthodes globales**

* **TreeExplainer** (pour les modèles basés sur des arbres) : Permet une interprétation globale des modèles en évaluant l’importance moyenne des features sur l’ensemble des prédictions.
* **Grad-CAM** (Gradient-weighted Class Activation Mapping) : Utilisé principalement pour les réseaux de neurones convolutifs, mais adaptable aux modèles de classification de graphes pour visualiser les parties importantes du graphe.

**Intégration dans notre approche**

Les méthodes d’explicabilité sont intégrées au processus de classification des CFGs pour fournir des explications compréhensibles par un humain. Les méthodes locales permettent d'expliquer des prédictions individuelles tandis que les méthodes globales fournissent une vision d'ensemble des décisions du modèle.

**Résultats**

Nous avons évalué notre méthodologie sur un ensemble de données d’applications Windows malveillantes et bénignes. Les résultats montrent que l’utilisation combinée des méthodes locales et globales permet une compréhension plus fine des comportements malveillants détectés.

**Conclusion**

Cet article présente une méthodologie d’explicabilité combinant des techniques locales et globales pour améliorer l’interprétation des modèles de classification de graphes appliqués aux CFGs. Cette approche permet non seulement de détecter des comportements malveillants mais aussi d’en fournir une explication compréhensible par un expert en cybersécurité.

**Références**

* Buijsman, S. (2023). *Defining Explanation and Explanatory Depth in XAI*. *Philosophy & Technology*, *36*, Article 39..
* Denis, C., & Varenne, F. (2022). *Interprétabilité et explicabilité pour l’apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique*. *Revue d’intelligence artificielle*, *36*(1), 5-29..
* Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. IEEE International Conference on Computer Vision.