PROPUESTA DE UN ENFOQUE BASADO EN EL APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE INTRUSIONES EN REDES DE TECNOLOGÍA DE OPERACIÓN

Isabel Herrera Montano[1], Mª Carmen Palacios[2], Jaime Garvía García[3], Isabel de la Torre Díez[1] y José Javier García Aranda[4]

[1]*Universidad de Valladolid, Paseo de Belén, 15, 47011 - Valladolid, Spain.*
[2]*TECNALIA, Basque Research and Technology Alliance (BRTA), Spain*
[3]*Servicios de Consultoría IT, Polígono Industrial Kurutz Gain 12-13| 20850 | Mendaro – Gipuzkoa, Spain*
[4]*Nokia, Maria Tubau Street, 9, 28050, Madrid, Spain*

RESUMEN

La detección de intrusiones en la infraestructura de organizaciones industriales ha sido de interés para múltiples investigadores en la actualidad. En este sentido, el tema que nos ocupa en este trabajo de investigación en curso, es la detección de intrusiones en las redes de tecnologías de las operaciones (OT), donde se ubican los sistemas de control industrial, y por tanto, un ataque puede traer graves consecuencias en el funcionamiento de los equipos industriales. El objetivo de este artículo es proponer una solución de detección de intrusiones en redes OT basada en técnicas de aprendizaje automático para detección de anomalías, tanto para ataques de día cero, como para ataques conocidos. A través del análisis de la información de cabeceras de los paquetes y datos estadísticos de los flujos de tráfico de red TCP/UDP, apoyada en un análisis realizado a las matrices ATT&CK Matrix. Para el desarrollo de nuestra propuesta se crearán conjuntos de datos específicos para este objetivo y se implementarán diferentes modelos de aprendizaje automático que nos permitirán extraer eventos anómalos para detectar diferentes ataques y tomar decisiones al respecto. Con la solución propuesta se intenta reducir las altas tasas de falsos positivos habituales en estos sistemas y detectar ataques tanto de día cero como conocidos.

PALABRAS CLAVE

Detección de Intrusiones, Inteligencia Artificial, Redes OT, Seguridad, Red de Tecnologías de Operaciones, Aprendizaje Automático

1. INTRODUCCIÓN

En entornos industriales se denomina red de tecnologías de las operaciones, o redes OT, a la interconexión directa y la gestión de sistemas y procesos físicos (Acarali *et al.*, 2022). Por motivos de seguridad, en estos entornos se está promoviendo cada vez más el uso de comunicaciones cifradas, que consecuentemente dificulta la detección de comportamientos maliciosos mediante técnicas basadas en Deep Packet Inspection (o DPI), debido a su necesidad de análisis del contenido de los paquetes (De La Torre Parra, Rad y Choo, 2019). El objetivo de este estudio es proponer una solución de detección de intrusiones en redes OT basada en técnicas de aprendizaje automático (o ML por sus siglas en inglés) para detección de anomalías, tanto para ataques de día cero como para ataques conocidos. La solución propuesta se apoya en un análisis realizado a las matrices ATT&CK Matrix (Mohamed, Jantan y Abiodun, 2018) (Mashima, 2022). MITRE ATT&CK es una base de conocimiento accesible a nivel mundial de tácticas y técnicas del adversario basadas en observaciones del mundo real (MITRE ATT&CK®, 2015). Las tácticas se refieren a qué intentan lograr los atacantes. Las técnicas, se corresponden a cómo el adversario pretende lograr ese objetivo. Para cada una de las técnicas, MITRE proporciona información sobre posibles mecanismos para su detección y/o mitigación. Estudios recientes como (Xiong et al., 2022), (Liu, Wang y Chen, 2022) y (Shin et al., 2022) también basan sus soluciones en estas matrices.

Un sistema de detección de intrusiones (en inglés, IDS "Intrusion Detection System") consiste en un conjunto de métodos y técnicas diseñadas para revelar actividad sospechosa sobre un recurso o recursos informáticos. Es decir, detectar eventos que sugieren un comportamiento anómalo, incorrecto o inapropiado según las políticas de seguridad consideradas en el entorno a proteger (Tidjon, Frappier y Mammar, 2019).

Nuestra solución es un trabajo de investigación en curso, que combinará una amplia variedad de técnicas de detección de intrusiones basadas en ML tanto supervisado como no supervisado, para intentar reducir las altas tasas de falsos positivos habituales en estos sistemas y detectar ataques tanto de día cero como conocidos. A continuación, se describe en qué se basan las técnicas de detección de anomalías y específicamente, las técnicas de día cero, así como, el estado del arte de estas técnicas en la literatura. En la sección 2 de este artículo se describe la arquitectura y el funcionamiento de la solución propuesta, los métodos utilizados para su desarrollo se describen en la sección 3 y posteriormente se concluye este artículo.

1.1 Detección de Anomalías

En el estado del arte actual se han descrito, implementado y probado una gran variedad de técnicas de detección de anomalías. La detección basada en anomalías, como sistema de detección de intrusiones basado en el comportamiento, crea un perfil de la actividad normal de un sistema para posteriormente supervisar su comportamiento y alertar sobre cualquier desviación del comportamiento considerado como normal o "baseline". Este perfil se genera durante un período de tiempo predefinido mientras el sistema se comporta en condiciones normales. Por tanto, su mayor ventaja es que pueden detectar ataques nunca vistos y presentan un alto porcentaje de detección de ataques. Sin embargo, estos sistemas ofrecen un alto número de falsos positivos, ya que cualquier actividad nueva en el sistema se considerará como un ataque. La principal vulnerabilidad de este tipo de IDS es que se produzca un ataque durante la fase de entrenamiento ya que éste se considerará en un futuro como tráfico legítimo. Asimismo, generalmente, presentan un alto coste computacional.

La detección de anomalías se puede clasificar en dos clases principales: Programadas y Autoaprendizaje, según el método utilizado para crear el perfil normal o básico de un sistema (Hodo *et al.*, 2017). La categoría "Programada" se refiere a modelos que necesitan que una persona le enseñe manualmente a detectar cambios en el comportamiento del sistema. Por lo tanto, el ser humano es quien decide el alcance de los comportamientos anormales en el sistema y los marca como una amenaza de intrusión. En la categoría de "Autoaprendizaje" se entrena un modelo con el tráfico de red recopilado durante un período de tiempo para aprender cómo se comportan los procesos subyacentes. Las Series Temporales y las técnicas de ML pertenecen a esta categoría.

En la literatura existen trabajos de investigación relevantes donde los algoritmos de ML han sido aplicados en el dominio de detección de intrusiones logrando resultados muy prometedores. Por ejemplo, los autores de (Sharafaldin, Habibi Lashkari y Ghorbani, 2018) utilizaron un regresor Bosque Aleatorio (RF por sus siglas en inglés) para determinar el mejor conjunto de características para detectar cada familia de ataques. Posteriormente, se examinó el rendimiento de estas características mediante la aplicación de diferentes algoritmos como K-Nearest Neighbours (KNN), AdaBoost, Multi-Layer Perceptron (MLP), Naïve Bayes, RF, Iterative Dichotomiser 3 (ID3) y Quadratic Discriminant Analysis (QDA). En (Vijayanand, Devaraj y Kannapiran, 2018), se evaluó un sistema de detección de intrusiones basado en un algoritmo genético para seleccionar características junto con múltiples Máquinas de Vectores de Soporte (en inglés SVM, *Support Vector Machines*) para su clasificación. Específicamente, este sistema combinó varios clasificadores SVM, ordenados en función de la gravedad de los ataques. En referencia a este trabajo, debe remarcarse que cada clasificador fue entrenado en un conjunto de características diferente para poder descubrir una categoría de ataque específica. Los autores de (Watson, 2018) aplicaron dos clasificadores, un Perceptrón Multicapa (en inglés MLP) y una red neuronal convolucional (en inglés CNN, *Convolutional Neural Network*). Para ello, utilizaban archivos de tráfico de red en formato "pcap" y seleccionan características específicas de las cabeceras de los paquetes de red. Asimismo, los modelos Recurrent Neural Network (RNN) y Long Short-Term Memory (LSTM) también se utilizan para detectar intrusiones en (Zhu *et al.*, 2017). En (Min *et al.*, 2018) propusieron SU-IDS; un sistema de detección de intrusiones en la red semi-supervisado y no supervisado que utilizaba un framework basado en un Auto-Encoder (o autocodificador). En este framework se tiene en cuenta tanto la pérdida habitual del agrupamiento (o clasificación) como una pérdida auxiliar del autocodificador, mejorando así su rendimiento.

1.2 Detección de Ataques de Día Cero

Actualmente, la detección de ataques de día cero (no vistos con anterioridad) es uno de los principales retos de los sistemas de detección de intrusiones. Los IDS actuales sufren altas tasas de falsos positivos, lo que limita su adopción práctica. Como consecuencia, los ataques de día cero pasan desapercibidos en la práctica, intensificando así sus impactos negativos. Un ataque de día cero puede definirse como "un patrón de tráfico de interés que, en general, no tiene patrones coincidentes en elementos de detección de ataques o malware en la red" (Hindy *et al.*, 2006).

Muchos investigadores se han centrado en abordar este problema. Ejemplo de ello son los autores de (Bilge y Dumitras, 2012), que analizaron el impacto de los ataques de día cero en el mundo real. Llegaron a descubrir que los ataques de día cero son más frecuentes de lo que se sospechaba, demostrando que de sus 18 ataques analizados, 11 eran ataques de día cero. Además, demostraron que los ataques de día cero pueden existir durante un largo período de tiempo (un promedio de 10 meses) antes de ser detectados, pudiendo comprometer los sistemas durante largos períodos de tiempo. En (Sharma *et al.*, 2018) se propuso un sistema de diagnóstico distribuido para detectar este tipo de ataques en redes de Internet de las cosas (IoT). En el trabajo de investigación (Sun *et al.*, 2018), los autores proponen un modelo probabilístico bayesiano para detectar rutas de ataque de día cero, visualizando los ataques en una estructura en forma de gráfico. En (Zhou y Pezaros, 2019) se evaluaron diferentes técnicas de ML supervisadas: árbol de decisión, RF, KNN, MLP, análisis discriminante cuadrático y clasificadores GaussianNB. Sin embargo, en este estudio no se aclara cómo se entrenan tales técnicas de ML para ser utilizadas para la detección de ataques desconocidos ni cómo se simulan y detectan los ataques de día cero. En otro trabajo se presentó Kitsune, un NIDS (Network Intrusion Detection System) plug and play capaz de detectar ataques en línea para dispositivos de red simples. El algoritmo central de Kitsune (KitNET) implementó un grupo de AutoEncoders para distinguir patrones de tráfico normales de anormales (Mirsky *et al.*, 2018). En (Shone *et al.*, 2018) se propuso un enfoque que combina el aprendizaje profundo y superficial, que puede analizar correctamente una amplia gama de tráfico de red. Específicamente, su solución aplicaba un autoencoder no simétrico (NDAE) junto con un RF. Desafortunadamente, el modelo no presentó la capacidad de detectar ataques de día cero. La técnica de aprendizaje por transferencia también se puede utilizar para detectar ataques de día cero, como en el caso de (Zhao *et al.*, 2019) que la utilizó para mapear la conexión entre los ataques conocidos y de día cero. En (Sameera y Shashi, 2020) se presentó la aplicación del aprendizaje de transferencia transductiva profunda para detectar ataques de día cero.

Las técnicas de aprendizaje profundo también se utilizan para abordar la detección de malware de día cero. En (Abri *et al.*, 2019) se evaluó la eficacia de varias técnicas de ML (SVM, Naïve Bayes, MLP, árboles de decisión, KNN y RF) para detectar malware de día cero, mientras que el método Deep-Convolutional Generative Adversarial Network (DCGAN) fue propuesto en (Kim, Bu y Cho, 2018).

2. ARQUITECTURA

En esta sección se describe la arquitectura y los métodos utilizados para la solución propuesta. La arquitectura y funcionamiento de esta solución se muestra en la figura 1, donde en primer lugar, se crearán conjuntos de datos para la detección de ataques de día cero y de ataques conocidos. Después de preprocesar ambos conjuntos de datos se pasarán a los modelos de ML para su entrenamiento y validación. Estos modelos permitirán la extracción de eventos de seguridad que serán clasificados en los diferentes tipos de ataques para generar alertas de seguridad para la toma de decisiones.

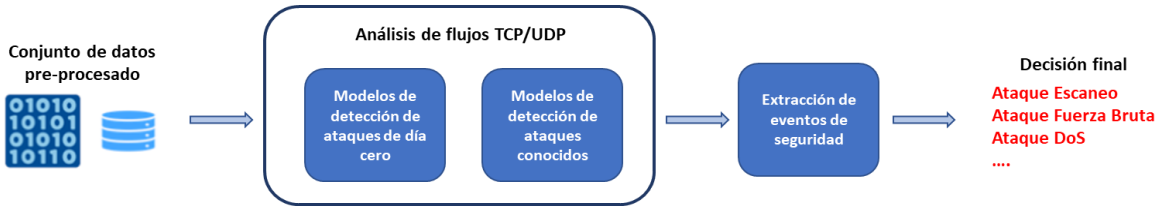

Figura 1. Arquitectura de la solución propuesta para la detección de intrusiones en redes OT

En el análisis realizado en MITRE ATT&CK, se constató que es posible identificar diferentes técnicas de ataque mediante el análisis de información extraída de las cabeceras de los paquetes y los flujos de tráfico de red TCP/UDP, tales como información estadística de protocolos, volúmenes de datos y secuencias temporales. Con la solución propuesta se pretende detectar las técnicas de ataque que se muestran en la Tabla 1, según el modo de amenazas propuesto por MITRE. Donde el "Dominio" indica el framework de referencia utilizado. Por una parte, ATT&CK Matrix for Enterprise se centra en redes IT. Y, sin embargo, el framework ATT&CK para ICS aborda las amenazas tanto a las personas como al entorno físico que se encuentran con las redes ICS (o Sistemas de Control Industrial e infraestructuras críticas).

Tabla 1. Técnicas de ataque que se pretenden detectar con la solución propuesta

Dominio	Técnica	Táctica
Enterprise	T1595 - Escaneo activo	Reconocimiento
	T1498 - Denegación de servicio de Red	Impacto
	T1110 - Fuerza bruta	Acceso a credenciales
ICS	T0840 - Enumeración de conexiones de red	Descubrimiento
	T0814 - Denegación de servicio	Función de inhibición de respuesta
	T0806 - Fuerza bruta E/S	Deterioro de control de procesos

3. MÉTODOS

Los métodos que se utilizarán para el desarrollo de esta propuesta se describen en esta sección. Se dan detalles de los conjuntos de datos en los que se basa el entrenamiento y validación de los modelos de aprendizaje automático, y cuáles son los modelos candidatos para probar el sistema.

3.1 Conjunto de Datos

Como se muestra en la figura 1, la herramienta se basa en el análisis de flujos TCP/UDP. La extracción y generación de los flujos bidireccionales de tráfico de red se realiza mediante la herramienta CICFlowMeter (*Applications | Research | Canadian Institute for Cybersecurity | UNB*, sin fecha) desarrollada por el Canadian Institute for Cybersecurity. Se generan dos conjuntos de datos para la detección de ataques de día cero. Uno contiene datos de comportamiento de referencia o normal, que almacena los flujos capturados durante varios días en los que el sistema opera de forma habitual y sin incidencias reseñables, y el otro es un conjunto de datos de evaluación que contiene todos los flujos registrados en los días en los que se realicen ciertos ataques. Para la detección de ataques conocidos, se crea un conjunto de datos que incluirá tanto flujos capturados con el sistema operando de forma habitual como flujos registrados bajo situaciones de ataque. En este caso, se realiza un muestreo estratificado proporcional a este conjunto de datos para obtener los conjuntos de datos de entrenamiento y de evaluación.

3.2 Modelos de Aprendizaje Automático

Se pretende diseñar y desarrollar un sistema de detección evolutivo que aborde la detección tanto de ataques conocidos como de ataques de día cero, para ello se han estudiado diferentes técnicas para su entrenamiento y prueba según el tipo de ataque a detectar.

En el caso de detección de ataques conocidos se aplicarán y evaluarán técnicas de aprendizaje automático y aprendizaje profundo supervisadas, con el objetivo de apoyar la detección de ataques conocidos y reducir las altas tasas de falsos positivos. Las técnicas de ML y aprendizaje profundo estudiadas en este sentido son: Naïve Bayes, RF y SVM, MLP y LSTM por sus resultados relevantes en la literatura. Así mismo, para la detección de ataques de día cero se aplicarán métodos de aprendizaje no supervisado que han obtenido tasas de aciertos significativas en trabajos similares encontrados en la literatura, tales como: Rango intercuartílico IQR, DBSCAN (Density-Based Spatial clustering of applications with noise), covarianza robusta o envolvente elíptica, One-Class SVM, Isolation Forest, Local Outlier Factor, AutoEncoders y redes neuronales de aprendizaje no supervisado. En este aspecto, a día de hoy, podemos destacar como una gran contribución de

este trabajo el uso de la mayoría de estas técnicas de aprendizaje no supervisado en la fase de preprocesamiento de los datos de comportamiento de referencia o normal, para así ser capaces de detectar los valores atípicos más ruidosos. De forma que se mejora sustancialmente el rendimiento de los modelos de detección de ataques de día cero.

4. CONCLUSIONES

Este es un trabajo de investigación en curso, que propone una solución novedosa para detectar intrusiones en la infraestructura de una organización industrial, a través del análisis de datos de la red de comunicaciones donde se encuentran elementos y sistemas con necesidades específicas que pueden beneficiarse de una aproximación basada en aprendizaje automático, principalmente en las redes OT, donde se ubican los sistemas de control industrial, y permitir así la detección de anomalías en su comportamiento.

Para ello, deben superarse varios desafíos: 1) Para detectar intrusiones es necesario analizar un gran volumen de datos en tiempo real, por lo que las técnicas de detección de anomalías a utilizar deberán ser computacionalmente eficientes. 2) Reducir el número de falsas alarmas emitidas, ya que se corre el riesgo de desbordar a los analistas de seguridad en su tarea diaria. 3) Los conjuntos de datos de entrenamiento suelen estar desequilibrados, dado a que en este dominio suele haber disponibles una gran cantidad de muestras de comportamiento normal o habitual y carencia de un número significativo de muestras de intrusiones.

Desde el punto de vista holístico esta solución puede considerarse como un sistema evolutivo por la posibilidad de mejorar su rendimiento incorporando retroalimentación de los ingenieros de ciberseguridad sobre nuevos incidentes. Además, tendrá la posibilidad de ser complementada con otros sistemas de análisis de ciberseguridad como la detección de cadenas de bloques de ataque por estar alineada con el modo de amenazas propuesto por MITRE.

Como líneas de trabajo futuro nos proponemos la implementación y prueba de la solución presentada en este artículo. Así como, la comparación de los resultados con otros sistemas de detección de intrusiones y la puesta en práctica de dicho sistema para su evolución y perfeccionamiento.

AGRADECIMIENTOS

Esta investigación ha sido apoyada por el "Centro para el Desarrollo Tecnológico Industrial (CDTI)" del Ministerio de Ciencia e Innovación en el marco del proyecto "Tecnologías para la seguridad de las relaciones digitales en un mundo hiperconectado (Secureworld)"

REFERENCIAS

Abri, F. *et al.* (2019) «The Performance of Machine and Deep Learning Classifiers in Detecting Zero-Day Vulnerabilities». Disponible en: http://arxiv.org/abs/1911.09586.

Acarali, D. *et al.* (2022) «Modelling smart grid IT-OT dependencies for DDoS impact propagation», *Computers & Security*, 112, p. 102528. doi: 10.1016/j.cose.2021.102528.

Applications | Research | Canadian Institute for Cybersecurity | UNB (sin fecha). Disponible en: https://www.unb.ca/cic/research/applications.html (Accedido: 14 de mayo de 2022).

Bilge, L. y Dumitras, T. (2012) «Before we knew it: an empirical study of zero-day attacks in the real world», en *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12*. New York, New York, USA: ACM Press, p. 833. doi: 10.1145/2382196.2382284.

Hindy, H. *et al.* (2006) «Towards an effective zero-day attack detection using outlier-based deep learning techniques».

Hodo, E. *et al.* (2017) «Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey», en. doi: 10.48550/arxiv.1701.02145.

Kim, J.-Y., Bu, S.-J. y Cho, S.-B. (2018) «Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders», *Information Sciences*, 460-461, pp. 83-102. doi: 10.1016/j.ins.2018.04.092.

De La Torre Parra, G., Rad, P. y Choo, K.-K. R. (2019) «Implementation of deep packet inspection in smart grids and industrial Internet of Things: Challenges and opportunities», *Journal of Network and Computer Applications*, 135, pp. 32-46. doi: 10.1016/j.jnca.2019.02.022.

Liu, C., Wang, J. y Chen, X. (2022) «Threat intelligence ATT&CK extraction based on the attention transformer hierarchical recurrent neural network», *Applied Soft Computing*, 122, p. 108826. doi: 10.1016/j.asoc.2022.108826.

Mashima, D. (2022) «MITRE ATT&CK Based Evaluation on In-Network Deception Technology for Modernized Electrical Substation Systems», *Sustainability*, 14(3), p. 1256. doi: 10.3390/su14031256.

Min, E. *et al.* (2018) «SU-IDS: A Semi-supervised and Unsupervised Framework for Network Intrusion Detection», en, pp. 322-334. doi: 10.1007/978-3-030-00012-7_30.

Mirsky, Y. *et al.* (2018) «Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection». Disponible en: http://arxiv.org/abs/1802.09089.

MITRE ATT&CK® (2015). Disponible en: https://attack.mitre.org/ (Accedido: 12 de mayo de 2022).

Mohamed, N. A., Jantan, A. y Abiodun, O. I. (2018) «Protect governments, and organizations infrastructure against cyber terrorism (mitigation and stop of server message block (SMB) remote code execution attack)», *International Journal of Engineering Research and Technology*, 11(2), pp. 261-272.

Sameera, N. y Shashi, M. (2020) «Deep transductive transfer learning framework for zero-day attack detection», *ICT Express*, 6(4), pp. 361-367. doi: 10.1016/j.icte.2020.03.003.

Sharafaldin, I., Habibi Lashkari, A. y Ghorbani, A. A. (2018) «Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization», en *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. SCITEPRESS - Science and Technology Publications, pp. 108-116. doi: 10.5220/0006639801080116.

Sharma, V. *et al.* (2018) «A framework for mitigating zero-day attacks in IoT», pp. 1-4. Disponible en: http://arxiv.org/abs/1804.05549.

Shin, Y. *et al.* (2022) «Focusing on the Weakest Link: A Similarity Analysis on Phishing Campaigns Based on the ATT&CK Matrix», *Security and Communication Networks*. Editado por I. You, 2022, pp. 1-12. doi: 10.1155/2022/1699657.

Shone, N. *et al.* (2018) «A Deep Learning Approach to Network Intrusion Detection», *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), pp. 41-50. doi: 10.1109/TETCI.2017.2772792.

Sun, X. *et al.* (2018) «Using Bayesian Networks for Probabilistic Identification of Zero-Day Attack Paths», *IEEE Transactions on Information Forensics and Security*, 13(10), pp. 2506-2521. doi: 10.1109/TIFS.2018.2821095.

Tidjon, L. N., Frappier, M. y Mammar, A. (2019) «Intrusion Detection Systems: A Cross-Domain Overview», *IEEE Communications Surveys & Tutorials*, 21(4), pp. 3639-3681. doi: 10.1109/COMST.2019.2922584.

Vijayanand, R., Devaraj, D. y Kannapiran, B. (2018) «Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection», *Computers & Security*, 77, pp. 304-314. doi: 10.1016/j.cose.2018.04.010.

Watson, G. (2018) «A Comparison of Header and Deep Packet Features when Detecting Network Intrusions». doi: 10.13016/M2G737680.

Xiong, W. *et al.* (2022) «Cyber security threat modeling based on the MITRE Enterprise ATT&CK Matrix», *Software and Systems Modeling*, 21(1), pp. 157-177. doi: 10.1007/s10270-021-00898-7.

Zhao, J. *et al.* (2019) «Transfer learning for detecting unknown network attacks», *EURASIP Journal on Information Security*, 2019(1), p. 1. doi: 10.1186/s13635-019-0084-4.

Zhou, Q. y Pezaros, D. (2019) «Evaluation of Machine Learning Classifiers for Zero-Day Intrusion Detection -- An Analysis on CIC-AWS-2018 dataset». doi: 10.48550/arXiv.1905.03685.

Zhu, J. *et al.* (2017) «Mechanism of situation element acquisition based on deep auto-encoder network in wireless sensor networks», *International Journal of Distributed Sensor Networks*, 13(3), p. 155014771769962. doi: 10.1177/1550147717699625.