Malware-XAI

Presentation by **Adolfo Mora Córdova**

Noviembre 11, 2025

Introducción

La IA en ciberseguridad funciona como caja negra: es dificil validar el porque una alerta es disparada o como se deciden las acciones.

Lo cual reduce la confianza y complica las auditorias ya las mejoras del modelo.

Se requiere transparencia, seguridad y fiabilidad en entornos sensibles (IDS, SOC, etc.)

PROPOSED FRAMEWORK

Malware-XAI

1 Explicabilidad

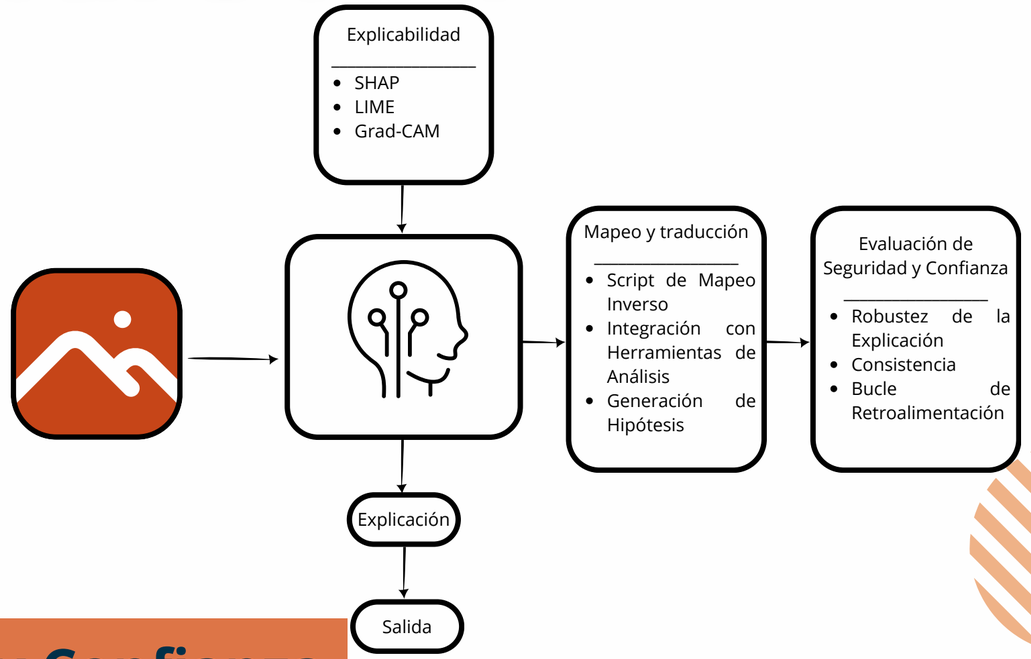
SHAP, LIME, Grad-CAM

2 Mapeo y traducción

Script de Mapeo Inverso, Integración con Herramientas de Análisis, Generación de Hipótesis

3 Evaluación de Seguridad y Confianza

Robustez de la Explicación, Consistencia, Bucle de Retroalimentación



Explicabilidad (XAI)

SHAP

Asignar un valor de "importancia" preciso a cada "píxel" (byte) individual de la entrada.

Un mapa de alta definición que te dice exactamente qué bytes contribuyeron positiva o negativamente a la clasificación.

LIME

Validar los hallazgos de SHAP/Grad-CAM. LIME crea un modelo interpretable simple que "explica" la predicción de la CNN solo para esa muestra específica.

Muestra qué "super-píxeles" (regiones de la imagen) fueron más importantes para el modelo local simple

Grad-CAM

Generar un mapa de calor (heatmap) de baja resolución superpuesto a tu "imagen" de malware.

Muestra las regiones generales que la CNN consideró más relevantes para su decisión.

Mapeo y traducción

Script de Mapeo Inverso

Script que tome las coordenadas (x, y) de los píxeles más importantes (identificados por SHAP o Grad-CAM) y las traduzca de nuevo a su dirección en el archivo binario original.

Integración con Herramientas de Análisis

Examinar en un desensamblador o editor hexadecimal

Generación de Hipótesis

Responder preguntas clave:

- ¿El byte resaltado es parte de una cadena de texto sospechosa?
- ¿Está al inicio de una sección de código ofuscado (packer)?

Evaluación de Seguridad y Confianza

Robustez de la Explicación

El framework debe incluir pruebas contra ataques de adversario. ¿Puede un malware ser diseñado para engañar a tu CNN y también a tus métodos XAI.

Consistencia

¿Las diferentes técnicas XAI (LIME, SHAP, Grad-CAM) apuntan a regiones similares? Si hay un gran desacuerdo, la confianza en la explicación disminuye.

Bucle de Retroalimentación

La explicación del (ej. "La CNN acertó, y Grad-CAM señaló correctamente la sección del packer") debe usarse para reentrenar y mejorar el modelo.

Thank You So Much!

Presentation by **Adolfo Mora Córdova**