Análisis de la Aplicabilidad del Machine Learning para la Detección, Prevención y Mitigación de Malware en Organizaciones.

Javier Alexander Anaya Moreno

Director

Hernando José Peña Hidalgo

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Maestría en Ciberseguridad

Bogotá D.C., 2026

Dedicatorias

A mi Madre querida, todo esto es para ti. Tu amor, tu sacrificio y ese ejemplo inquebrantable de no rendirte nunca fueron, son y serán el pilar de mi vida. Gracias por creer en mí incluso cuando yo dudaba. Cada cosa que aprendí sobre el valor del trabajo, la humildad y la constancia, y que hoy me permite celebrar este logro, te lo debo a ti y a tu esfuerzo incansable.

A mi lado, en cada paso, mi pareja, gracias por ser mi refugio incondicional. En este camino, a veces agotador, tu apoyo, tu paciencia infinita y tu comprensión fueron el motor que me mantuvo en pie. Tu presencia constante y tus palabras de aliento me salvaron en los momentos más duros.

Y a mis hijos, mi motor y mi inspiración, ustedes son la razón de fondo de todo. Cada esfuerzo, cada noche sin dormir y cada pequeño avance en este trabajo, tiene un único propósito, ustedes. Su amor, su inocencia y su fortaleza me dieron el impulso final para terminar esta meta con la certeza de que estamos construyendo un futuro mejor para nuestra familia.

Este trabajo es de ustedes. Con todo mi corazón y una gratitud que no tiene fin.

Agradecimientos

Mi casa de estudios quiero darles las gracias de corazón a la Universidad Nacional Abierta y a Distancia (UNAD). Más que una institución, fue el espacio que me permitió transformarme: no solo aprendí sobre mi carrera, sino que crecí muchísimo como persona y como profesional. Su modelo, tan flexible e inclusivo, me dio las bases y las herramientas perfectas para poder llegar hasta aquí.

A mis tutores, mi gratitud es enorme para todos los docentes que estuvieron conmigo en el camino. Su compromiso, esa orientación tan valiosa y su dedicación constante no solo me enseñaron la materia, sino que me ayudaron a desarrollar mi mente, a pensar de forma crítica y a investigar. Cada tutoría, cada comentario y cada retroalimentación fue un regalo que trascendió el aula y se convirtió en una pieza clave de este logro.

Una mención especial a mis guías del proyecto, un agradecimiento muy especial a los tutores y asesores de este proyecto. Gracias por su guía experta, por la paciencia que me tuvieron y por ese acompañamiento tan riguroso. Su apoyo fue absolutamente decisivo para tomar una simple idea y convertirla en la investigación sólida y aplicable que ven hoy.

Finalmente, gracias a toda la comunidad académica de la UNAD. Gracias por crear un ambiente tan colaborativo, respetuoso y comprometido con el conocimiento. Este trabajo no es solo mío; es el reflejo del esfuerzo colectivo de todos ustedes, que, desde sus diferentes roles, aportaron inmensamente a mi formación y me ayudaron a alcanzar esta gran meta.

Resumen

La Inteligencia Artificial (IA) se ha convertido en una aliada fundamental en la protección frente al malware, aportando nuevas formas de detectar, analizar y responder a las amenazas digitales con una rapidez y precisión sin precedentes. En un mundo donde los ciberdelincuentes desarrollan ataques cada vez más sofisticados y disponen de las mismas herramientas avanzadas que los equipos de seguridad, la IA representa una respuesta inteligente y adaptable.

Hoy, las organizaciones emplean la IA para fortalecer sus sistemas de ciberseguridad, dándoles la capacidad de aprender, anticiparse y reaccionar en tiempo real. Gracias a técnicas como el Machine Learning y el Deep Learning, los sistemas pueden analizar enormes volúmenes de datos para descubrir patrones ocultos y comportamientos que delatan actividades sospechosas o maliciosas.

El análisis de malware basado en IA combina métodos de aprendizaje supervisado donde los algoritmos se entrenan con ejemplos previamente clasificados como seguros o dañinos y no supervisado, que permite a los sistemas identificar anomalías sin instrucciones previas. Además, la IA automatiza tareas críticas como el aislamiento de equipos comprometidos o la ejecución inmediata de protocolos de corrección, lo que mejora considerablemente los tiempos de respuesta y reduce el impacto de los ataques.

No obstante, esta misma tecnología supone nuevos retos. Los atacantes también emplean la IA para ocultar sus rastros, adaptar su comportamiento y evadir los controles de seguridad. Esta dualidad revela una verdad ineludible: el poder de la IA exige un compromiso constante con la innovación, la ética y la vigilancia.

En este contexto, el presente estudio analiza el papel de la Inteligencia Artificial en la ciberseguridad como un paradigma emergente que está revolucionando la forma en que

concebimos la defensa digital. Aunque aún se encuentra en evolución, la IA ya demuestra su enorme potencial para transformar las estrategias de protección, convirtiéndose no solo en una herramienta tecnológica, sino en el núcleo de una nueva era de seguridad proactiva y resiliente.

Palabras clave**:** Análisis, Deep Learning, Inteligencia Artificial, Machine Learning, Malware, Ransonware.

Abstract

Artificial Intelligence (AI) has become a fundamental ally in malware protection, providing new ways to detect, analyze, and respond to digital threats with unprecedented speed and precision. In a world where cybercriminals develop increasingly sophisticated attacks and have the same advanced tools as security teams, AI represents an intelligent and adaptable response.

Today, organizations employ AI to strengthen their cybersecurity systems, giving them the ability to learn, anticipate, and react in real time. Thanks to techniques such as Machine Learning and Deep Learning, systems can analyze huge volumes of data to uncover hidden patterns and behaviors that reveal suspicious or malicious activity.

AI-based malware analysis combines supervised learning methods, where algorithms are trained with examples previously classified as safe or harmful, and unsupervised learning, allowing systems to identify anomalies without prior instruction. Furthermore, AI automates critical tasks such as isolating compromised devices or immediately executing remediation protocols, significantly improving response times and reducing the impact of attacks.

However, this same technology poses new challenges. Attackers also use AI to hide their tracks, adapt their behavior, and evade security controls. This duality reveals an inescapable truth: the power of AI demands a constant commitment to innovation, ethics, and vigilance.

In this context, this study analyzes the role of Artificial Intelligence in cybersecurity as an emerging paradigm that is revolutionizing the way we think about digital defense. Although still evolving, AI already demonstrates its enormous potential to transform protection strategies, becoming not just a technological tool, but the core of a new era of proactive and resilient security.

Keywords**:** Analysis, Deep Learning, Artificial Intelligence, Machine Learning, Malware, Ransomware.

Tabla de contenido

Introducción 16

Planteamiento del problema 18

Formulación del problema 19

Pregunta problema 19

Justificación 20

Objetivos 22

Objetivo general 22

Objetivos específicos 22

Marco Referencial 23

Antecedentes 23

Marco Legal 28

Marco Conceptual 32

Marco teórico 47

Diseño metodológico 63

Revisión sistemática sobre el uso del Machine Learning en el análisis de malware 65

Amenazas Cibernéticas Emergentes 71

Aplicaciones de Aprendizaje Automático en Ciberseguridad 74

Fundamentos y Herramientas de Aprendizaje Automático 76

Casos Relevantes y Estadísticas 77

Investigación y Desarrollo en Ciberseguridad 78

Análisis de la Revisión Sistemática 80

Estrategias de análisis de malware basadas en ML: evaluación estática, dinámica e híbrida y su aplicabilidad en entornos organizacionales 85

Análisis Estático 85

Análisis Dinámico 88

Análisis Híbrido 89

Aplicabilidad y efectividad en la organización 95

Casos reales y simulados del uso de Machine Learning en detección y mitigación de malware 97

Diseño de modelos predictivos basados en Machine Learning para detección de malware con calidad, representatividad y equilibrio 105

Mejora de la calidad y representatividad del dataset 106

Normalización contextual y selección de características 107

Selección de características orientadas a amenazas modernas 109

Propuesta de pipeline predictivo mejorado para la detección de malware 110

Diseño para la Construcción de un Sistema Predictivo de Detección de Malware Basado en Machine Learning 123

Validación adaptativa de modelos de detección de malware 131

Marco conceptual para la validación adaptativa 132

Protocolo experimental propuesto 134

Validación de Modelos Predictivos en Detección de Malware 157

Discusión 164

Conclusiones 167

Recomendaciones 170

Referencias Bibliográficas 172

Apéndices 181

Lista de Tablas

Tabla 1 *Resumen artículos seleccionados* 69

Tabla 2 *Comparativa de casos* 101

Tabla 3 *Dataset simulado* 120

Lista de Figuras

Figura 1 *La Inteligencia Artificial, el Machine Learning y el Deep Learning* 35

Figura 2 *Selección de artículos.* 64

Figura 3 *Resumen de la metodología RSL* 66

Figura 4 *Palabras clave identificadas.* 67

Figura 5 *Descripción general* 68

Figura 6 *Comparativa de precisión entre algoritmos de ML*..... 75

Figura 7 *Aplicaciones de Aprendizaje Automático en Ciberseguridad.* 79

Figura 8 *Aplicabilidad en organizaciones.* 93

Figura 9 *Comparación de los análisis* 94

Figura 10 *Desafíos en el Análisis de Malware.* 103

Figura 11 *Diagrama visual del pipeline.* 113

Figura 12 *Tipo de malware.* 123

Figura 13 *Construcción de un Sistema Predictivo de Detección de Malware.*..... 128

Figura 14 *Preparación y recolección de datos de modelos predictivo.* 136

Figura 15 *Entrenamiento Inicial para la validación adaptativa.* 139

Figura 16 *Laboratorio de modelos predictivos de malware.* 151

Figura 17 *Pruebas en pre-producción.* 153

Figura 18 *Proceso de Actualización Incremental y Control de Calidad de Modelos de Detección de Malware.* 156

Glosario

Anti-debugging: En el contexto de malware o análisis forense, se refiere a mecanismos que el software malicioso implementa para dificultar o impedir que sea depurado, analizado o se ejecute en un entorno controlado (sandbox/VM).

Análisis de comportamiento: Método de detección de amenazas que se basa en identificar actividades inusuales o sospechosas en el comportamiento de usuarios o sistemas, en lugar de depender únicamente de firmas conocidas de malware.

Aumentación: Se refiere al mismo proceso: incrementar la variedad/representatividad del conjunto de entrenamiento mediante transformaciones controladas de las muestras originales.

APT (Amenaza Persistente Avanzada): Ataque cibernético prolongado y dirigido que busca infiltrarse en una red para robar información de forma continua y sin ser detectado.

Attribute-Based Encryption (ABE): La encriptación basada en atributos es un enfoque criptográfico que va más allá del tradicional cifrado con clave pública.

Clasificación supervisada: Técnica de ML en la que el modelo se entrena con datos etiquetados para aprender a predecir la categoría o clase de nuevas entradas de datos.

Class Activation Maps (CAM): Se trata de una técnica de visualización asociada a redes neuronales convolucionales (CNN) que permite identificar qué partes de una imagen fueron más relevantes para que la red tomara una decisión determinada.

Concept Drift: Se refiere al fenómeno por el cual, con el tiempo, la relación entre los datos de entrada y la variable objetivo cambia.

Dataset: Conjunto de datos recopilados y organizados para su análisis, utilizado en el entrenamiento y evaluación de modelos de ML y DL en la detección de malware.

Data Augmentation: Consiste en ampliar artificialmente un conjunto de datos original aplicando transformaciones, por ejemplo, rotaciones, recortes, escalados, adición de ruido que conservan la etiqueta de las muestras.

Deep Learning (DL): Subcampo del ML que emplea redes neuronales artificiales con múltiples capas para modelar y entender patrones complejos en grandes volúmenes de datos.

Drift-Aware Retraining: Es una estrategia que reconoce la existencia del concept drift y activa el reentrenamiento del modelo cuando se detectan caídas de rendimiento, desviaciones estadísticamente significativas o aparición de nuevas familias de amenazas.

Explainable Artificial Intelligence (XAI): Agrupa técnicas y métodos cuyo objetivo es hacer que los modelos de IA no sólo "funcionen", sino que también puedan explicar por qué toman determinadas decisiones.

Feature Elimination: Es la técnica mediante la cual se eliminan del conjunto de variables o "features" aquellas que aportan poco valor al modelo o que introducen ruido, redundancia o sesgos.

Ingeniería de características (Feature Engineering): Proceso de selección y transformación de variables relevantes a partir de datos brutos para mejorar el rendimiento de los modelos de ML.

Inteligencia Artificial (IA): es una rama de la informática que se centra en el desarrollo de sistemas y programas capaces de realizar tareas que normalmente requieren inteligencia humana. Estas tareas incluyen el aprendizaje, el razonamiento, la resolución de problemas, el reconocimiento de patrones, la comprensión del lenguaje natural y la toma de decisiones.

IoT: describe la interconexión de objetos físicos a través de internet, permitiéndoles recoger, enviar y recibir datos sin intervención humana directa.

Machine Learning (ML): Rama de la IA que utiliza algoritmos para analizar datos y aprender de ellos, permitiendo a los sistemas mejorar su rendimiento en tareas específicas sin ser programados explícitamente.

Malware: Software malicioso diseñado para dañar, interrumpir o acceder sin autorización a sistemas informáticos. Incluye virus, gusanos, troyanos, ransomware y spyware.

Phishing: Técnica de ingeniería social que busca engañar a las personas para que revelen información confidencial, como contraseñas o datos bancarios, mediante correos electrónicos o sitios web falsos.

Random Forest: Es un algoritmo de aprendizaje supervisado que agrupa varios árboles de decisión un "bosque" para mejorar la precisión y reducir el sobreajuste de cada árbol individual.

Ransomware: Tipo de malware que cifra los archivos de la víctima y exige un rescate para restaurar el acceso a los datos.

Red Neuronal Convolucional (CNN): Tipo de red neuronal profunda especialmente eficaz en el procesamiento de datos con una estructura en forma de cuadrícula, como imágenes, y utilizada en la detección de malware mediante la identificación de patrones en los datos.

Red Neuronal Recurrente (RNN): Red neuronal que utiliza conexiones recurrentes para procesar secuencias de datos, permitiendo el análisis de información temporal o secuencial, útil en la detección de comportamientos maliciosos en flujos de datos.

Rootkit: Conjunto de herramientas que permiten el acceso no autorizado a un sistema informático y ocultan su presencia, facilitando el control remoto del sistema comprometido.

Spyware: Software que recopila información sobre una persona u organización sin su conocimiento, a menudo con fines maliciosos como el robo de datos personales o corporativos.

SOC: Centro de Operaciones de Seguridad es una unidad centralizada dentro de una organización encargada de monitorear, detectar, analizar y responder a incidentes de ciberseguridad en tiempo real.

Zero-day: Vulnerabilidad desconocida para los desarrolladores del software que es explotada por atacantes antes de que se publique una solución o parche de seguridad.

Introducción

En el panorama digital actual, las organizaciones enfrentan una creciente sofisticación y volumen de amenazas cibernéticas, siendo el malware una de las más persistentes y dañinas. Desde virus y gusanos hasta ransomware y spyware, el malware puede comprometer la confidencialidad, integridad y disponibilidad de la información y los sistemas, generando pérdidas financieras significativas, daño reputacional e interrupción de las operaciones. Como es mencionado por (Al-rimy et al., 2018) y (Razgallah et al., 2021), la creciente sofisticación del malware, caracterizada por técnicas de evasión avanzadas y la velocidad exponencial con la que emergen nuevas variantes, representa un reto crítico y constante para la ciberseguridad en las organizaciones.

En este contexto dinámico y amenazante, en el estudio de (Gibert et al., 2020) (Ucci et al., 2019) el ML se posiciona como una herramienta transformadora con el potencial de revolucionar la detección, clasificación y mitigación de estas amenazas persistentes. Con apoyo de (Gibert et al., 2020) y (Kalambe et al., 2025), la inherente capacidad de los algoritmos del ML para aprender de vastos volúmenes de datos heterogéneos y adaptarse proactivamente a patrones emergentes, incluso aquellos desconocidos para las defensas tradicionales basadas en firmas, ofrece una solución prometedora y cada vez más necesaria para contrarrestar la creciente complejidad de los ataques informáticos.

Diversos estudios menciona (Gibert et al., 2020) de manera concluyente que la integración estratégica de técnicas avanzadas de ML y DL permite la detección temprana y precisa de actividades maliciosas, incluyendo variantes polimórficas y metamórficas que eluden las defensas convencionales, mejorando sustancialmente tanto la eficiencia operativa de los equipos de seguridad al automatizar tareas de análisis y priorización, como la resiliencia general de los sistemas y la infraestructura organizacional frente a las embestidas cibernéticas.

El malware ha evolucionado de amenazas simples a campañas sofisticadas como BlackCat (ransomware-as-a-service) y Emotet (malware polimórfico). Según en el estudio realizado por Al-Rimy (2018), el 67% de las organizaciones afectadas por ransomware reportan pérdidas superiores a $1 millón de dólares, mientras que el 45% enfrenta interrupciones operativas críticas. Las soluciones tradicionales basadas en firmas son insuficientes frente a técnicas evasivas como la ofuscación dinámica y el cifrado homomórfico.

Planteamiento del problema

El malware en la actualidad es una de las formas que los ciberdelincuentes utilizan para realizar sus diferentes ataques con las diferentes técnicas que son utilizadas, como lo menciona (Anand et al., 2022) "*El ransomware que evade la detección pasa a la fase de ejecución después del acceso inicial y la instalación*", con lo anterior se observa uno de los tipos de malware que sigue evolucionando para que no sean detectados generando inconvenientes a las organizaciones para lograr así su detección y posterior erradicación.

Ya que las organizaciones dentro de sus infraestructuras tecnológicas no están contemplando los daños que pueden ocasionar estos ataques de malware con las diferentes técnicas utilizadas, en el estudio que menciona (Tafur-Arciniegas et al., 2023) "*Los mejores clasificadores para identificar claramente el ransomware del software benigno fueron Random Forest y SVM con una puntuación f1 del 86% y una puntuación f1 del 82%, así como un 85% en precisión general para Random Forest*", es así que se encuentra para clasificar los diferentes ransomware son Random Forest y SVM haciendo que su efectividad sea alta y con el apoyo del ML y DL se puede fortalecer y eficaz con estas clasificaciones.

De acuerdo con lo anterior, se evidencia que la falta de un análisis adecuado del malware, proveniente de los diversos vectores de entrada en las organizaciones, puede generar impactos significativos como la degradación del rendimiento de la red, el robo de información sensible y la indisponibilidad de los servicios. Esta situación, sumada a una respuesta tardía frente a los ataques, incrementa el riesgo de afectaciones reputacionales, pérdida de oportunidades de negocio y consecuencias económicas derivadas de sanciones o pagos asociados a incidentes como el secuestro de la información. En este contexto, la incorporación de técnicas basadas en IA, particularmente mediante el uso de ML y DL, permite fortalecer las capacidades de detección,

análisis y respuesta, facilitando la identificación temprana de amenazas y la implementación de medidas correctivas que contribuyan a proteger la infraestructura tecnológica y la información organizacional.

Formulación del problema

La falta de un análisis adecuado de las muestras de malware que ingresan a las redes corporativas ocasiona ralentizaciones en el tráfico, indisponibilidad de servicios, filtración de datos confidenciales y, en última instancia, la pérdida de confianza de clientes y socios de negocio, además de las sanciones económicas asociadas al incumplimiento de normativas de protección de datos frente a este escenario, es imperativo desarrollar estrategias de detección y mitigación en donde el ML y DL puedan incorporar mecanismos de retroalimentación y actualización continua de sus propios modelos, permitiendo a las organizaciones responder de manera ágil y efectiva ante la emergencia de nuevas amenazas y reduciendo drásticamente las tasas de falsos positivos y negativos en sus sistemas de ciberseguridad.

En consecuencia, el planteamiento central de este estudio se fundamenta en la necesidad de diseñar y validar modelos predictivos capaces de clasificar, mitigar el malware y las diferentes familias con alta precisión y adaptabilidad, garantizando así la resiliencia de las infraestructuras organizacionales frente a los ciberataques más avanzados.

Pregunta problema

¿Cómo puede analizarse la aplicabilidad de las técnicas de Machine Learning para la detección, prevención y mitigación del malware en entornos organizacionales?

Justificación

La IA en la actualidad sigue transformando diversas industrias como son los avances en robots semiautónomos que perfeccionan el ensamblaje, innovaciones en fintech impulsadas por IA que rejuvenecen el sector financiero. Al mismo tiempo, el National Health Service en Inglaterra implementa escaneos con IA con el objetivo de mejorar los diagnósticos de fracturas, mientras que IBM genera modelos generativos compactos y eficientes. La computación neuromórfica y la evolución de redes 5G e IoT también están marcando hitos importantes en el campo.

Es así, que la IA está evolucionando en todos los campos, los sistemas informáticos de las organizaciones en la actualidad la ven como una gran herramienta para poder hacer que la seguridad de las redes, servidores y cualquier activo informático de la organización se encuentre dentro de los análisis realizados. La protección de los sistemas informáticos ante ataques con malware es uno de los mayores retos de seguridad, tanto para organizaciones como para individuos. Por eso es tan importante su análisis, porque nos permite estudiar la estructura, el funcionamiento y la interacción de dicho software malicioso, con el propósito de evaluar el daño causado, diseñar técnicas para su defensa y valorar las intenciones y capacidad de un atacante. (Hernández-Pereira, 2022).

Las diferentes técnicas que aplican los atacantes informáticos también son muy sofisticadas ya que también utilizan la IA para su beneficio y así poder lograr su objetivo principal que poder ingresar a las redes y así hacerse de la información de las organizaciones, donde posteriormente solicitan rescate de la información que fue hurtada por los delincuentes. El personal de seguridad de las organizaciones debe trabajar arduamente para defender sus redes y los activos informáticos, capacitándose, buscando nuevos modelos de defensa, identificar las amenazas, catalogarlas y así poder tomar las medidas necesarias.

Para reducir esta problemática es fundamental, primero la implementación de herramientas para la identificación del software malicioso y después para su clasificación. Para ello se puede hacer uso de algoritmos de ML, una de las herramientas más eficientes y empleadas para dicha tarea. Este tipo de aprendizaje se configura como un conjunto de técnicas relacionadas con la IA que permiten a los sistemas resolver de manera autónoma gran cantidad de cuestiones, sin necesidad alguna de intervención humana (Hernández-Pereira, 2022).

Objetivos

Objetivo general

Analizar el uso del Machine Learning en la detección y mitigación del malware en las organizaciones mediante una revisión documental y de casos prácticos, para proponer soluciones innovadoras de ciberseguridad.

Objetivos específicos

Examinar la documentación existente sobre el uso del Machine Learning en el análisis de malware, mediante una revisión sistemática de literatura que permita identificar tendencias, enfoques innovadores y prácticas destacadas para combatir el malware en las organizaciones.

Evaluar técnicas de análisis estático y dinámico utilizadas en la identificación y caracterización del comportamiento del malware, determinando su aplicabilidad y efectividad en entornos organizacionales mediante la revisión de casos reales y simulados bajo un marco conceptual, donde sea posible establecer criterios técnicos y estratégicos que orienten la selección e implementación de métodos de análisis adecuados.

Diseñar modelos predictivos basados en Marchine Learning para la detección de malware que mejoren la calidad, representatividad y equilibrio de los análisis de los datasets, incorporando técnicas de aumento de datos, normalización contextual y selección de características orientadas a patrones de amenazas modernas.

Validar modelos basados en la detección de malware donde se integrarán mecanismos de Machine Learning que permita identificar las nuevas variantes de malware, utilizando retroalimentación de incidentes reales y actualizaciones incrementales de datasets en entornos controlados, que permita a los equipos de seguridad comprender las decisiones del algoritmo, identificando falsos positivos/negativos y priorizar acciones de respuesta.

Marco Referencial

Antecedentes

La actual era digital cada vez es más afectada por diferentes métodos de los ciberataques que arrecian por diversos frentes, el ransomware está siendo el gran protagonista, pero también los ataques a los sistemas de criptomonedas, los de denegación de servicio (DoS y DDoS) o los ataques a la cadena de suministro. La mayoría de estos ataques son la continuidad de otros grandes ciberataques de la década pasada que marcaron el devenir de los actuales. Los ciberataques más relevantes de la última década en el Mundo.

WikiLeaks, creado en 2006 por el australiano Julián Assange, ganó su popularidad en el 2010 cuando se publicaron 251.287 telegramas diplomáticos, intercambiados entre más de 250 embajadas de los Estados Unidos y el Departamento de Estado de los Estados Unidos en Washington. De los aproximadamente 250.000 documentos revelados por este sitio web de WikiLeaks, hay 55.000 cables emitidos desde España o dirigidos a las delegaciones de Estados Unidos. Y casi 40.000 más que mencionan a España en comunicaciones entre terceras partes (Contreras & Contreras, 2023).

En la brecha de datos resultaron comprometidos los nombres, correos electrónicos, datos de acceso y otros datos personales de aproximadamente 77 millones de personas con la cuenta de PlayStation Network (el servicio de PlayStation que permite la compra de juegos online), este servicio dejó de funcionar durante una semana. En ese momento, la empresa japonesa no descartó la posibilidad de que los datos bancarios de los usuarios hubieran sido robados por los ciberdelincuentes (Contreras & Contreras, 2023).

Este ataque tuvo lugar en 2012, pero la magnitud de este se conoció cuatro años más tarde. Dropbox en el 2012 confirmó que los correos electrónicos de los usuarios habían sido expuestos, pero fue en 2016 cuando Leakbase descubrió que también habían sido robadas las contraseñas. En total, aproximadamente 68 millones de usuarios se vieron afectados. Los ciberdelincuentes pudieron entrar en estas cuentas porque uno de los empleados de Dropbox usó su contraseña profesional en LinkedIn, cuando a principios de ese año LinkedIn sufrió un ataque, los crackers tuvieron acceso a la contraseña del empleado y la usaron para acceder a la red interna de Dropbox para poder realizar el robo de la información (Contreras & Contreras, 2023).

Ahora, el gigante estadounidense de venta minorista Target fue objeto de un ataque histórico en 2013 que afectó a más de 70 millones de clientes. Además del robo de información personal como (nombres, direcciones, números de teléfono y correos electrónicos), se presentó al menos 40 millones de víctimas que también vieron cómo les robaban sus datos bancarios. Los ciber criminales ingresaron al sistema de Target a través de un malware PoS, el cual afecto a los dispositivos del punto de venta en este caso, los lectores de tarjetas de crédito/débito y cajas registradoras. El ataque adquirió proporciones aún mayores al estar diseñado para la temporada de compras previas a la Navidad, entre el 27 de noviembre y el 15 de diciembre del 2013 (Contreras & Contreras, 2023).

Por otra parte, en mayo de 2014 eBay emitió un comunicado en el que pedía a sus 145 millones de usuarios que cambiaran su contraseña ya que descubrieron que su red había sido objeto de un ciberataque. Los ciberdelincuentes pudieron ingresar en el sistema de la empresa mediante el acceso no autorizado a las contraseñas de algunos de los empleados, se hicieron con nombres de clientes, contraseñas cifradas, correos electrónicos, direcciones, números de teléfono y fechas

de nacimiento. eBay fue muy criticado por el tiempo que tardó en notificar a sus clientes sobre el incidente que se los presento entre febrero y marzo de 2014 (Contreras & Contreras, 2023).

Por otro lado, la información de 191 millones de votantes estadounidenses, alrededor del 60% de la población, fue expuesta en Internet debido a un error de la empresa de marketing contratada por el Comité Nacional Republicano durante la campaña de Donald Trump. Esto género que los registros de los votantes, incluyendo nombres, direcciones, números de teléfono, fechas, afiliaciones a partidos, fechas de nacimiento, incluso religión y posicionamiento en temas controvertidos, fueran accesibles en la web por estos ciberdelincuentes (Contreras & Contreras, 2023).

En igual forma LeakedSource público un caso donde lo clasificó en su momento como el mayor robo de datos de la historia. Aproximadamente 412 millones de cuentas en la red de sitios para adultos y pornografía Friend Finder fueron expuestas en el mercado negro o Dark web, incluyendo correos electrónicos y contraseñas. Como estos datos estaban asociados a sitios de contenido para adultos, el impacto del ataque también implicó la extorsión y vergüenza de los usuarios implicados (Contreras & Contreras, 2023).

De la misma forma, los medios de comunicación destacaron una noticia donde un Uber tuvieron afectaciones tanto en el número de víctimas afectadas que fueron alrededor de 57 millones, sino también porque pagó cien mil dólares a dos ciberdelincuentes para eliminar los datos robados y ocultar el ciberataque, manteniéndolo en secreto. El ataque tuvo lugar en octubre de 2016 un año antes de su publicación e incluyó la exposición de nombres, correos electrónicos y números de teléfono de 57 millones de clientes en todo el mundo, así como la información personal de 7 millones de conductores de esa empresa de transporte (Contreras & Contreras, 2023).

En 2018, Cambridge Analytica mostró al mundo cómo el robo de datos puede ser usado en política en este caso, para influir en las elecciones presidenciales de Estados Unidos de América en el 2016. Cambridge Analytica es una empresa de análisis de datos que trabajó con el equipo de Donald Trump donde utilizó sin consentimiento la información de 50 millones de perfiles de Facebook para identificar los patrones de comportamiento y gustos de los usuarios para utilizarlos en la difusión de propaganda política (Contreras & Contreras, 2023).

Un año después del escándalo de Cambridge Analytica, Facebook se vio involucrado una vez más en un caso de exposición de datos. Aproximadamente 419 millones de números de teléfono y de identificación de usuario en Facebook fueron almacenados en un servidor online que no estaba protegido por una contraseña. Aunque no son tan sensibles como los datos financieros, los números de teléfono pueden ser utilizados por los crackers para spam, phishing o fraudes asociados a la tarjeta SIM de los celulares. Los Estados Unidos de América, el Reino Unido y Vietnam fueron los países más afectados por este ataque (Contreras & Contreras, 2023).

En los últimos dos años, los ciberdelincuentes secuestraron la información de Sanitas, Salud Total y Audifarma, entre otras instituciones como el INVIMA, exponiendo los datos confidenciales de cerca de 300.000 personas. Estos hechos causaron que cerca de 5.5 millones de afiliados del sistema de salud presentaran problemas para pedir citas, programar procedimientos y reclamar medicamentos, entre otras dificultades (Económica, 2024).

Los efectos de una futura reforma a la salud no es la única preocupación que tiene nerviosos los cerca de 51 millones de usuarios afiliados al sistema. A raíz de los ataques cibernéticos que en el último tiempo recibieron Sanitas, Salud Total y la red de droguerías Audifarma, entre otras entidades, muchos de ellos hoy se preguntan qué tan segura está y en que manos se encuentra su información (Económica, 2024).

Vale recordar que la intrusión de los sistemas que afectó a Sanitas en diciembre de 2022 y que duró 45 días, según la Superintendencia Nacional de Salud (Supersalud), causó la vulneración de las historias clínicas, cédulas, teléfonos y direcciones, entre otros, de aproximadamente 242.000 afiliados. Esto sin contar que también se perjudicó el servicio por más de un mes de 5.5 millones de usuarios (Económica, 2024).

Un mes después, el 22 de enero de 2023, la red de droguerías Audifarma fue víctima de un ataque cibernético que, apagó su infraestructura tecnológica. Y, finalmente, el hecho más reciente, se presentó el pasado 27 de enero de 2024, cuando los sistemas de Salud Total fueron alterados, dejando en jaque sus canales de comunicación y generando traumatismos en la atención que reciben los más de 4.6 millones de usuarios que tiene esta EPS (Económica, 2024).

Sin embargo, estos problemas de seguridad no sólo han afectado a la red de atención del sistema. Autoridades como el INVIMA fueron atacadas en octubre de 2022 cuando su información fue secuestrada, lo que hizo sus plataformas dejarán de operar (Económica, 2024).

Estos incidentes reflejan una seria falencia en la estrategia de ciberseguridad de las instituciones de salud; de hecho, el Índice de Inteligencia de Amenazas X-Force de IBM para 2024, señala que el país durante 2024 ha concentrado el 17 % de los ataques cibernéticos de Latinoamérica, siendo estas entidades las más afectadas. Esta cifra es preocupante, además que se tiene en cuenta que el 51 % de las empresas de esta industria aún no cuentan con un plan robusto de respaldo y de recuperación de datos que les permita responder a las amenazas cibernéticas, como lo evidenció la encuesta Digital Trust Insights de 2024 de PwC. Ante este difícil panorama, la tecnología, a través de diferentes soluciones en la nube, viene transformando las estrategias de ciberseguridad de las compañías del sector salud (Económica, 2024).

De acuerdo con Ricardo González Vargas, Chief Technology Officer (CTO) de Clouxter, “almacenar la información en la nube permite gestionar con mayor seguridad los datos, incluyendo los de los usuarios, las credenciales para pagos y las historias médicas, entre otros. Igualmente, es una excelente alternativa para evitar alojar la información en servidores o datacenters que no cuentan con una seguridad robusta, facilitando la presencia de robos, filtraciones, secuestros y extorciones que afectan directamente la reputación de las empresas” (Económica, 2024).

Marco Legal

El desarrollo de modelos de ML para la detección, prevención y mitigación de malware en organizaciones requiere un marco legal sólido que oriente su implementación responsable. Este marco articula políticas nacionales, normas internacionales y legislación vigente en Colombia, asegurando que los sistemas predictivos operen bajo principios de seguridad, protección de datos, ética y gobernanza tecnológica.

CONPES 3995 – Política Nacional de Confianza y Seguridad Digital

El CONPES 3995 de 2020 establece la hoja de ruta para fortalecer la seguridad digital del país, promoviendo capacidades en análisis avanzado, gestión del riesgo tecnológico y protección de infraestructuras críticas. Para sistemas basados en ML, esta política resalta la necesidad de implementar mecanismos de detección y respuesta temprana ante amenazas, así como adoptar buenas prácticas en vigilancia tecnológica e innovación en ciberseguridad (Departamento Nacional de Planeación, 2020).

CONPES 4144 de 2023 – Política Nacional de Inteligencia Artificial

El CONPES 4144 amplía el enfoque hacia el uso estratégico y responsable de la inteligencia artificial en Colombia, estableciendo lineamientos para su desarrollo ético, seguro y sostenible. En el contexto de la ciberseguridad, esta política impulsa:

- La implementación de sistemas de IA confiables y transparentes.
- La gestión de riesgos asociados a decisiones automatizadas.
- El fortalecimiento de capacidades institucionales para el uso de ML en la protección digital. (Departamento Nacional de Planeación, 2023).

Política Nacional de Inteligencia Artificial – Minciencias (Proyecto de Ley IA)

El documento de lineamientos de IA promovido por Minciencias refuerza la necesidad de establecer un marco regulatorio que garantice el desarrollo responsable de tecnologías basadas en ML. Este plantea:

- Principios de ética, transparencia y explicabilidad en sistemas automatizados.
- Protección de derechos fundamentales frente a decisiones algorítmicas.
- Promoción de la investigación y desarrollo en IA aplicada a sectores estratégicos como la ciberseguridad. (Ministerio de Ciencia Tecnología e Innovación, 2025)

Estrategia Nacional de Seguridad Digital de Colombia

La Estrategia Nacional de Seguridad Digital, en sus actualizaciones vigentes, orienta los esfuerzos del país hacia una mayor madurez digital, promoviendo:

- La adopción de tecnologías emergentes como el ML para fortalecer capacidades defensivas.
- El desarrollo de Centros de Operación de Seguridad (SOC) con procesos automatizados.
- La protección de infraestructuras críticas mediante análisis avanzado y gestión del riesgo. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2023).

Lineamientos de Política Digital – MinTIC

Los lineamientos actualizados del Ministerio TIC refuerzan la transformación digital segura, destacando:

- La integración de analítica avanzada e inteligencia artificial en la gestión pública y privada.
- La necesidad de garantizar la seguridad de la información en entornos digitales.
- La promoción de marcos de gobernanza tecnológica para la adopción de soluciones basadas en datos. (Ministerio de Tecnologías de la Información y las Comunicaciones, 2024)

ISO/IEC 27032:2012 – Directrices para la Ciberseguridad

La norma ISO/IEC 27032:2012 proporciona lineamientos para la protección de ecosistemas digitales, destacando la importancia de prevenir ataques como malware, phishing, botnets o intrusiones avanzadas. Para modelos de ML, esta norma fortalece:

- Prácticas para la gestión de incidentes.
- Coordinación entre actores involucrados en la protección digital.
- Requerimientos de seguridad en sistemas que analizan grandes volúmenes de datos sensibles.(International Organization for Standardization, 2012).

ISO/IEC 27001:2022 – Sistema de Gestión de Seguridad de la Información

La ISO/IEC 27001:2022 es el estándar internacional más reconocido para la gestión de seguridad de la información. En el contexto de la detección de malware con ML, esta norma exige que las organizaciones:

- Gestionen adecuadamente los riesgos asociados a datos y modelos.
- Mantengan la trazabilidad y auditoría de la información procesada.

- Apliquen controles para proteger la confidencialidad, integridad y disponibilidad de los datos utilizados en el entrenamiento. (International Organization for Standardization, 2022).

ISO/IEC 42001:2023 – Sistemas de Gestión para Inteligencia Artificial

La reciente ISO/IEC 42001:2023 introduce lineamientos para el desarrollo ético, transparente y gobernable de sistemas de IA. Es especialmente relevante para modelos de ML orientados a ciberseguridad, ya que establece:

- Prácticas para garantizar la explicabilidad y mitigación de sesgos.
- Controles para monitorear y actualizar modelos predictivos.
- Políticas de rendición de cuentas en sistemas automatizados.(International Organization for Standardization, 2023).

Ley 1273 de 2009 – Delitos Informáticos

La Ley 1273 de 2009 tipifica los delitos relacionados con la seguridad de la información, incluyendo la creación, propagación o uso indebido de malware. Para proyectos que manejan muestras maliciosas, esta ley exige:

- Manejo controlado y seguro de muestras.
- Prevención de cualquier divulgación accidental que pueda constituir delito.
- Responsabilidad penal por conductas que comprometan la integridad de sistemas informáticos. (Congreso de la República de Colombia, 2009).

Ley 1581 de 2012 – Protección de Datos Personales

La Ley 1581 de 2012 regula el tratamiento de datos personales y exige que cualquier sistema que procese datos, incluidos logs o telemetría utilizados para entrenar modelos, cumpla con:

- Principios de seguridad, confidencialidad y autonomía del titular.
- Finalidad específica, autorización y minimización de datos.
- Medidas técnicas adecuadas para evitar accesos no autorizados.(Congreso de la República de Colombia, 2012).

El conjunto normativo analizado proporciona los lineamientos esenciales para la implementación responsable de técnicas de ML en ciberseguridad. La articulación de políticas nacionales, estándares internacionales y legislación sobre delitos informáticos y protección de datos garantiza que los sistemas diseñados no solo sean eficaces, sino también seguros, auditables y éticamente gestionados. En este contexto, la adopción de modelos basados en ML se enmarca en una estructura legal que promueve la confiabilidad, la transparencia y la protección integral de la información organizacional.

Marco Conceptual

Inteligencia Artificial

La IA es un campo de la ciencia relacionado con la creación de computadoras y máquinas que pueden razonar, aprender y actuar de una manera la cual requiere inteligencia humana o se vea involucrado los datos cuya escala excede lo que los seres humanos pueden llegar a analizar.

Adicionalmente es un campo amplio que incluye muchas disciplinas, como la informática, la ingeniería de hardware y software, el análisis y la estadística de datos, la neurociencia, la lingüística abarcando campos como la filosofía y la psicología.

A nivel operativo para el uso organizacional, la IA son un conjunto de tecnologías que se basan en el ML y el DL, porque se usan para los análisis de datos, la categorización de objetos, la generación de predicciones y previsiones, las recomendaciones, el procesamiento de lenguaje

natural, la recuperación inteligente de datos y una infinidad de procesos que son realizados por los seres humanos.

Los tipos de IA que son organizados de varias maneras, según la etapa de desarrollo o la acciones que se están realizando los investigadores, desarrolladores o las personas que deseen evolucionar en el proceso.

- Máquinas reactivas: La IA limitada la cual reacciona a diferentes tipos de estímulos basados en reglas preprogramadas. No usa memoria y, por lo tanto, no puede aprender con datos nuevos.
- Memoria limitada: Se considera que la mayor parte de la IA moderna es de memoria limitada. Esta puede usar la memoria para mejorar con el tiempo mediante el entrenamiento con nuevos datos, por lo general, a través de una red neuronal artificial o por otro modelo de entrenamiento. El DL, es un subconjunto del ML, la cual se considera IA con memoria limitada.
- Teoría de la mente: En la actualidad no existe IA con teoría de la mente, pero se investiga con distintas posibilidades. El término hace referencia a la IA que puede emular la mente humana y tiene capacidades de toma de decisiones similares a la de los seres humanos, lo cual puede reconocer y recordar emociones, reaccionar en situaciones sociales como lo hace los seres humanos.
- Autoconocimiento: Es un paso más de la IA con la teoría de la mente, el concepto de la IA con autoconocimiento describe una máquina mítica que contiene conocimiento de su propia existencia y tiene las capacidades intelectuales y emocionales de los seres humanos. Al igual que la IA con teoría de la mente, la IA con autoconciencia no existe en la actualidad.

Machine Learning

El ML es la ciencia que desarrolla algoritmos y modelos estadísticos que son utilizados en los sistemas de computación con el fin de llevar tareas sin instrucciones explícitas, en vez de establecer patrones o deducciones. Los sistemas de computación utilizan algoritmos de ML en donde procesa grandes cantidades de datos históricos logrando identificar patrones de datos analizados. Esto ayuda a generar los resultados con una precisión a partir de un conjunto de datos de entrada.

Los algoritmos se clasifican en cuatro tipos de aprendizaje distintos en función de la salida esperada y del tipo de entrada.

- ML supervisado
- ML sin supervisar
- Aprendizaje semisupervisado
- ML por refuerzo

Deep Learning

Es un tipo de ML la cual entrena a una computadora para que realice tareas a manera que las hacen los seres humanos, como el reconocimiento del habla, la identificación de imágenes o hacer predicciones. La naturaleza dinámica de los métodos de la DL es:

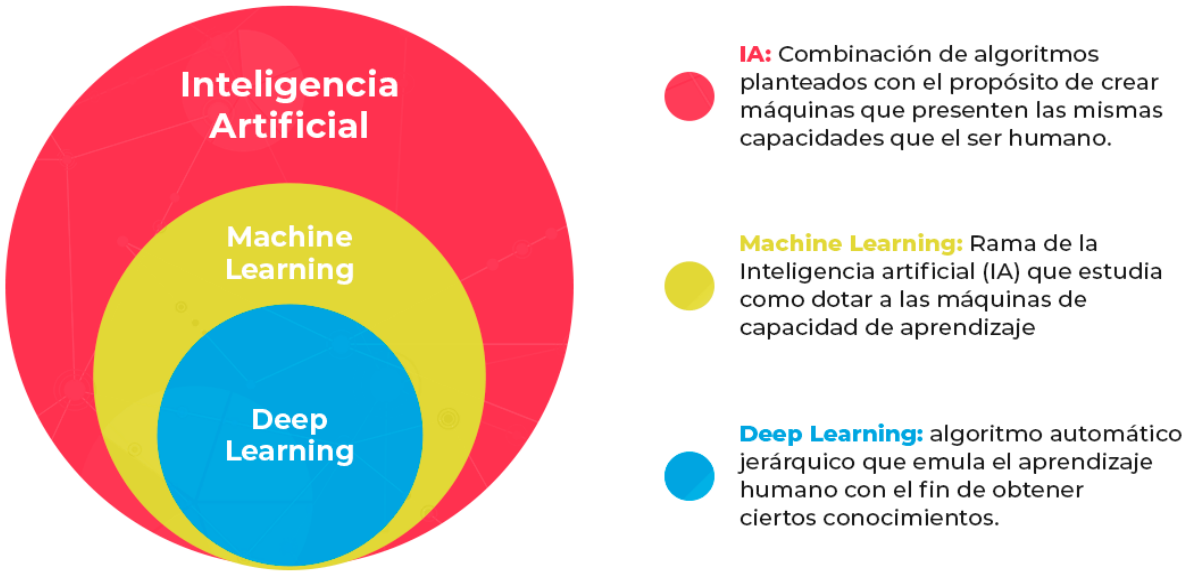
- Su capacidad de mejorar y adaptarse continuamente a cambios en el patrón de información implícito,
- Presenta oportunidades para introducir un comportamiento más dinámico a la analítica.

En la actualidad el enfoque del mercado en las técnicas de DL son las aplicaciones de cómputo cognitivo, pero adicionalmente se encuentra un gran potencial en las aplicaciones

analíticas más tradicionales. Otra oportunidad para ser más eficiente y simplificado en operaciones analíticas existentes. Como, por ejemplo, una empresa realizo experimentos con redes neurales profundas en problemas de transcripción de habla a texto, esto se compara con las técnicas estándares, lo que nos demuestra que el índice de errores en palabras disminuyó más de 10% cuando se aplicaron redes neurales profundas. Adicional a lo anterior se eliminaron cerca de 10 pasos del procesamiento de datos, ingeniería de características y modelado. Como se visualiza en la figura 1 la integración del IA, ML y DL.

Figura 1

La Inteligencia Artificial, el Machine Learning y el Deep Learning



Nota*:* Adaptado de ¿Qué es el Machine Learning? [Fotografía], por Universidad Complutense, (s.f.), (https://www.masterdatascienceucm.com/que-es-machine-learning).

Aprendizaje Supervisado

En el aprendizaje supervisado es aquel que utiliza un conjunto de datos etiquetados que entrenan algoritmos para clasificar datos o predecir resultados con precisión. Un ejemplo de aprendizaje supervisado es predecir los tiempos de vuelo según las horas de mayor actividad en

un aeropuerto, la congestión del tráfico aéreo y las condiciones climáticas (adicionalmente teniendo en cuenta otros parámetros posibles).

Sin embargo, el ser humano tiene que proceder para etiquetar o enumerar el conjunto de datos a fin de adiestrar al modelo sobre cómo estos factores generan afectación en los horarios de los diferentes vuelos. Un modelo supervisado depende bastante de la capacidad de estudiar los resultados para concluir que los diferentes factores climáticos como (lluvia o la nieve) generar los retrasos de los vuelos.

Aprendizaje No Supervisado

Por el lado del aprendizaje no supervisado se utiliza los algoritmos de ML para examinar y agrupar los conjuntos de datos no etiquetados en las bases. Los modelos funcionan sin interferencias humanas constantemente, donde encuentran una estructura de clasificaciones y llegan a ella mediante datos no etiquetados.

La intervención humana es para validar las variables de los resultados que va arrogando los algoritmos. Como ejemplo, cuando una persona compra un electrodoméstico nuevo en línea, el modelo de aprendizaje no supervisado determinará que la persona pertenece a cierto grupo de compradores que adquieren un conjunto de productos. No obstante, es tarea de los analistas de datos validar que el motor de recomendación ofrezca las diferentes opciones para adquirir más electrodomésticos para la cocina.

Clasificación

En la clasificación un problema que utilizan algoritmos para clasificar datos en los segmentos particulares que se han determinado por el analista de datos. Entre los algoritmos de clasificación más comunes son la regresión logística, cierta cantidad de vecinos más cercanos, bosque aleatorio, clasificador bayesiano simple, descenso de gradiente estocástico y árboles de

decisión. Por ejemplo, cotidiano es un algoritmo que ayuda a rechazar correo no deseado (SPAM) en una bandeja de entrada de correo electrónico principal ya sea para una empresa o personal, un algoritmo que permita a los usuarios bloquear o restringir a las personas en las redes sociales.

Regresión

Es un método estadístico y de ML que se basa en algoritmos para medir la relación entre una variable dependiente y una o más variables independientes. En los modelos de regresión, el usuario puede generar predicciones de causa y efecto basadas en numerosos puntos de datos. Como ejemplo, una empresa podría implicar como predecir la trayectoria de crecimiento de los ingresos de publicidad que generan por semana o al mes.

Clustering

Son los datos no etiquetados que se clasifican mediante técnicas de agrupación de clústeres. Esto crea agrupaciones en función de la ubicación, los niveles de ingresos, la edad de los compradores o alguna cualquier otra variable. De ejemplo, una empresa que trabaja en la segmentación de mercado, el algoritmo que utilizaría en la agrupación de clústeres es el de K-medias asignará puntos de datos similares a los grupos que representa un conjunto de parámetros que se establecen.

Análisis de Datos

Es el proceso de exploración, transformación y examinación de datos para lograr identificar las tendencias y los patrones que revelen las perspectivas más importantes y que aumentan la eficiencia para respaldar la toma de decisiones. Una estrategia moderna para el análisis de datos empodera a los sistemas y organizaciones para trabajar a partir de los análisis automatizados en tiempo real de la información garantizando los resultados inmediatos y de gran impacto. Es así, que existen varios tipos diferentes de análisis de datos.

- Análisis descriptivo,
- Análisis de diagnóstico,
- Análisis predictivo,
- Análisis prescriptivo.

Los análisis descriptivos y de diagnóstico permiten a los analistas de datos y a los líderes nivelar el conjunto, estos procesos son bloques de creación que preparan el camino para obtener perspectivas más sofisticados a partir de análisis predictivos y prescriptivos.

La gran varias de casos prácticos en el análisis de datos donde un mundo que prioriza lo digital son casi infinitos, donde la predicción del comportamiento de los clientes a partir de las interacciones omnicanal hasta la anticipación de los cambios en una cadena de suministro debido a catástrofes naturales que se puedan generar.

Detección de Anomalías

Esta examina los puntos de datos específicos y detecta las incidencias menos comunes que parecen ser sospechosas con los diferentes patrones establecidos del comportamiento. La detección de anomalías no es algo novedoso, pero a medida que se incrementa el volumen de datos en la industria el seguimiento manual ya no resulta práctico. Las personas que son más probables que utilicen este proceso son los administradores de seguridad y plataformas, los desarrolladores de aplicaciones y los ingenieros de seguridad.

Una estrategia para la detección de anomalías comienza por la identificación de los diferentes indicadores para el rendimiento de los KPI. Estos indicadores están vinculados al problema empresarial que está resolviendo en el momento.

Adicional se debe comprender las características de los datos, es por eso por lo que surgen las siguientes preguntas, ¿Cómo fluyen hacia la red?, ¿Son continuos o en lote?, ¿De qué puntos

de datos está realizando un seguimiento?, el poder resolver estas preguntas contribuye a generar una estrategia, ya que los datos son una parte fundamental en este proceso.

La detección de anomalías es principalmente importante en las finanzas, la venta minorista y la ciberseguridad, pero deberían todas las empresas tener presente en implementar una solución de detección de anomalías. Ya que estas soluciones proporcionan un medio automatizado para detectar valores extraños perjudiciales y para proteger los datos.

Red Neuronal

Es un método de la IA que les enseña a las computadoras a procesar datos de una manera similar a como lo hace el cerebro del ser humano. Es un tipo de proceso de ML lo que se le llama DL, este utiliza las neuronas interconectadas en las estructuras de capas que se compara al cerebro del ser humano.

El cerebro del ser humano inspira una arquitectura en las redes neuronales. Las células del cerebro del ser humano se llaman neuronas y se forma una red compleja con un alto nivel de interconexión la cual envían señales eléctricas entre sí para ayudar a los humanos a procesar la información. Es así, cómo funciona la red neuronal artificial la cual se forma por neuronas artificiales donde trabajan juntas para resolver los diferentes problemas. Las neuronas artificiales son módulos de software, la cual se llama nodos, las redes neuronales artificiales son programas de software o algoritmos que utilizan los sistemas informáticos para resolver cálculos matemáticos.

Estas redes neuronales pueden apoyar a las computadoras a tomar decisiones inteligentes con asistencia humana limitada. Esto es ya que puede aprender y modelar las relaciones entre los diferentes datos de entrada como de salida que no sean lineales y que sean complejos. Las redes neuronales están presentes en varios casos de uso en diferentes sectores:

- Diagnóstico médico mediante la clasificación de imágenes médicas,

- Marketing orientado mediante el filtrado de redes sociales y el análisis de datos de comportamiento,
- Predicciones financieras mediante el procesamiento de datos históricos de instrumentos financieros,
- Previsión de la carga eléctrica y la demanda de energía,
- Proceso y control de calidad,
- Identificación de compuestos químicos.

Malware

Es un software malicioso que contiene cualquier código de software o programa informático, incluidos ransomware, troyanos y spyware, escrito de manera intencional para dañar los sistemas informáticos de las organizaciones o a sus usuarios.

Los ciberdelincuentes desarrollan y utilizan estos malware para algunas de las siguientes actividades:

- Secuestrar los dispositivos, datos o redes empresariales enteras por grandes sumas de dinero.
- Obtener acceso no autorizado a datos confidenciales o activos digitales de las organizaciones o de los usuarios.
- Robar credenciales de inicio de sesión, números de tarjetas de crédito, propiedad intelectual u otra información valiosa de las personas.
- Interrumpir los sistemas críticos en los que confían las empresas y las entidades gubernamentales.

Tipos de malware. Los cibercriminales o crackers desarrollan constantemente nuevas cepas de malware con nuevas características y funcionalidades generando evolución de los

mismos. Con el pasar del tiempo estas cepas individuales de malware generaron nuevas variantes para así eludir mejor el software o herramienta de seguridad dificultando a los profesionales de la ciberseguridad.

Los crackers en la mayoría de los casos comparten su malware haciéndolo un código abierto o vendiéndolo a otros ciberdelincuentes. Los acuerdos de malware como servicio prevalecen entre los desarrolladores o creadores de ransomware de tal manera los ciberdelincuentes con poca experiencia técnica pueden cosechar los beneficios de los delitos cibernéticos. Las cepas de malware pueden clasificarse en algunos tipos comunes de la siguiente manera:

***Virus de computadoras*.** Los términos "malware" y "virus de computadora" se utilizan a menudo como sinónimos, pero un virus es una técnica de un tipo muy particular de malware. Concretamente, un virus es un código malicioso que secuestra software legítimo para hacer daño y difundir copias de sí mismo.

Los virus no pueden actuar por sí solos. En cambio, estos ocultan fragmentos de su código en otros programas que se ejecutan. Un usuario inicia el programa, el virus alternativamente comienza a ejecutarse. Los virus están diseñados para eliminar datos importantes de los usuarios, interrumpir las operaciones normales de las organizaciones y propagar copias de sí mismos a otros programas en la computadora infectada y pasarlos por la red.

Botnets. Es una red de dispositivos conectados a Internet e infectados con un malware el cual está bajo el control de un hacker. Las botnets pueden incluir PC, dispositivos móviles, dispositivos de Internet de las Cosas, entre otros. Las víctimas normalmente no se dan cuenta que sus dispositivos forman parte de una botnet. Los hackers suelen utilizar botnets para lanzar ataques

DDoS – Denegación de Servicio, que bombardean una red objetivo con bastante tráfico para lograr ralentizar o lograr apagar por completo la red

Criptojackers. Es un tipo malware que toma el control de los dispositivos y los utiliza para extraer criptomonedas, como bitcoin, sin el conocimiento del propietario. La minería de criptomonedas es una tarea extremadamente costosa y que requiere mucha computación. Los ciberdelincuentes se lucran mientras que los usuarios de computadoras infectadas experimentan ralentización de rendimiento y bloqueos constante de las maquinas. Los criptojackers suelen apuntar a la infraestructura de la nube empresarial, lo que les permite obtener más recursos para la criptografía que para atacar computadoras individuales.

Malware sin archivos. Este malware sin archivos es un tipo de ataque que se utiliza en vulnerabilidades de programas de software legítimos, como navegadores web y procesadores de texto, para inyectar el código malicioso directamente en la memoria de las maquinas. Como el código se ejecuta en la memoria, no deja rastro en el disco duro. Como el que utiliza el software legítimo, suele eludir esta detección.

Muchos de los ataques de malware sin archivos usan PowerShell, la cual es una interfaz de línea de comandos y una herramienta de scripts integrada en los sistemas operativos Microsoft Windows. Los crackers pueden ejecutar scripts de PowerShell para cambiar las configuraciones, robar contraseñas o causar otros daños a las organizaciones.

Las macros maliciosas son otro tipo de vector común para los ataques sin archivos. Las aplicaciones como Microsoft Word y Excel permiten a los usuarios definir macros, conjuntos de comandos para automatizar tareas simples como formatear texto o realizar diferentes cálculos. Los crackers pueden almacenar scripts maliciosos en estas macros; cuando un usuario abre el archivo, esos scripts se ejecutan automáticamente.

Otros tipos de malware. Los gusanos son programas maliciosos auto replicantes que pueden propagarse entre aplicaciones y/o dispositivos sin una interacción humana, a diferencia de un virus, que solo puede propagarse si un usuario ejecuta un programa comprometido. Si bien algunos gusanos no hacen más que propagarse, muchos tienen consecuencias más severas.

Los caballos de Troya se disfrazan de programas útiles y/o se ocultan dentro de software legítimo para engañar a los usuarios para que los instalen en las maquinas. Un troyano de acceso remoto o (RAT) crea una puerta trasera secreta en el dispositivo infectado. Otro tipo de troyano (dropper) instala malware adicional una vez que tiene un punto de apoyo. Ryuk es una de las cepas de ransomware recientes más devastadoras, utilizó el troyano Emotet para infectar los dispositivos.

Los rootkits son paquetes de malware que permiten a los crackers obtener acceso privilegiado, a nivel de usuario administrador en el sistema operativo de las computadoras o a otros activos de las organizaciones. Los crackers pueden utilizar estos permisos elevados para hacer prácticamente lo que se desee en las maquinas, como añadir y eliminar usuarios o reconfigurar aplicaciones. Los crackers suelen utilizar los rootkits para ocultar procesos maliciosos o desactivar el software de seguridad que podría detectarlos en las organizaciones.

El scareware asusta a los usuarios para que descarguen malware o pasen información confidencial a un estafador. El scareware suele aparecer como una ventana emergente repentina con un mensaje urgente, que suele advertir al usuario de que infringió alguna ley o de que su dispositivo tiene un algún virus. La ventana emergente indica al usuario que pague un tipo de multa o que descargue un software de seguridad falso que resulta ser malware real.

El spyware se oculta en los equipos infectados, recopilando información confidencial y transmitiéndola a un atacante. Un tipo común de spyware, llamado registrador de las claves es el que registra toda la información de un usuario y permite a los crackers obtener nombres de usuario,

contraseñas, cuentas bancarias y números de tarjetas de crédito para así poder ingresar posteriormente.

El adware envía spam al dispositivo con anuncios emergentes no deseados. El adware a menudo se incluye con el software gratis, sin que el usuario lo sepa. Cuando el usuario realiza la instalación el programa, también instala el adware sin querer. La mayoría de los programas publicitarios son más que una molestia. Sin embargo, algunos programas publicitarios recopilan datos personales, redireccionan los navegadores sitio web a sitios web maliciosos o incluso descargan más malware en el dispositivo del usuario si este hace clic en las ventanas emergentes que le van saliendo.

Ransomware. Este tipo de malware genera un bloqueo de los dispositivos o datos de las víctimas y se exige un pago de rescate, generalmente en forma de criptomoneda, para generar el desbloquearlos. Según el índice X-Force Threat Intelligence de IBM, el ransomware es el segundo tipo más común de ciberataque y representa el 17 % de los ataques en el mundo.

Los ataques de ransomware más básicos hacen que los activos sean inutilizables hasta que se pague el rescate, pero los ciberdelincuentes pueden utilizar tácticas adicionales para aumentar la presión sobre las víctimas.

En un doble ataque de extorsión, los ciberdelincuentes roban datos y amenazan con filtrarlos en la dark web si no se les paga. En un ataque de triple extorsión, los crackers cifran los datos de la víctima, los roban y amenazan con desconectar los sistemas de las organizaciones a través de un ataque de denegación distribuida del servicio (DDoS). Las demandas de rescate pueden oscilar entre decenas de miles y millones de dólares estadounidenses.

Malware de acceso remoto. Los crackers utilizan el malware de acceso remoto para obtener acceso a las computadoras, servidores u otros dispositivos de las organizaciones mediante

la creación o el aprovechamiento de puertas traseras. Según el índice X-Force Threat Intelligence, sembrar puertas traseras es el objetivo más común de los crackers lo cual representa el 21 % de los ataques en el mundo.

Las puertas traseras permiten a los ciberdelincuentes hacer mucho daño como robar datos o credenciales, tomar el control de los dispositivos o instalar malware aún más peligroso como ransomware. Algunos crackers utilizan malware de acceso remoto para crear puertas traseras que pueden vender a otros crackers, lo que puede darles varios miles de dólares estadounidenses cada uno de ellos.

Algunos malware de acceso remoto, como Back Orifice o CrossRAT, se crean intencionalmente con fines maliciosos. Los crackers también pueden modificar o dar un mal uso a software legítimo para acceder de forma remota a los dispositivos de las organizaciones. En particular, los ciberdelincuentes utilizan credenciales robadas para el protocolo de escritorio remoto (RDP) de Microsoft como puertas traseras.

Ofuscación de malware. Es una técnica que se utiliza por los ciberdelincuentes para ocultar la verdadera intención y funcionalidad del software malicioso donde implica la modificación del código y la estructura del malware para dificultar que las diferentes herramientas de seguridad y los analistas lo detecten y comprendan. La ofuscación de malware utiliza varias tácticas para impedir la detección y el análisis.

Modificación de Código. Los creadores del malware constantemente alteran el código del software malicioso para que sea ilegible para los métodos de detección tradicionales. Se utilizan técnicas como el polimorfismo, encriptación y metamorfismo. El malware polimórfico cambia constantemente su código para evadir la detección, lo que dificulta su identificación y seguimiento. Al encriptar el código, los creadores maliciosos dificultan que las herramientas de seguridad lean

las instrucciones maliciosas directamente. Por último, el malware metamórfico altera su código cada vez que infecta un nuevo sistema, para evitar así la detección basada en firmas conocidas.

Empaquetadores y Criptoanalizadores. Son herramientas que ofuscan aún más el malware comprimiendo o encriptando el archivo que se ejecuta. La compresión reduce el tamaño del malware, lo que hace más difícil su respectivo análisis, mientras que la encriptación asegura que este archivo permanezca encriptado hasta el momento de su ejecución. Cuando el malware se ejecuta, se desempaqueta y desencripta, lo que dificulta el análisis y su detección. Los empaquetadores y criptoanalizadores suelen incorporar técnicas de anti-análisis, lo que aumenta la complejidad de descifrar la intención de estos malware.

Técnicas de Anti-Análisis. Para dificultar el trabajo de los investigadores de seguridad, los diseñadores de malware emplean métodos de anti-análisis que como objetivo es dificultar la comprensión del comportamiento del malware generado. Dichas técnicas incluyen la inclusión de código basura, que son instrucciones irrelevantes o que no tienen ningún sentido si son agregadas al código malicioso. Adicionalmente, los comandos de suspensión introducen retrasos en la ejecución del malware, lo que logra confundir a las herramientas de análisis y retrasar así su detección. Por finalizar, los creadores del malware emplean técnicas de evasión de sandbox para lograr evitar el análisis en los entornos controlados donde los investigadores de seguridad pueden observar y analizar los diferentes comportamientos del malware de manera segura.

Carga y Ejecución Dinámicas. Algunos malware cargan y ejecutan su código de manera dinámica, lo cual dificulta identificar la funcionalidad completa de inmediato. Al cargar este código de manera dinámica durante la ejecución o con disparadores específicos, los creadores de malware pueden ofuscar la verdadera naturaleza y capacidades del dicho malware. Esta técnica

permite que el malware evolucione con el pasar del tiempo, agregando nuevas funcionalidades sin necesidad de recompilación o redespliegue.

Comunicación Ofuscada. El malware puede ofuscar su comunicación con el servidor de comando y control "C2", haciéndolo más difícil de detectar y rastrear hasta el origen. A menudo se logra mediante el uso de las técnicas de encriptación o codificación para lograr ocultar los datos que se transmiten. Al disfrazar esta comunicación, los creadores de malware buscan evitar la monitorización de redes y así prevenir la detección de sus actividades maliciosas.

Marco teórico

El uso de técnicas de ML y DL se ha consolidado como un pilar en la defensa cibernética moderna. Estas técnicas permiten que sistemas automatizados aprendan a distinguir entre comportamiento normal y malicioso mediante el análisis de grandes volúmenes de datos, superando en muchos casos las limitaciones de los métodos tradicionales basados en firmas estáticas. En el ámbito de la ciberseguridad, ML/DL se emplea para reforzar la detección de intrusiones, la detección de anomalías, el análisis de malware y la inteligencia de amenazas, entre otras tareas críticas. La compilación reciente de (Elhanashi A.; Dini P., 2024) enfatiza que las técnicas avanzadas de DL, procesamiento de lenguaje natural e IA explicable están revolucionando la detección de intrusos, detección de anomalías e inteligencia de amenazas, abordando aplicaciones como el análisis de malware, la seguridad en IoT y nube, la protección de blockchain y la defensa contra ataques adversarios. En este sentido, la confidencialidad, integridad y disponibilidad de los sistemas conectados a redes es protegida mediante sistemas que aprenden de los datos para adaptarse a ataques nuevos y evolucionados.

Fundamentos del Marchine Learning y Deep Learning

En su forma más básica, el ML implica algoritmos que extraen patrones de datos para tomar decisiones o hacer predicciones, minimizando la intervención humana. El DL, por su parte, es un subcampo de ML basado en redes neuronales artificiales de múltiples capas (por ejemplo, redes neuronales convolucionales y recurrentes) que pueden aprender representaciones jerárquicas complejas. Como señalan (Halbouni et al., 2022), términos como IA, ML y DL a menudo se usan indistintamente, aunque conceptualmente el aprendizaje profundo está contenido dentro del aprendizaje automático, y éste dentro de la IA. En ciberseguridad, la motivación principal para usar ML/DL radica en la capacidad de procesar grandes volúmenes de datos heterogéneos (registros de red, binarios de malware, tráfico IoT, etc.) y detectar patrones que indiquen ataques, frente a los métodos tradicionales que dependen de firmas predefinidas.

El ML puede dividirse en enfoques supervisados y no supervisados. En los métodos supervisados, el modelo se entrena con datos etiquetados (por ejemplo, tráfico “maligno” vs “benigno”) y aprende a clasificar nuevos datos según esas etiquetas. En contraposición, los métodos no supervisados buscan estructurar los datos sin etiquetas, siendo útiles para detección de anomalías (identificando datos que se desvían del patrón normal) o clustering. La literatura reciente revisada por (Dixit & Silakari, 2020) indica que en ciberseguridad se emplean ambos enfoques: se han aplicado tanto redes convolucionales y recurrentes supervisadas como autoencoders o entrenados sin etiquetas para identificar comportamientos inusuales. Los enfoques supervisados suelen usarse en tareas como clasificación de muestras de malware o ataques conocidos, mientras que los no supervisados son populares en detección de anomalías, donde los patrones de ataque pueden no estar etiquetados previamente. Además, el DL por refuerzo ha

surgido como herramienta para aprender políticas de defensa de forma autónoma, aunque su uso en ciberseguridad aún es incipiente comparado con las CNN y RNN tradicionales.

En el detalle arquitectónico, el DL se realiza mediante redes neuronales artificiales. Entre las más usadas en seguridad informática destacan las CNN y las RNN. Las CNN son eficaces para procesar datos con estructura espacial como imágenes, y se han aplicado innovadoramente al problema de detección de malware al convertir binarios en imágenes de píxeles, de modo que las CNN detectan patrones de bytes maliciosos. Las RNN y sus variantes como LSTM o GRU son útiles para datos secuenciales, por ejemplo, para analizar secuencias de llamadas al sistema o tráfico de red a lo largo del tiempo. Estudios comparativos han encontrado que combinaciones híbridas CNN-LSTM pueden superar a modelos puros: por ejemplo, un IDS basado en CNN 1D combinado con LSTM alcanzó una precisión del 91,8%. En general, las redes profundas aprenden automáticamente representaciones "features" relevantes a partir de los datos sin necesidad de ingeniería manual intensiva, lo que mejora la precisión y generalización respecto a métodos tradicionales. De hecho, (Dixit & Silakari, 2020) concluyen que el uso de modelos de DL aumenta la precisión, escalabilidad y confiabilidad de las soluciones de seguridad con respecto a técnicas previas.

Algoritmos Supervisados vs No Supervisados y Híbridos

El aprendizaje supervisado clásico abarca algoritmos como SVM, árboles de decisión, bosques aleatorios, etc., que han sido empleados con éxito en ciertas áreas de seguridad, por ejemplo, filtrado de spam o detección de intrusiones de red. Por otro lado, los métodos no supervisados incluyen clustering K-means, DBSCAN, etc. y técnicas de reducción de dimensionalidad para identificar desviaciones o grupos en los datos. Adicionalmente, se desarrollan enfoques híbridos que combinan estática y dinámicamente los datos o combinan

múltiples clasificadores. Así como lo menciona (Gibert et al., 2020) clasifican los métodos de ML para detección de malware en tres categorías: basados en análisis estático, dinámico y híbrido. En paralelo, en sistemas de detección de intrusiones en redes (IDS) es común distinguir entre detección de uso (basada en firmas o reglas) y detección de anomalías (usualmente basada en ML). (Halbouni et al., 2022) explica que la detección basada en firmas tiene una baja tasa de falsas alarmas, pero sólo detecta ataques conocidos, mientras que la detección de anomalías basada en ML incurre en más falsas alarmas, pero puede descubrir ataques nuevos e inesperados. Así, en la práctica muchas soluciones combinan ambos enfoques: el ML supervisado detecta patrones maliciosos previos, y algoritmos no supervisados monitorizan comportamientos en tiempo real para señalizar anomalías sin sesgos de firma.

En el marco del DL, las arquitecturas se usan tanto en modo supervisado CNN entrenada con ejemplos de malware vs benignos, como no supervisado autoencoders o GAN para modelar tráfico normal y detectar outliers. Por ejemplo, un autoencoder (AE) entrenado con tráfico de red normal puede identificar intrusiones cuando se observa un alto error de reconstrucción. Dixit y (Dixit & Silakari, 2020) ilustran que las diferentes clases de redes (CNN, AE, DBN, RNN, GAN, aprendizaje por refuerzo, etc.) cubren tanto modelos supervisados como no supervisados. En suma, la elección del algoritmo supervisado/no supervisado depende de la disponibilidad de datos etiquetados y del tipo de problema (clasificación vs detección de anomalías), mientras que hoy en día los avances en DL permiten incluso una mezcla híbrida que extrae múltiples tipos de información de los datos.

Redes Neuronales Profunda

Las redes neuronales profundas han permitido saltos cualitativos en la detección de amenazas. En particular, las CNN han sido muy exitosas en clasificación de malware cuando el

código malicioso se representa como imagen, pues pueden aprender características espaciales invariantes a transformaciones de bytes. Por ejemplo, transformar un ejecutable malicioso en una imagen permite que la CNN identifique patrones intrínsecos incluso si el código cambia superficialmente, aprovechando que las estructuras globales de la imagen permanecen estables. Con esta técnica, (Zhu et al., 2023) reportan una precisión del 98,7% en conjuntos de datos populares de malware *(Malvis, Malimg, MS Big2015*) usando un modelo CNN ajustado. Esto ilustra la potencia de las CNN para capturar características complejas de la señal maliciosa de forma automática.

Por su parte, las RNN (y variantes LSTM/GRU) se aplican a flujos de datos secuenciales: secuencias de llamadas al sistema en un malware, registros de red de sesiones, o incluso comportamiento temporal de dispositivos IoT. Estas redes recuerdan información previa en la secuencia y la utilizan para clasificar el estado actual. En la práctica, combinaciones como CNN+LSTM han mostrado desempeño superior en escenarios de IDS: un estudio revisado por (Halbouni et al., 2022) informa que un modelo CNN-LSTM alcanzó 91,8% de exactitud en detección de intrusos sobre tráfico de red, superando a modelos CNN o LSTM puros. Otras arquitecturas frecuentes en ciberseguridad son las autoencoders para compresión o detección de anomalías y las Generative Adversarial Networks (GAN), especialmente en generación de ejemplos adversarios o refuerzo del modelo. En general, el DL permite extraer automáticamente representaciones jerárquicas ricas a partir de los datos de entrada, aspecto fundamental para lidiar con la variedad y complejidad de las amenazas actuales.

Análisis Estático y Dinámico de Malware

El análisis de malware puede realizarse de dos formas complementarias: estática y dinámica. En el análisis estático se examina el ejecutable (binario) sin ejecutarlo, extrayendo

características como cadenas de texto, cabeceras de archivos, instrucciones de ensamblador, hash de secciones, o representaciones de imágenes de bytes. Esto suele ser rápido y seguro, pero los ataques pueden ofuscar el código (por ejemplo, empaquetadores, cifrado, polimorfismo) para evadirlo. (Gibert et al., 2020) señalan que, aunque el análisis estático es eficiente, los malwares ofuscados pueden ocultar sus firmas, dificultando su detección en esta fase. Por otro lado, el análisis dinámico ejecuta el malware en un entorno controlado (sandboxes, máquinas virtuales) observando su comportamiento: monitoriza modificaciones en archivos, conexiones de red, cambios en el sistema, etc. (Gibert et al., 2020) resaltan que este enfoque captura el comportamiento real y tiende a detectar variantes ofuscadas, pues el monitoreo en tiempo de ejecución revela las acciones verdaderas del código malicioso. Sin embargo, el análisis dinámico es más lento y requiere más recursos, y algunos malwares pueden detectar la sandbox y permanecer inactivos o comportarse de forma diferente para evadirla.

De hecho, una práctica avanzada es combinar ambos enfoques en un sistema híbrido: por ejemplo, se puede realizar un filtrado inicial estático rápido para descartar archivos claramente benignos, y pasar al análisis dinámico aquellos sospechosos. El marco teórico apunta a que los modelos de ML/DL aprovechan características de ambos tipos de análisis. Así, por ejemplo, una CNN puede procesar "imágenes" binarias extraídas estáticamente, mientras que una RNN supervisa secuencias de eventos dinámicos. (Gibert et al., 2020) sintetizan que tradicionalmente los antivirus basados en firmas y heurísticas eran incapaces de seguir el ritmo de nuevas variantes, lo cual abrió la oportunidad para ML, capaz de aprender patrones complejos de análisis estático y dinámico de malware. En resumen, el análisis de malware en la literatura actual reconoce la complementariedad entre estática (rápida pero vulnerable a ofuscación) y dinámica (robusta ante

obfuscación pero costosa), y explora arquitecturas híbridas y multicapa que integran resultados de ambos para mejorar la detección global.

Ofuscación y Ransomware

Entre las amenazas avanzadas, la ofuscación de código es una técnica común usada por los atacantes: por ejemplo, empaquetadores ejecutables, cifrado de secciones de código, mutaciones polimórficas o metamórficas. Estas técnicas dificultan la tarea estática y exigen métodos robustos. Como advierten los estudios, el ransomware se ha destacado últimamente como un caso particularmente dañino de malware ofuscado: es un tipo de malware que explota mecanismos criptográficos para secuestrar los archivos del usuario y exigir rescate monetario. (Al-rimy et al., 2018) describen el ransomware como malware que secuestra recursos del usuario mediante criptografía, demandando dinero a cambio de la llave de descifrado. Resaltan que, a diferencia de virus tradicionales, el ransomware deja efectos irreversibles incluso tras la eliminación del código malicioso, causando enormes pérdidas por inactividad y daños colaterales. De hecho, casos como WannaCry, Locky o Cerber son ejemplos de ransomware que han paralizado sistemas enteros a gran escala.

Otros malwares avanzados incluyen rootkits altamente encubiertos, criptominadores ocultos o malware de día cero (que explota vulnerabilidades desconocidas). Estos plantean retos especiales: requieren que los modelos de ML/DL aprendan características muy sutiles. Por ello, en el marco actual se estudian técnicas de autoaprendizaje y transferencia para captar patrones emergentes. Además, la ofuscación puede inferir a los investigadores que se use preprocesamiento robusto: extracción de features resistentes a transformaciones, por ejemplo, análisis de entropía o Graph Neural Networks que capturan la estructura de llamadas internas. (Gibert et al., 2020) mencionan que el aprendizaje de características profundas por redes neuronales permite extraer

automáticamente atributos discriminantes incluso ante ofuscación o cloaking, al contrario de los detectores estáticos convencionales. En síntesis, las amenazas avanzadas como el ransomware obligan a sistemas basados en ML/DL a incorporar análisis estáticos extensivos, técnicas de DL multi-modal (combinando código y comportamiento) y capacidades para adaptarse a malware metamórfico.

Detección de Anomalías

La detección de anomalías es un componente crítico en ciberseguridad para descubrir ataques desconocidos o desviaciones en el comportamiento de sistemas. En ML, la detección de anomalías se aborda típicamente con modelos no supervisados o con aprendizaje supervisado entrenado con datos de tráfico normal. Como se ha mencionado, esta modalidad suele incurrir en un mayor porcentaje de falsas alarmas, pero tiene la ventaja de identificar intrusiones nuevas. En redes, un IDS basado en ML intentará modelar el "tráfico normal" y señalará como anomalías patrones estadísticamente raros. (Halbouni et al., 2022) ilustran que los sistemas basados en firmas (misuse) tienen bajas falsas alarmas, pero detectan sólo ataques conocidos, mientras que la detección de anomalías (ML) baja los falsos negativos al poder señalar comportamientos atípicos, aunque se arriesga a más falsos positivos. Este enfoque es esencial en entornos dinámicos: por ejemplo, en IoT o redes empresariales distribuidas, es imposible enumerar todas las firmas, de modo que la detección de anomalías permite descubrir actividades maliciosas emergentes.

En la práctica, se emplean redes profundas (autoencoders, redes neuronales recurrentes, etc.) para esta tarea. Un autoencoder entrenado con tráfico de red normal intentará reconstruir entrada vs salida; si un malware es inyectado, la reconstrucción falla significativamente señalando anomalía. Además, algunas soluciones híbridas combinan detección de uso (firmas derivadas del aprendizaje supervisado) con detección de anomalías, implementando múltiples capas de defensa.

Con la explosión de datos como es el big data en ciberseguridad, la capacidad de procesar y analizar volúmenes masivos es clave: se recurre entonces a enfoques de ML basado en streaming y aprendizaje incremental para detectar anomalías en tiempo real. (Qureshi et al., 2024) apuntan la necesidad de soluciones "escalables en tiempo real" para entornos IoT debido a la enorme cantidad de datos generados. En resumen, la detección de anomalías utiliza ML para complementar la defensa tradicional e identificar patrones de ataque aún no codificados, aprovechando modelos no supervisados o semisupervisados entrenados en comportamiento benigno para revelar lo inusual.

Vulnerabilidades y Seguridad en IoT, Nube y Sistemas Distribuidos

La interconexión masiva de dispositivos IoT y la creciente adopción de servicios en la nube han ampliado la superficie de ataque en ciberseguridad. Los dispositivos IoT suelen ser heterogéneos, con recursos limitados y protocolos variados, lo que plantea desafíos únicos de seguridad. Además, la integración con la nube introduce riesgos adicionales: por ejemplo, servicios IoT basados en PaaS/IaaS pueden manejar datos sensibles sin regulación uniforme, exponiendo interfaces vulnerables. (Singh et al., 2024) destacan que la convergencia de sistemas tradicionales y basados en la nube crea amenazas nuevas y singulares en la ausencia de estándares globales de seguridad. Por su parte, (Qureshi et al., 2024) señalan que a medida que los entornos IoT se expanden, la complejidad del malware específico de IoT aumenta, y faltan conjuntos de datos adecuados para entrenar modelos en estos entornos.

Frente a esto, el ML/DL ofrece soluciones prometedoras en IoT y nube. (El-Sofany et al., 2024) desarrollaron un sistema de detección de ataques para IoT que usa múltiples algoritmos ML, alcanzando una precisión del 99,9% en la identificación de actividades maliciosas y superando modelos previos. Esta implementación empleó siete clasificadores diferentes y configuró un

agente inteligente para identificar patrones de ataque en tráfico IoT en tiempo real, logrando casi un 100% de detección y un AUC perfecto. Casos de uso documentados incluyen la monitorización de sensores en redes industriales, detección de botnets en redes celulares IoT, análisis de tráfico de dispositivos médicos conectados, etc.

En entornos distribuidos como sistemas en la nube o redes globales, los desafíos incluyen gestionar datos provenientes de múltiples fuentes y asegurar comunicaciones seguras. El aprendizaje federado y la criptografía homomórfica se exploran como enfoques para aplicar ML sin exponer datos sensibles. Adicionalmente, se utilizan técnicas de ML para fortalecer la autenticación, por ejemplo, sistemas biometricos o basados en comportamiento del usuario, y para el threat intelligence el análisis de grandes volúmenes de registros en la nube buscando indicadores de compromiso. (Singh et al., 2024) también remarcan que tecnologías emergentes como fog computing y blockchain se combinan con ML/AI para mejorar la seguridad de IoT distribuido. Mohanta et al. (citado en (Singh et al., 2024)) proponen usar ML/AI junto con blockchain para autenticar dispositivos IoT y analizar tráfico en busca de anomalías, todo en un ecosistema IoT-cloud, lo cual refuerza la defensa contra accesos no autorizados.

En síntesis, los sistemas distribuidos e IoT se benefician de modelos ML capaces de procesar datos distribuidos y detectar patrones de ataque en diversos escenarios. Queda, sin embargo, el reto de garantizar que estos modelos funcionen eficientemente en dispositivos con recursos limitados y bajo conectividad intermitente. Aquí cobra relevancia el preprocesamiento de datos, la reducción de dimensionalidad y la optimización de modelos, por ejemplo, a través de redes neuronales ligeras o poda de modelos, para implementaciones en el borde (edge computing).

Preprocesamiento de Datos y Calidad de Conjuntos de Datos

Un paso crítico para cualquier solución basada en ML es la preparación y calidad de los datos. En ciberseguridad esto implica limpiar registros (quitar campos corruptos, anonimizar información sensible), normalizar escalas, por ejemplo, estandarizar tiempos entre eventos, etiquetar correctamente muestras de ataques frente a benignas. (Vourganas & Michala, 2024) enfatizan que los avances en ML para ciberseguridad están fuertemente limitados por la heterogeneidad y calidad de los datasets disponibles. Notan que existe una falta de uniformidad en las características incluidas en los datasets, por ejemplo, formatos de red distintos, en las metodologías de recolección (datos sintéticos vs reales, entornos simulados vs reales) y en los requisitos de preprocesamiento. Esta disparidad dificulta la comparación entre estudios y puede introducir sesgos: un modelo entrenado con datos desbalanceados o mal etiquetados puede presentar falsas alarmas o ignorar tipos de ataque críticos.

Por ello, se debe dedicar atención al preprocesamiento: técnicas comunes incluyen normalización de atributos numéricos, codificación de categorías, por ejemplo, transformar direcciones IP a vectores, manejo de valores faltantes (imputación o filtrado), y extracción de features relevantes (cálculo de entropía de secciones de un ejecutable, ingeniería de características específicas de red como tasa de paquetes, etc.). En escenarios de DL, a menudo se emplea representación bruta (como imágenes de bytes), lo cual reduce la ingeniería manual, pero aun así puede requerir filtrado previo de ruido. Además, se presta especial atención a la calidad de las etiquetas en los datos supervisados, evaluando que no existan ejemplos etiquetados incorrectamente que confundan al modelo.

Otro aspecto crucial es la escala de datos y la actualización continua. En ciberseguridad, los datos crecen constantemente (millones de registros diarios en redes corporativas). Los modelos

deben reentrenarse periódicamente con nuevos datos para adaptarse a cambios. Técnicas como el aprendizaje incremental permiten actualizar los modelos sin reentrenar desde cero. Finalmente, la disponibilidad de datasets de código abierto es limitada para algunos ámbitos especializados, por ejemplo, malware IoT o tráfico cifrado moderno, lo que refuerza la recomendación de (Vourganas & Michala, 2024) de crear repositorios más completos y consistentes.

Retos, Ataques Adversarios, Deriva de Concepto e Interpretabilidad

A pesar de sus éxitos, los sistemas basados en ML/DL enfrentan varios retos en ciberseguridad. Uno de ellos son los ataques adversarios: pequeños cambios maliciosos en la entrada pueden engañar al modelo para que lo clasifique como benigno, sin alterar su funcionalidad nociva. (Elhanashi A.; Dini P., 2024) incluye los ataques adversarios como un tópico central en seguridad de IA. Los atacantes pueden generar ejemplos adversarios específicamente para redes neuronales, explotando la sensibilidad de los clasificadores profundos. Por ello, se investiga el entrenamiento adversario (inyectar ejemplos adversarios en el entrenamiento) y el desarrollo de modelos robustos.

Otro reto es la deriva de concepto (concept drift). Esto ocurre cuando las características del tráfico o del malware cambian con el tiempo (por ejemplo, nuevos protocolos de comunicación o nuevas técnicas de ofuscación), de modo que el modelo entrenado previamente se vuelve obsoleto. (Gibert et al., 2020) identifican explícitamente el drift como un problema crítico en el dominio del malware. Por ejemplo, un detector entrenado con muestras de malware de 2018 puede fallar al detectar variantes de 2023 cuyas propiedades han evolucionado. Para contrarrestar esto, se implementan técnicas de aprendizaje continuo o adaptativo, donde el modelo se actualiza periódicamente con datos nuevos, o se utilizan detectores de cambio para desencadenar reentrenamientos.

La interpretabilidad es otro desafío importante. Los modelos de DL son a menudo cajas negras difíciles de explicar: no está claro por qué clasifican una muestra como maliciosa. Esto es problemático en ciberseguridad, donde los analistas requieren justificar alertas y entender el comportamiento del sistema. Un estudio de (Bensaoud et al., 2024) subraya que los clasificadores basados en DL "no pueden explicar sus decisiones", recomendando el uso de técnicas de XAI o ML interpretable. Mediante métodos como la visualización de activaciones, análisis de sensibilidad o logging detallado, se busca entender qué características están influyendo en la detección. (Zhu et al., 2023) mencionan concretamente que es importante emplear registros (logging), perfilado, pruebas unitarias y herramientas de visualización para diagnosticar y asegurar la estabilidad y equidad del modelo. Por ejemplo, se pueden generar mapas de calor sobre una imagen de malware para ver qué regiones byte son más determinantes en la predicción. Estas técnicas ayudan a identificar sesgos (por ejemplo, si el modelo aprende un atajo espurio) y a corregir problemas éticos como false positives en usuarios inocentes.

En conjunto, estos retos motivan enfoques de investigación complementarios: robustecer modelos contra ejemplos adversos, incorporar detección de deriva (adaptación del modelo cuando cambian los datos), e integrar la interpretabilidad desde el diseño. Aunque aún no existe una solución única, se está trabajando en marcos de ML "a prueba de adversarios" y en arquitecturas multiagente que permitan explicaciones por partes del modelo. La actualización dinámica de modelos con nuevos datos streaming learning y la participación de expertos humanos en el ciclo de entrenamiento son tendencias recientes en este sentido.

Casos de Uso y Resultados Comparativos

El uso de ML/DL en ciberseguridad se ha probado en numerosos estudios y escenarios reales, con resultados frecuentemente superiores a métodos clásicos. A continuación, se mencionan algunos casos ilustrativos reportados:

- Detección de intrusiones en redes (IDS). Varios trabajos han aplicado CNN, RNN o modelos híbridos a conjuntos de datos públicos (por ejemplo, NSL-KDD, CICIDS). (Halbouni et al., 2022) reportan que un modelo CNN-LSTM alcanzó 91,8% de precisión en clasificación de tráfico de red por ataque vs normal. Otros estudios muestran que las RNN puras suelen preformar ligeramente peor que las CNN-LSTM en este tipo de datos tabulares de red. En general, los modelos DL tienden a mejorar las métricas de detección al aprender mejor las características de los ataques de red (por ejemplo, patrones temporales en los flujos de paquetes).
- Detección de malware (Windows/Linux). Los modelos basados en redes profundas entrenados con múltiples características (cabeceras PE, llamadas a sistema, etc.) han logrado altos niveles de clasificación. Por ejemplo, usando técnicas de visión por computador (conversión de código a imagen) se han obtenido precisiones cercanas al 99%. En general, los enfoques de DL reportan precisión y recall superiores en detección de malware frente a SVM o árboles de decisión tradicionales, gracias a su capacidad de modelar relaciones no lineales complejas.
- Seguridad en IoT. El trabajo de (El-Sofany et al., 2024) mostró resultados muy prometedores en dispositivos IoT: su sistema de ML logró 99,9% de precisión y 99,9 de F1 detectando ciberataques en redes IoT, superando claramente los métodos previos. Otro ejemplo es la detección de botnets IoT con redes neuronales anchas y

aislamiento de anomalías, donde se han reportado tasas de detección superiores al 95%. Además, se han aplicado autoencoders en tráfico de sensores para alertar de interferencias anómalas con buenos resultados de sensibilidad.

- Ransomware y malware móvil. En Android, por ejemplo, la combinación de análisis estático (permisos, llamadas al sistema) y dinámico (monitorización de comportamiento) con ML ha permitido construir apps de detección de ransomware móvil con más del 90% de acierto. (Guerra-Manzanares, 2024) examinan estos métodos y señalan que, pese a la variedad de malware Android, los sistemas ML actualizados logran identificar familias nuevas con buen rendimiento.
- Prevención de ataques adversarios. Algunos casos de uso documentados incluyen entrenamiento adversario: entrenar el modelo con muestras ligeramente perturbadas para aumentar su robustez. Un método con red neuronal ligera mejorada para detectar malware incluso frente a muestras adversarias, demostrando mejores resultados que 19 otros modelos DL estándar. Esto indica que incorporar adversarios en el entrenamiento puede mejorar la detección real.

En síntesis, los estudios comparativos muestran consistentemente que las técnicas de ML/DL aportan mejoras en métricas clave (precisión, recall, F1) respecto a enfoques antiguos. (Dixit & Silakari, 2020) concluyen que, en la mayoría de los casos, los modelos profundos superan a los algoritmos convencionales en seguridad, al margen de que requieran más potencia de cálculo. No obstante, destacan que las ganancias dependen de un buen ajuste de hiperparámetros, la calidad del dataset y la arquitectura utilizada. También se observa que las implementaciones multinivel (por ejemplo, CNN seguido de RNN) suelen entregar resultados óptimos en clasificación de malware complejo o de intrusiones avanzadas.

De cara al futuro, la evaluación de estas técnicas en entornos reales (más allá de datasets académicos) es un punto importante. La producción de comparativas entre trabajos (en conjunto o de manera meta-analítica) es todavía incipiente, pero crece el número de estudios documentando ataques recientes (LockBit, REvil, cryptojacking, exploits de día cero) y mostrando cómo los modelos actuales los detectan. En este contexto, el marco teórico revisado indica que el ML y el DL son hoy herramientas esenciales en la detección, prevención y mitigación de malware, siempre que se combinen con un riguroso preprocesamiento de datos y una continua adaptación a nuevas amenazas.

Diseño metodológico

Para el desarrollo del presente trabajo, se llevó a cabo una búsqueda exhaustiva en diversas bases de datos académicas y repositorios especializados en tecnología y ciberseguridad. El propósito fue recopilar información científica actualizada y relevante sobre la aplicación del ML y el DL en la detección, prevención y mitigación de malware en entornos organizacionales.

El proceso metodológico comprendió varias etapas claramente definidas. En primer lugar, se formularon las preguntas de investigación, orientadas a identificar las técnicas más efectivas de aprendizaje automático utilizadas en el análisis estático y dinámico de malware, así como los desafíos asociados a su implementación práctica.

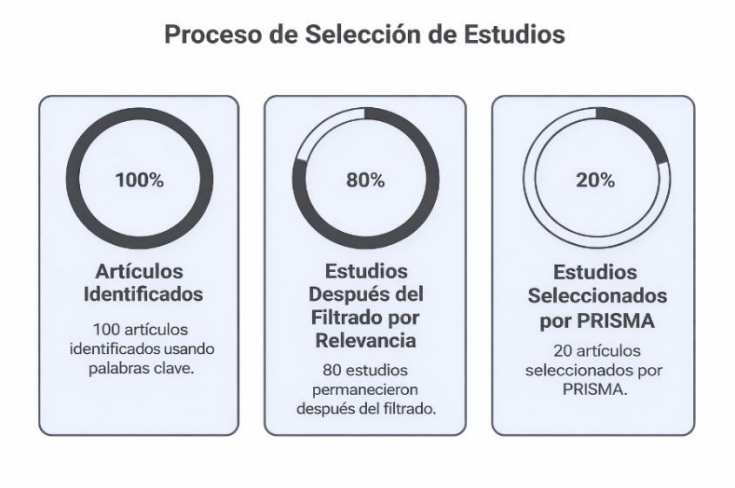
Posteriormente, se diseñó la revisión sistemática considerando criterios de inclusión y exclusión rigurosos. Las bases de datos consultadas fueron ACM Digital Library, Google Scholar, IEEE Digital Library, Nature, Research Gate, Science@Direct, Scopus, Springer Link, Springer Open, por su relevancia en publicaciones científicas del ámbito tecnológico. Los criterios de búsqueda incluyeron:

- Estudios publicados entre 2018 y 2025.
- Trabajos que abordaran la aplicación de ML o DL en análisis estático (como la extracción de opcodes) o dinámico (como el monitoreo de llamadas a APIs).
- Publicaciones que presentaran métricas cuantitativas de desempeño, tales como precisión (accuracy) o F1-score.

El proceso de selección consistió en la lectura crítica de títulos, resúmenes y contenidos completos, eliminando duplicados y filtrando los estudios que cumplían con los criterios establecidos, así como se visualiza en la figura 2.

Figura 2

Selección de artículos.



Nota: El proceso de selección de estudios reduce significativamente el grupo inicial de artículos a un conjunto enfocado de estudios y reportes relevantes. Fuente: elaboración propia.

Finalmente, los artículos seleccionados se clasificaron temáticamente en cuatro categorías principales:

- Análisis estático: investigaciones centradas en la detección de malware mediante la extracción y comparación de características binarias.
- Análisis dinámico: estudios basados en la observación del comportamiento del software en entornos controlados o sandboxes.
- Modelos predictivos: trabajos enfocados en el uso de algoritmos supervisados para la identificación automatizada de amenazas.
- Interpretabilidad: publicaciones que integran técnicas de explicabilidad, como los valores SHAP, para comprender las decisiones de los modelos.

Esta metodología permitió obtener una visión integral y actualizada del estado del arte, sirviendo como base para el diseño de modelos predictivos y la identificación de estrategias innovadoras en materia de ciberseguridad.

Revisión sistemática sobre el uso del Machine Learning en el análisis de malware

En la presente revisión sistemática se realiza un análisis profundo sobre el papel del ML y el DL dentro del panorama actual de la ciberseguridad. La metodología empleada se fundamenta en la recopilación y síntesis de investigaciones recientes, estudios teóricos y aplicaciones prácticas, con el propósito de ofrecer una visión integral del avance tecnológico en este campo.

El enfoque principal de la revisión está orientado a la identificación de las amenazas cibernéticas emergentes y al análisis de cómo las técnicas basadas en ML y DL contribuyen a su detección, prevención y mitigación. Asimismo, se examinan los principios fundamentales de estas tecnologías y su aplicación en diferentes áreas de la ciberseguridad, como la detección de intrusiones, la autenticación de usuarios, el análisis de comportamiento y la respuesta ante incidentes.

Tras una Revisión Sistemática de Literatura (RSL), se puede indicar que el uso del ML y DL se ha consolidado como una estrategia altamente eficaz para la detección, análisis, mitigación y prevención de malware y otras amenazas cibernéticas. La revisión abarcó más de 100 fuentes académicas, artículos científicos, informes técnicos y documentos institucionales recientes que han documentado avances significativos en el campo de la ciberseguridad basada en IA, en la figura 3 se puede observar la metodología RSL.

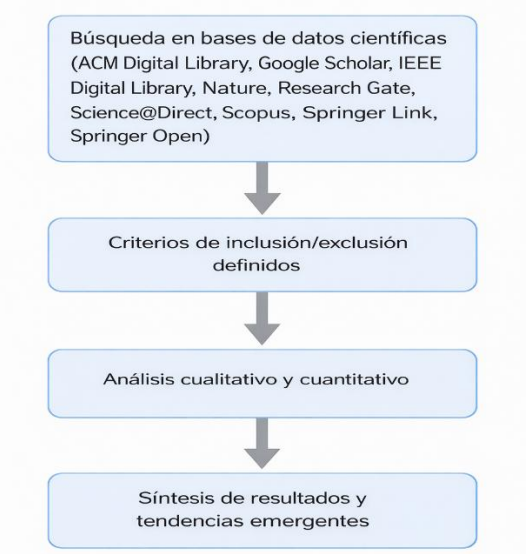
El primer paso de este RSL es definir la pregunta "*¿Cómo puede analizarse la aplicabilidad de las técnicas de Machine Learning para la detección, prevención y mitigación del malware en entornos organizacionales?"* en la cual se pretende proporcionar respuesta sobre la

detección, análisis, mitigación y prevención de malware y otras amenazas cibernéticas para mejorar la seguridad en las organizaciones.

El segundo paso consiste en recopilar todos los trabajos de investigación relacionados según los términos de búsqueda específicos. Las palabras clave identificadas se pueden consultar en la figura 4.

Figura 3

Resumen de la metodología RSL



Nota: Explicación del Metodología RSL. Fuente: elaboración propia.

Para la selección de las bases de datos bibliográficas, que proporciona en el campo de la Informática y la Ingeniería: ACM Digital Library, Google Scholar, IEEE Digital Library, Nature, Research Gate, Science@Direct, Scopus, Springer Link, Springer Open. La Figura 5 muestra el número de trabajos obtenidos y categorizados por cada base de datos.

Figura 4

Palabras clave identificadas.

Nota: Popularidad de las palabras en el campo de la tecnología. Fuente: elaboración propia

A continuación, se definen los criterios de inclusión (I) y exclusión (E) para añadir/eliminar los trabajos que sean relevantes/irrelevantes para este estudio:

Criterios de inclusión

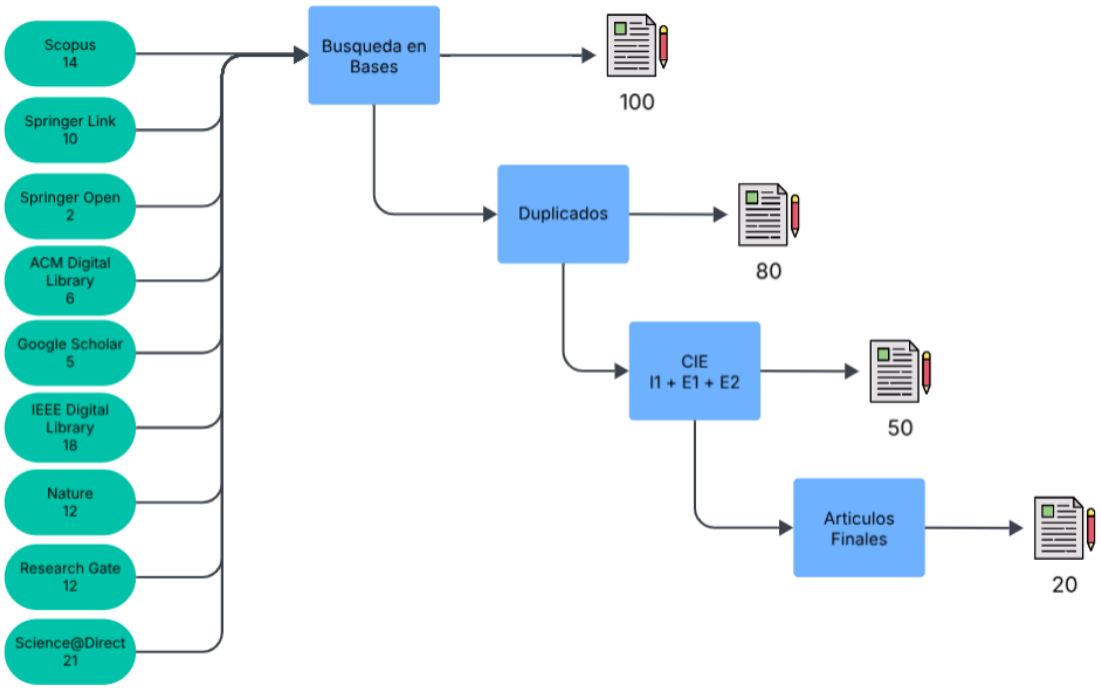
- I1: Artículos publicados en los últimos cinco años.

Criterios de exclusión

- E1: Informes técnicos, resúmenes, encuestas (literatura gris) y estudios secundarios (SMS).
- E2: Trabajos escritos en idiomas diferentes a inglés y español.

Figura 5

Descripción general



Nota: Descripción general del proceso de filtrado de artículos. Fuente: elaboración propia.

Es así, que la revisión del título, el resumen y las palabras clave para identificar aquellos artículos que puedan ser de interés y descartar aquellos que no cumplan con los criterios de inclusión y exclusión que fueron definidos. Una vez completado este paso, se analizaron y clasificaron todos los artículos seleccionados. El conjunto final de trabajos relevantes para este estudio y su análisis se presentan en la tabla 1 en la cual encontramos un extracto del archivo total el cual se encuentra en el anexo A - Artículos seleccionados.

Tabla 1 *Resumen artículos seleccionados*

Título	Año	Autores
Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review	2024	Shanjita Akter Prome and Neethiahnanthan Ari Ragavan and Md Rafiqul Islam and David Asirvatham and Anasuya Jegathevi Jegathesan
Evolving techniques in cyber threat hunting: A systematic review	2024	Arash Mahboubi and Khanh Luong and Hamed Aboutorab and Hang Thanh Bui and Geoff Jarrad and Mohammed Bahutair and Seyit Camtepe and Ganna Pogrebna and Ejaz Ahmed and Bazara Barry and Hannah Gately
Machine learning and deep learning for user authentication and authorization in cybersecurity: A state-of-the-art review	2024	Zinniya Taffannum Pritee and Mehedi Hasan Anik and Saida Binta Alam and Jamin Rahman Jim and Md Mohsin Kabir and M.F. Mridha
A comprehensive literature review on ransomware detection using deep learning	2025	Er. Kritika
Systematic review of deep learning solutions for malware detection and forensic analysis in IoT	2024	Siraj Uddin Qureshi and Jingsha He and Saima Tunio and Nafei Zhu and Ahsan Nazir and Ahsan Wajahat and Faheem Ullah and Abdul Wadud
Distributed denial of service attacks in cloud: State-of-the-art of scientific and commercial solutions	2021	Aanshi Bhardwaj and Veenu Mangat and Renu Vig and Subir Halder and Mauro Conti
A comprehensive plane-wise review of DDoS attacks in SDN: Leveraging detection and mitigation through machine learning and deep learning	2025	Dhruv Kalambe and Divyansh Sharma and Pushkar Kadam and Shivangi Surati

Nota: Resumen de los artículos seleccionados. Fuente: elaboración propia

El desarrollo de esta revisión sistemática permitió realizar un análisis profundo y actualizado sobre el papel que desempeñan el ML y el DL en el panorama contemporáneo de la ciberseguridad. A través de la recopilación y comparación de estudios recientes, tanto teóricos como aplicados, fue posible comprender cómo estas tecnologías están transformando los mecanismos de detección, prevención y respuesta frente a amenazas cada vez más sofisticadas. Con lo cual se puede verificar el análisis en el anexo B - Matriz Revisión Bibliográfica.

Los resultados obtenidos muestran que el uso de modelos inteligentes no solo fortalece las defensas digitales, sino que también promueve una gestión más proactiva y adaptativa del riesgo. Las investigaciones revisadas demuestran que técnicas como las redes neuronales profundas, los algoritmos de clasificación supervisada y los modelos híbridos (estático–dinámico) han alcanzado niveles de precisión superiores al 90% en la identificación de ransomware, malware polimórfico y ataques DDoS, mejorando significativamente los tiempos de detección y respuesta.

Asimismo, la revisión evidencia la transición hacia modelos de seguridad más contextuales y dinámicos, donde el acceso se regula en función del riesgo y la confianza del usuario. Tecnologías emergentes como la ABE y la integración con blockchain consolidan un enfoque más granular, auditable y seguro, especialmente en entornos distribuidos como IoT, SDN y cloud computing.

De igual forma, se identificaron avances importantes en el desarrollo de herramientas de aprendizaje federado y análisis explicable XAI, que permiten mejorar la transparencia de los sistemas y proteger la privacidad de los datos sin sacrificar capacidad predictiva. Estos hallazgos confirman que el futuro de la ciberseguridad está orientado hacia soluciones inteligentes, colaborativas y éticamente responsables.

Desde una perspectiva práctica, esta revisión también deja entrever el valor estratégico del ML y el DL para las organizaciones, su implementación favorece la detección temprana de ataques, la priorización de alertas en SOCs, la automatización de respuestas y el fortalecimiento de la resiliencia operativa. No obstante, persisten desafíos relevantes, como el alto costo computacional, la necesidad de datasets balanceados y actualizados, y la formación continua de talento especializado.

En términos productivos, el proceso de revisión no solo permitió mapear el estado actual del conocimiento, sino también identificar líneas de investigación emergentes con gran potencial de aplicación. Entre ellas destacan el desarrollo de modelos adversariales defensivos, el aprendizaje continuo para contrarrestar el concept drift, y la integración de IA explicable en entornos de decisión crítica.

Finalmente, el estudio incorpora la revisión de casos relevantes, estadísticas recientes y tendencias de investigación, con el fin de comprender cómo la IA está transformando las estrategias de defensa digital en las organizaciones modernas.

Amenazas Cibernéticas Emergentes

En la actualidad, las amenazas cibernéticas evolucionan constantemente, desafiando la seguridad de sistemas y datos a nivel global. Investigaciones recientes han abordado diversas facetas de estas amenazas, proponiendo estrategias y soluciones innovadoras para su detección y mitigación.

- (Al-rimy et al., 2018) identifican factores clave que facilitan el éxito del ransomware, como la explotación de vulnerabilidades humanas (ingeniería social) y técnicas de evasión de detección. Su taxonomía clasifica estos ataques en categorías como crypto-

ransomware (cifrado de datos) y scareware (extorsión psicológica). Para mitigarlos, proponen estrategias multicapa, incluyendo:

- Backups automatizados en entornos aislados.
- Segmentación de redes para limitar la propagación.

- (Kritika, 2025) y (Anand et al., 2022) se centran en la detección de ransomware utilizando aprendizaje profundo y análisis de llamadas a API, respectivamente.
    - Deep Learning: Kritika emplea redes neuronales profundas (DNN) para analizar firmas de ransomware en tiempo real, logrando una precisión del 92% en detección.
    - Análisis de llamadas a API: Anand et al. detectan patrones sospechosos (acceso masivo a archivos) mediante secuencias de llamadas al sistema en Windows, con una tasa de falsos positivos <5%.
- (Bala & Behal, 2024) y (Kalambe et al., 2025) analizan técnicas de IA para detectar ataques DDoS en entornos IoT y SDN, destacando desafíos y soluciones actuales:
    - Redes Neuronales Recurrentes (RNN): Efectivas para predecir picos de tráfico anómalo en SDN, reduciendo el tiempo de respuesta en un 30%.
    - Federated Learning: (Kumar et al., 2024) proponen modelos distribuidos para proteger servicios en la nube sin centralizar datos, ideal para infraestructuras multicloud.
- (Bhardwaj et al., 2021) revisan soluciones científicas y comerciales para ataques DDoS en la nube:
    - Herramientas comerciales: Cloudflare y AWS Shield destacan por su escalabilidad, pero su costo limita su adopción en PYMEs.

- Soluciones basadas en ML: Algoritmos como Isolation Forest detectan anomalías con un 89% de precisión en entornos cloud.

- (Birthriya et al., 2025) analizan el spear phishing, una variante dirigida que utiliza información personalizada para engañar a víctimas específicas (empleados de alto perfil). Sus hallazgos incluyen técnicas de detección:
    - Procesamiento de lenguaje natural (NLP): Modelos como BERT identifican inconsistencias en correos electrónicos (dominios falsos).
    - Análisis de metadatos: Verificación de fuentes y rutas de redirección.
- (Gibert et al., 2020) y (Ucci et al., 2019) discuten el uso de aprendizaje automático para la detección y clasificación de malware, destacando los avances y desafíos en este campo. Es así que al examinar el rol de ML en la lucha contra el malware:
    - Clasificación basada en características:
        - Análisis estático: Detección de firmas en código binario.
        - Análisis dinámico: Monitoreo de comportamientos en sandboxes.
    - Visualización de malware: (Pinhero et al., 2021) convierten código en imágenes para aplicar redes neuronales convolucionales (CNN), alcanzando un 98% de precisión.
- (Hanif et al., 2021) y (Kalouptsoglou et al., 2023) se enfocan en la predicción de vulnerabilidades de software utilizando enfoques de aprendizaje automático, resaltando la importancia de identificar debilidades en el software antes de que puedan ser explotadas. Donde al utilizar ML para identificar vulnerabilidades en código fuente:
    - Modelos como Random Forest predicen riesgos en repositorios de GitHub con un 85% de exactitud.

- Integración de grafos de dependencias para mapear relaciones entre módulos de software.

Las amenazas emergentes exigen soluciones adaptativas que combinen ML, políticas proactivas y colaboración intersectorial. Si bien los estudios demuestran avances significativos en detección de ransomware, mitigación de DDoS y análisis de malware, persisten retos como la escalabilidad en IoT, la sofisticación de ataques adversarios y la necesidad de educación en ciberseguridad. La integración de marcos de zero-trust y el desarrollo de modelos explicables XAI serán claves para construir defensas resilientes en un panorama digital en constante evolución.

Aplicaciones de Aprendizaje Automático en Ciberseguridad

En el contexto de las crecientes amenazas cibernéticas, el ML y el DL han emergido como herramientas fundamentales en la ciberseguridad moderna. Estas tecnologías permiten detectar patrones complejos, automatizar respuestas y anticipar ataques, fortaleciendo así la defensa de sistemas informáticos.

La detección temprana de amenazas es esencial para mitigar riesgos cibernéticos. Investigaciones recientes han explorado modelos de ML y DL para identificar actividades maliciosas en tiempo real, abordando desafíos como la escasez de datos etiquetados y la necesidad de interpretabilidad en los modelos, (Halbouni et al., 2022) y (Vourganas & Michala, 2024) revisan enfoques de aprendizaje automático y profundo para la detección de amenazas cibernéticas, destacando su eficacia y desafíos

La seguridad en la autenticación y autorización de usuarios ha evolucionado con la integración de ML y DL. Estos enfoques mejoran la precisión en la verificación de identidades, aunque enfrentan retos relacionados con la privacidad de los datos y la adaptabilidad a comportamientos de usuarios cambiantes. (Pritee et al., 2024) examinan el uso de aprendizaje

automático y profundo en la autenticación y autorización de usuarios, resaltando avances y limitaciones.

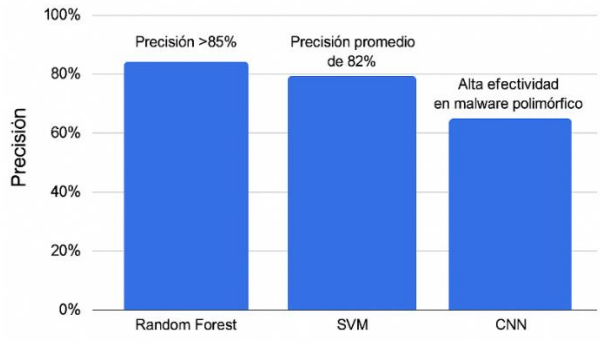
El análisis del comportamiento del usuario es crucial para detectar amenazas internas y actividades engañosas. Modelos de DL han sido aplicados para identificar desviaciones en patrones de uso, permitiendo una detección más efectiva de anomalías que podrían indicar compromisos de seguridad. (Yuan & Wu, 2021) exploran el uso de aprendizaje profundo para la detección de amenazas internas, mientras que (Prome et al., 2024) revisan técnicas de detección de engaños utilizando aprendizaje automático y profundo.

Se confirmó que modelos como Random Forest, Support Vector Machines (SVM) y Redes Neuronales Convolucionales (CNN) ofrecen altos niveles de precisión en la clasificación de software malicioso como es mencionado por (Anand et al., 2022; Dixit & Silakari, 2020; Kritika, 2025), en la figura 6 se realiza una comparación entre los algoritmos del ML.

- Random Forest: Precisión >85%
- SVM: Precisión promedio de 82%
- CNN: Alta efectividad en malware polimórfico

Figura 6

Comparativa de precisión entre algoritmos de ML



Nota: Se muestra la precisión que se tiene los diferentes modelos. Fuente: elaboración propia.

Estas aplicaciones del aprendizaje automático en ciberseguridad destacan su potencial para fortalecer las defensas digitales, aunque también subrayan la necesidad de abordar desafíos técnicos y éticos asociados con su implementación.

Fundamentos y Herramientas de Aprendizaje Automático

Para comprender de manera adecuada las aplicaciones del aprendizaje automático en el ámbito de la ciberseguridad, es indispensable conocer sus fundamentos teóricos y las herramientas que lo sustentan. Esto implica familiarizarse con conceptos esenciales, como los tipos de aprendizaje (supervisado y no supervisado), así como con las tecnologías que hacen posible su funcionamiento, entre ellas las redes neuronales y el DL.

El aprendizaje automático puede dividirse en dos enfoques principales. En el aprendizaje supervisado, los modelos se entrenan con datos previamente etiquetados, lo que les permite predecir resultados específicos a partir de ejemplos conocidos. Este método se aplica con éxito en tareas como la detección de correos electrónicos de spam o el análisis de sentimientos. En contraste, el aprendizaje no supervisado trabaja con datos sin etiquetas, buscando descubrir patrones ocultos o agrupar la información de manera significativa. Este tipo de aprendizaje es particularmente útil en la ciberseguridad para detectar comportamientos anómalos o identificar amenazas desconocidas.

Las redes neuronales artificiales, inspiradas en el funcionamiento del cerebro humano, están compuestas por nodos interconectados que procesan información en diferentes capas. Dentro de este marco, el DL representa una evolución del ML, ya que utiliza redes con múltiples capas ocultas para analizar grandes volúmenes de datos y extraer características complejas. En el campo de la ciberseguridad, estas redes son capaces de identificar patrones sofisticados, adaptarse a

amenazas emergentes y mejorar la precisión de los sistemas de detección. No obstante, su implementación requiere una considerable cantidad de datos y recursos computacionales.

En síntesis, dominar los principios del ML y DL resulta esencial para el desarrollo de soluciones de ciberseguridad más inteligentes, resilientes y adaptativas, capaces de responder de forma efectiva ante el cambiante panorama de amenazas digitales.

Casos Relevantes y Estadísticas

La creciente sofisticación de los ciberataques a nivel global ha dejado una huella profunda en la última década, evidenciada por incidentes de gran impacto como las filtraciones que se presentaron en WikiLeaks en el 2010, la vulneración de la red de Sony PlayStation Network en el 2011, las brechas de seguridad en Dropbox que fueron en el 2012 y Facebook en el 2019. Estos eventos no solo comprometieron millones de datos personales, sino que también marcaron un punto de inflexión en la percepción y gestión de la ciberseguridad a nivel mundial.

En el contexto latinoamericano, Colombia se ha convertido en un objetivo prioritario para los ciberdelincuentes. Según el artículo de la (Económica, 2024) donde los informes de IBM, en 2024 el país concentró el 17% de los ciberataques registrados en la región, siendo el sector salud el más afectado, con un 60% de los incidentes dirigidos a instituciones médicas. Casos emblemáticos incluyen el ataque a Sanitas en diciembre de 2022, que comprometió datos sensibles de aproximadamente 242,000 afiliados y afectó los servicios de más de 5.5 millones de usuarios durante 45 días.

Estos acontecimientos subrayan la urgente necesidad de fortalecer las estrategias de ciberseguridad, especialmente en sectores críticos como el de la salud, donde la protección de la información sensible y la continuidad operativa son fundamentales para garantizar la confianza y seguridad de los ciudadanos.

Investigación y Desarrollo en Ciberseguridad

La investigación en ciberseguridad enfrenta un panorama dinámico marcado por la evolución tecnológica y la aparición de amenazas cada vez más sofisticadas. Los estudios de (Díaz et al., 2019) y (Díaz et al., 2021) ofrecen una visión crítica sobre los desafíos y las adaptaciones necesarias en este campo, especialmente en contextos de crisis global.

(Díaz et al., 2019) identifican que la rápida adopción de tecnologías emergentes, como el Internet de las Cosas (IoT), la IA y la computación en la nube, ha generado vectores de ataque complejos y superficies de riesgo expandidas. Entre los desafíos clave destacan:

- Tecnologías disruptivas: La integración de IA en sistemas críticos introduce vulnerabilidades asociadas a sesgos algorítmicos o ataques adversarios.
- Amenazas híbridas: Combinación de ransomware, phishing dirigido y ataques de día cero (zero-day), que requieren enfoques multidisciplinarios para su mitigación.
- Escasez de talento especializado: La falta de profesionales capacitados en ciberseguridad dificulta la respuesta ágil a incidentes.

Los autores subrayan la necesidad de paradigmas proactivos, como la seguridad por diseño y la colaboración público-privada, para construir sistemas resilientes desde su concepción.

La pandemia de COVID-19, analizada por (Díaz et al., 2021), actuó como un catalizador para redefinir prioridades en ciberseguridad. Entre los hallazgos clave:

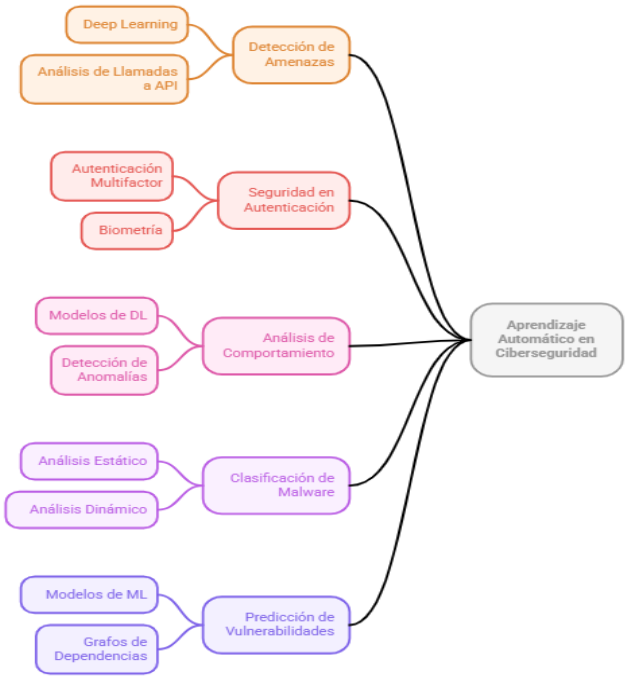
- Aumento de ataques: Se reportó un incremento del 40% en ciberataques a infraestructuras críticas, especialmente en sectores sanitarios y educativos, explotando la transición abrupta al teletrabajo.

- Cambio en enfoques de investigación: La urgencia por proteger redes remotas impulsó el desarrollo de herramientas de detección en tiempo real basadas en ML y protocolos de autenticación multifactor robustos.
- Resiliencia organizacional: Las instituciones adoptaron marcos de cyber hygiene para entrenar a empleados en identificar amenazas como phishing relacionado con la pandemia.

Además, el estudio revela que la pandemia aceleró la adopción de modelos híbridos de trabajo, generando nuevos retos como la protección de dispositivos personales no administrados y la seguridad en entornos cloud-first. Donde en la figura 7 se puede visualizar un resumen del ML en la ciberseguridad.

Figura 7

Aplicaciones de Aprendizaje Automático en Ciberseguridad.



Nota: Se muestra las diferentes fuentes para que el ML en la ciberseguridad. Fuente: elaboración propia.

Análisis de la Revisión Sistemática

En el desarrollo de la revisión sistemática de la literatura permitió identificar de manera estructurada las principales técnicas, algoritmos, datasets, métodos y logros alcanzados en el campo de la detección de malware mediante el uso de ML y DL. A partir de este análisis, se evidencian tres enfoques predominantes para el análisis de malware: el análisis estático, el análisis dinámico y el análisis híbrido.

El análisis estático se fundamenta en la inspección del código sin necesidad de ejecutarlo, permitiendo la extracción de características relevantes como las librerías importadas, cadenas de texto, firmas y niveles de entropía. Este enfoque presenta ventajas importantes en términos de rapidez y bajo costo computacional; sin embargo, su efectividad puede verse limitada frente a técnicas de ofuscación y empaquetamiento utilizadas por el malware moderno. Por otro lado, el análisis dinámico se basa en la ejecución controlada de muestras en entornos sandbox, permitiendo observar comportamientos maliciosos a través del monitoreo de llamadas a API, tráfico de red y actividades del sistema. Aunque este enfoque ofrece una mayor capacidad de detección basada en comportamiento, también enfrenta desafíos asociados a técnicas de evasión avanzadas que impiden la correcta identificación del malware. En este contexto, el análisis híbrido emerge como una solución más robusta al combinar características estáticas y dinámicas, logrando mejorar significativamente la precisión y la capacidad de generalización de los modelos.

En cuanto a los algoritmos utilizados, la literatura evidencia una evolución desde modelos tradicionales de ML hacia arquitecturas más complejas basadas en DL. Algoritmos como Random Forest, Support Vector Machines, K-Nearest Neighbors y modelos de ensamble como XGBoost y LightGBM han demostrado eficiencias superiores al 80%, destacándose por su interpretabilidad y menor costo computacional. No obstante, los enfoques basados en DL, tales como CNN,

RNN/LSTM y GNN, han alcanzado niveles de precisión superiores al 95%, gracias a su capacidad para modelar patrones complejos en datos estructurados, secuenciales y gráficos. Adicionalmente, los modelos híbridos que integran características estáticas y dinámicas han mostrado un mejor equilibrio entre precisión, robustez y capacidad de adaptación frente a amenazas emergentes.

Respecto a los datasets, se identificaron múltiples fuentes de datos ampliamente utilizadas en la comunidad científica, entre las que se destacan EMBER, CICIDS2017, CTU-13 y Malimg. Estos datasets permiten abordar diferentes perspectivas del problema, incluyendo análisis de archivos ejecutables, tráfico de red y representaciones visuales del malware. Sin embargo, uno de los principales desafíos identificados es el desbalance de clases y la falta de representatividad de amenazas actuales, lo que puede afectar la capacidad de generalización de los modelos desarrollados.

En relación con el preprocesamiento de datos y la ingeniería de características, se evidenció la aplicación de técnicas como la normalización Min-Max y Z-score, la eliminación de valores atípicos y la codificación de variables categóricas. Asimismo, métodos de selección de características como PCA, Recursive Feature Elimination y técnicas basadas en la importancia de variables han permitido reducir la dimensionalidad de los datos y mejorar la eficiencia de los modelos. De igual forma, el uso de representaciones avanzadas como embeddings de secuencias y grafos de flujo de control ha contribuido significativamente a capturar patrones complejos asociados al comportamiento del malware.

Un aspecto relevante identificado en la literatura es el uso de técnicas de data augmentation para abordar problemas de desbalance y mejorar la capacidad de generalización. Estas técnicas incluyen transformaciones en imágenes de malware, generación de secuencias sintéticas, métodos programáticos como MARVOLO y el uso de GANs para la creación de muestras evasivas. Estos

enfoques han demostrado ser especialmente efectivos frente a malware polimórfico, aumentando la robustez de los modelos frente a nuevas variantes.

En cuanto a la evaluación de los modelos, las métricas más utilizadas incluyen accuracy, precision, recall, F1-score y AUC-ROC. No obstante, en el contexto de ciberseguridad, el recall adquiere un papel prioritario, ya que permite minimizar la cantidad de amenazas no detectadas. Los estudios revisados evidencian que, si bien es posible alcanzar altos niveles de precisión, existe un trade-off con la tasa de falsos positivos, lo cual representa un desafío importante en entornos operacionales.

Los logros reportados en la literatura reflejan avances significativos en términos de rendimiento, robustez y explicabilidad. Los modelos basados en DL han alcanzado precisiones superiores al 95%, mientras que los modelos de ensamble han logrado eficiencias superiores al 85% con menor costo computacional. Asimismo, la incorporación de técnicas de IA Explicable, como SHAP, LIME y Grad-CAM, ha permitido mejorar la interpretabilidad de los modelos, facilitando su adopción en entornos organizacionales. Adicionalmente, el uso de redes neuronales de grafos y enfoques adversariales ha permitido fortalecer la detección frente a técnicas de evasión avanzadas.

Sin embargo, a pesar de estos avances, la literatura también evidencia importantes brechas de investigación. Entre ellas se destacan la falta de datasets actualizados, la presencia de fenómenos de concept drift, la vulnerabilidad frente a ataques adversariales y la limitada interpretabilidad de modelos complejos. Estas limitaciones justifican la necesidad de desarrollar enfoques más robustos, adaptativos y explicables.

En este sentido, los hallazgos obtenidos a partir de la revisión sistemática proporcionan una base sólida para el desarrollo de la presente investigación, orientando la selección de técnicas,

modelos y estrategias metodológicas. En particular, se evidencia la pertinencia de utilizar enfoques híbridos, incorporar técnicas de data augmentation, aplicar métodos avanzados de ingeniería de características y evaluar los modelos mediante métricas robustas que prioricen la detección efectiva de amenazas. De esta manera, se establece un fundamento técnico que respalda la propuesta planteada y permite avanzar hacia el desarrollo de soluciones más eficientes para la detección de malware en entornos organizacionales.

Para finalizar, la revisión sistemática de literatura desarrollada permitió examinar de manera estructurada y crítica la documentación existente sobre el uso del ML en el análisis de malware, cumpliendo así el objetivo propuesto. A partir de este proceso, fue posible identificar tendencias relevantes, como el crecimiento del uso de modelos híbridos, el incremento en la aplicación de técnicas de DL (CNN, RNN y GNN) y la incorporación de enfoques basados en explicabilidad que fortalecen la confianza en los sistemas de detección.

De igual manera, se evidenciaron enfoques innovadores orientados a mejorar la eficacia de los modelos, tales como el uso de data augmentation semántico y visual, la integración de análisis estático, dinámico e híbrido, y la utilización de representaciones avanzadas como grafos de flujo de control y secuencias de llamadas API. Estas prácticas han permitido alcanzar mejoras significativas en métricas de desempeño como precisión, recall y F1-score, especialmente en la detección de amenazas complejas como ransomware, malware polimórfico y ataques distribuidos.

Asimismo, la revisión permitió consolidar un conjunto de prácticas destacadas que sirven como base técnica para el desarrollo de soluciones en entornos organizacionales, entre ellas la adecuada curación y balanceo de datasets, la selección de características relevantes, la validación robusta de modelos y la integración con infraestructuras de seguridad como SIEM, EDR e IDS/IPS.

En conjunto, los hallazgos obtenidos no solo evidencian el avance y la madurez del uso del ML en la ciberseguridad, sino que también proporcionan una base sólida, estructurada y actualizada para el diseño de propuestas orientadas a la detección, prevención y mitigación del malware en las organizaciones. De esta manera, el capítulo no solo cumple con el análisis del estado del arte, sino que establece un fundamento técnico y metodológico claro que guía el desarrollo de los objetivos posteriores de la investigación.

Estrategias de análisis de malware basadas en ML: evaluación estática, dinámica e híbrida y su aplicabilidad en entornos organizacionales

En el contexto actual de la ciberseguridad, las organizaciones enfrentan una evolución constante de amenazas digitales que desafían los métodos tradicionales de detección. Los ciberataques modernos más sofisticados, evasivos y persistentes, exigen estrategias analíticas capaces de anticipar, identificar y mitigar comportamientos maliciosos con rapidez y precisión. En este marco, los enfoques de análisis estático, dinámico e híbrido se consolidan como pilares fundamentales para comprender el comportamiento del malware y fortalecer las defensas organizacionales.

Se presenta una revisión integral de estos tres enfoques, abordando sus principios, técnicas, ventajas, limitaciones y aplicaciones prácticas en entornos corporativos. A través del estudio de diversas investigaciones recientes, se busca destacar cómo las técnicas de ML y DL han potenciado cada tipo de análisis, contribuyendo a la detección temprana, la prevención de infecciones y la mitigación de impactos en los sistemas organizacionales. Asimismo, se examina el papel del análisis híbrido como estrategia clave para enfrentar amenazas avanzadas, garantizando una defensa más adaptativa, explicable y resiliente frente a un panorama de ciberataques en constante transformación.

Análisis Estático

El análisis estático consiste en examinar el software malicioso sin ejecutarlo, permitiendo identificar componentes como funciones, cadenas de texto y estructuras de archivo, normalmente mediante ingeniería inversa y descompilación. Esto ayuda a descubrir posibles amenazas, indicadores de compromiso y vectores de ataque sin riesgo de infectar sistemas productivos. El análisis estático es útil para encontrar variantes conocidas mediante firmas, pero puede ser menos

efectivo ante malware ofuscado o desconocido. Siendo así se puede observar cada uno de los modelos que trabajan:

Extracción de características PE / Drebin / manifest

- Windows PE: En ejecutables Windows, las características estáticas más utilizadas incluyen: tamaño y entropía de secciones, tabla de imports, hashes de secciones, cadenas estáticas e histogramas de bytes. Estas características alimentan modelos tradicionales (Random Forest, XGBoost) así como redes neuronales. Varios estudios señalan que, sobre conjuntos públicos de referencia, los clasificadores basados en estas representaciones alcanzan buenos niveles de precisión y son eficientes para pre-filtrar grandes volúmenes de ficheros (Gibert et al., 2020; Ucci et al., 2019). En la práctica, conjuntos públicos como EMBER disponibles en repositorios tipo Kaggle se utilizan frecuentemente para entrenamiento y benchmarking.
- Android: Para aplicaciones Android la extracción estática suele considerar: permisos, intents, componentes del manifiesto, llamadas a API declaradas y secuencias de opcodes. Estos atributos han permitido entrenar clasificadores (SVM, árboles y modelos CNN/MLP) orientados a detectar familias móviles (Guerra-Manzanares, 2024; Razgallah et al., 2021). Conjuntos como CICMalDroid o colecciones anotadas de APK sirven para validar estos enfoques.

Grafos y flujo de control

El uso de representaciones gráficas (Control Flow Graphs, Call Graphs) y la extracción de subestructuras (walks, embeddings) ha demostrado ser especialmente útil para capturar relaciones semánticas que no se aprecian en vectores planos. Modelos profundos que consumen CFGs o

trazas de flujo (ya sea desde el binario o desde la ejecución controlada) muestran mayor robustez frente a pequeñas variaciones sintácticas del malware y reducen falsos negativos en familias Windows complejas(Qiang et al., 2022). Estas técnicas, no obstante, requieren mayor coste computacional y sofisticación en el preprocesamiento.

Visualización y Explainable AI (XAI)

Algunas líneas de investigación convierten binarios en representaciones visuales (mapas de bytes) y aplican CNNs, lo que facilita además la interpretación mediante mapas de activación y técnicas XAI (por ejemplo, SHAP, LIME). La combinación de visualización y métodos explicables ayuda a los analistas a entender por qué una muestra fue clasificada como maliciosa, favoreciendo la confianza operativa y la trazabilidad de decisiones en SOCs (Elhanashi A.; Dini P., 2024; Pinhero et al., 2021).

Las fortalezas en el análisis estático son rápido, escalable y seguro (no ejecuta código peligroso), lo que lo hace idóneo para pre-screening masivo en correo electrónico, repositorios o endpoints. Las técnicas estáticas permiten establecer pipelines automatizados que filtran la mayoría de material benigno antes de costosas detonaciones dinámicas (Gibert et al., 2020).

Las debilidades son la principal vulnerabilidad es la evasión de las técnicas de packing, cifrado de secciones, polimorfismo/metamorfismo y manipulación de características pueden degradar drásticamente el rendimiento de detectores estáticos. Además, los atacantes pueden lanzar ataques dirigidos al espacio de características (feature-space attacks) para engañar a modelos ML (Ling et al., 2023). Por ello, se recomienda enfoques híbridos (estático + dinámico), evaluaciones adversariales y el uso de features resistentes para mejorar la robustez.

Análisis Dinámico

El análisis dinámico, por otro lado, ejecuta el malware en un entorno virtualizado (sandbox) para observar su comportamiento en tiempo real, por ejemplo, interacciones con el sistema de archivos, intentos de cifrado, comunicación con servidores externos o movimientos laterales en la red. Sirve para detectar actividad sospechosa y mecanismos evasivos, aunque el malware avanzado puede camuflarse para no activar sus funciones en sandboxes.

Secuencias de llamadas API y comportamiento del sistema

El monitoreo de llamadas API es uno de los enfoques más sólidos para detectar ransomware y troyanos avanzados, ya que permite observar accesos a archivos, rutinas criptográficas, modificaciones en el registro y comunicaciones de red.

Plataformas como Cuckoo Sandbox o módulos de Endpoint Detection and Response (EDR) generan trazas detalladas de llamadas que, al ser analizadas con ML, permiten detectar fases de cifrado o comportamientos anómalos antes de que se produzca un daño irreversible.

Estudios recientes muestran que combinar firmas tradicionales con modelos ML entrenados sobre secuencias de API aumenta la tasa de detección temprana de ransomware en escenarios de producción (Anand et al., 2022).

Tráfico de red y telemetría

Muchas familias de malware dependen de comunicación con servidores de Comando y Control (C2). El análisis de tráfico (NetFlow, pcap, logs de red) se convierte en una fuente rica para modelos de ML y DL. Conjuntos de datos públicos como CTU-13, compilado por Stratosphere IPS, permiten entrenar modelos que identifican botnets, anomalías y patrones de DDoS (García et al., 2014).

Estos métodos son especialmente útiles para la detección en SOCs/IDS, al integrarse con flujos de red reales y permitir segmentar dispositivos comprometidos, reducir superficie de ataque y mejorar la respuesta (Bala & Behal, 2024).

Trazas de ejecución / flujo de control

El registro de trazas de ejecución (instrucciones, ramas de control, dependencias de datos) permite alimentar modelos profundos (DNNs, RNNs) que identifican comportamientos maliciosos incluso si el binario fue modificado superficialmente.

Estos modelos son más robustos frente a variantes polimórficas porque detectan patrones de ejecución en lugar de atributos superficiales. (Qiang et al., 2022) demostraron que aplicar redes neuronales profundas sobre trazas de flujo de control logra alta precisión en la detección de malware de Windows PE.

La evasión dinámica en el malware moderno incluye técnicas para detectar entornos virtualizados o de depuración (anti-VM, anti-debugging). Algunas variantes retrasan su activación (delayed execution) o requieren interacción humana para desplegar su carga. Estas limitaciones obligan a las organizaciones a utilizar sandboxes realistas, idealmente en entornos bare-metal o con instrumentación ligera, y a diseñar escenarios de ejecución interactivos que simulen condiciones reales (Afianian et al., 2018; Ling et al., 2023). El coste computacional es mayor que en el análisis estático, y la escalabilidad depende de la capacidad de orquestación en infraestructura de seguridad.

Análisis Híbrido

El análisis híbrido combina las fortalezas del análisis estático y dinámico para mejorar la cobertura, la robustez y la precisión en la identificación de software malicioso. Esta integración

responde a la necesidad de superar las limitaciones de cada enfoque cuando se usan de manera aislada (Ucci et al., 2019; Gibert et al., 2020).

Preprocesamiento estático

Extracción de features como cabeceras PE, permisos en Android, cadenas de texto y vectores de importación. Esto permite clasificar grandes volúmenes de muestras de forma rápida (Bensaoud et al., 2024; Dixit & Silakari, 2020).

Validación dinámica

Ejecución controlada en sandbox para observar el comportamiento real de la muestra (llamadas API, tráfico de red, cifrado de archivos, persistencia). Este paso identifica técnicas de evasión o cargas diferidas (Anand et al., 2022).

Integración en un pipeline automatizado

- Fase 1: Filtrado masivo con clasificadores estáticos → bajo costo computacional.
- Fase 2: Validación de casos sospechosos en entornos dinámicos.
- Fase 3: Enriquecimiento con inteligencia de amenazas (TI) y retroalimentación de modelos ML/DL.

Ventajas del análisis híbrido

El análisis híbrido surge como una estrategia superior al combinar las ventajas del análisis estático y dinámico, mitigando de forma significativa sus limitaciones individuales.

Reducción de falsos positivos y negativos. La capacidad de los atacantes para evadir los mecanismos de detección es un desafío constante en ciberseguridad. El análisis estático puede ser fácilmente engañado mediante técnicas de empaquetado, ofuscación y generación polimórfica de código, lo que reduce su fiabilidad (Kritika, 2025). Por su parte, el análisis dinámico se ve limitado por técnicas de anti-VM y anti-debugging, donde el malware retrasa o modifica su ejecución al

detectar un entorno controlado (Ling et al., 2023). La combinación híbrida reduce estas brechas, permitiendo correlacionar indicadores de estructura con patrones de comportamiento, incrementando así la tasa de detección real.

Escalabilidad y profundidad en la evaluación. El análisis híbrido integra la escalabilidad del análisis estático con la profundidad del análisis dinámico. Por un lado, el procesamiento estático permite evaluar miles de muestras en pocos minutos, generando un prefiltrado inicial con técnicas como extracción de opcodes, CFGs o permisos. Por otro lado, la ejecución en sandboxes dinámicos provee un nivel de detalle mayor, revelando trazas de API, comunicaciones de red y persistencia del malware que complementan los hallazgos estáticos (Anand et al., 2022; Hernández-Pereira, 2022).

Soporte integral para investigación forense y respuesta temprana. Otra ventaja del enfoque híbrido es su aplicabilidad tanto en respuesta inmediata a incidentes como en investigación forense posterior. Durante la fase de ataque, los modelos híbridos permiten detectar comportamientos anómalos en tiempo cercano al real, mientras que en la fase de análisis forense contribuyen a identificar indicadores de compromiso más robustos, incluyendo hashes, opcodes, CFGs, patrones de tráfico C2, rutinas de cifrado y mecanismos de persistencia (Gibert et al., 2020; Ucci et al., 2019).

Aplicabilidad en organizaciones

El análisis híbrido de malware tiene un alto potencial de implementación en entornos organizacionales, donde la seguridad informática debe equilibrar la rapidez en la detección con la profundidad en el análisis. Su aplicabilidad se manifiesta en distintos niveles, en la figura 8, se puede validar la aplicabilidad en las organizaciones:

Gateways de correo electrónico y endpoints. En esta capa, el análisis híbrido fortalece la protección contra vectores de ataque comunes como adjuntos maliciosos y descargas no autorizadas. El proceso inicia con un análisis estático basado en ML, capaz de filtrar de forma rápida y escalable grandes volúmenes de archivos. Posteriormente, las muestras sospechosas se someten a sandboxing dinámico, donde se confirma su comportamiento malicioso, especialmente útil para la detección de ransomware y troyanos de acceso remoto(Anand et al., 2022; Kritika, 2025).

Centros de Operaciones de Seguridad (SOC) y Equipos de Respuesta a Incidentes (CERT). En el ámbito de gestión centralizada, los modelos híbridos se integran en sistemas SIEM (Security Information and Event Management). Allí permiten correlacionar indicadores estáticos, como firmas, hashes y estructuras de código, con indicadores dinámicos, tales como eventos de red, secuencias de llamadas API y patrones de persistencia. Esta correlación facilita una detección más temprana y una respuesta orquestada y priorizada (Gibert et al., 2020; Ucci et al., 2019).

Sectores de infraestructura crítica (salud, energía, banca). Las organizaciones de sectores estratégicos requieren un nivel de protección superior frente a amenazas avanzadas, particularmente ransomware dirigido y ataques persistentes avanzados (APT). En estos escenarios, el análisis híbrido se implementa con sandboxes bare-metal, que evitan las técnicas de evasión de entornos virtualizados, y se combina con modelos de ML entrenados sobre tráfico de red y telemetría de sistemas. Esto proporciona defensas proactivas y adaptativas, reforzando la continuidad del negocio y minimizando el impacto de ciberataques(Al-rimy et al., 2018; Económica, 2024).

Figura 8

Aplicabilidad en organizaciones.

Nota: Aplicabilidad en organizaciones reforzando la continuidad del negocio y minimizando el impacto de ciberataque. Fuente: elaboración propia.

Limitaciones y desafíos

A pesar de los avances que supone el análisis híbrido, su implementación en entornos organizacionales y de investigación enfrenta varias limitaciones y retos técnicos que deben ser considerados:

Costo computacional. La ejecución de muestras en entornos dinámicos, como sandboxes o plataformas de emulación, requiere un alto consumo de recursos de hardware y software. Esto incluye infraestructura de aislamiento, orquestación de pruebas y capacidad de almacenamiento para trazas y registros generados. En organizaciones con gran volumen de muestras diarias, los costos de escalabilidad pueden representar un obstáculo significativo (Anand et al., 2022; Elhanashi A.; Dini P., 2024).

Complejidad de integración. La fusión de pipelines estáticos y dinámicos implica estandarizar los formatos de características extraídas (features), lo cual no siempre es sencillo. Por

ejemplo, la normalización de atributos como vectores de bytes, llamadas API, grafos de flujo de control y métricas de red requiere arquitecturas de procesamiento unificadas, lo que incrementa la complejidad de desarrollo y mantenimiento de los sistemas (Bensaoud et al., 2024; Hernández-Pereira, 2022).

Evasión avanzada. Los atacantes evolucionan constantemente sus técnicas de evasión, utilizando malware polimórfico, empaquetadores avanzados y estrategias de adversarial ML que manipulan las características de entrada para confundir a los clasificadores. Estas técnicas reducen la efectividad de los modelos entrenados y obligan a diseñar sistemas adaptativos y resilientes (Kritika, 2025; Ling et al., 2023).

El análisis híbrido representa un enfoque robusto y completo, su adopción masiva depende de superar retos de costo, integración y resistencia frente a evasión avanzada, aspectos que actualmente concentran los mayores esfuerzos de investigación en ciberseguridad basada en ML. En la figura 9, se observa una comparación entre el análisis estático, dinámico e hibrido.

Figura 9

Comparación de los análisis

Nota: Comparación del análisis estático, dinámico e hibrido. Fuente: elaboración propia.

Aplicabilidad y efectividad en la organización

La adopción de técnicas de análisis de malware en entornos organizacionales debe responder tanto a criterios técnicos como a consideraciones estratégicas de negocio. La literatura coincide en que no existe un método único y universal; la efectividad depende de los objetivos institucionales, el tipo de amenaza predominante y los recursos disponibles (Elhanashi A.; Dini P., 2024; Halbouni et al., 2022). Donde se valida las diferentes aplicabilidades de acuerdo al análisis que se utilizan:

Elección según objetivo y entorno

- Objetivo del análisis
    - Detección temprana: análisis estático con modelos de clasificación masiva, que permiten filtrar rápidamente muestras sospechosas (Ucci et al., 2019).
    - Respuesta a incidentes: análisis dinámico ágil para identificar vectores de ataque y frenar la propagación.
    - Investigación forense: análisis estático profundo con ingeniería inversa, binarios y CFGs.
    - Gestión proactiva: monitoreo de telemetría y análisis híbrido para anticipar ataques (Gibert et al., 2020).
- Entorno organizacional
    - Empresas con alto grado de madurez en ciberseguridad → despliegue de sandboxes avanzados, integración con SIEM y automatización de análisis en SOC con capacidades de Threat Hunting.

    - Empresas con recursos limitados → uso de servicios externos como VirusTotal, Hybrid Analysis o soluciones en la nube con ML integrado (Singh et al., 2024).
- Tipo de amenaza predominante
    - Ransomware: requiere monitoreo dinámico de llamadas API y trazas de cifrado (Anand et al., 2022; Kritika, 2025).
    - Troyanos y spyware: análisis estático de PE y tráfico en red.
    - DDoS: análisis dinámico de datasets de tráfico (CTU-13, CIC-IDS2017).
    - Spear phishing: ML para clasificación de correos, adjuntos y enlaces maliciosos (Birthriya et al., 2025).

Criterios técnicos y estratégicos para la selección de métodos

- Requisito de profundidad: herramientas avanzadas como IDA Pro, Ghidra o Radare2 para ingeniería inversa, frente a plataformas rápidas como VirusTotal o Hybrid Analysis para escaneo masivo.
- Tipo de evasión empleada: en malware con anti-VM o anti-debugging, conviene usar sandboxes bare-metal con orquestación dinámica (Afianian et al., 2018).
- Nivel de experiencia del personal: analista sénior pueden aplicar técnicas manuales y detalladas; personal júnior puede apoyarse en servicios automatizados y flujos en la nube.
- Impacto en el negocio:
    - Respuesta rápida: análisis dinámico.

- Investigación postmortem: análisis estático profundo.
- Planificación estratégica: modelos híbridos y Threat Intelligence para reforzar defensas.

Casos reales y simulados bajo marco conceptual

- Ransomware en simulaciones controladas: (Anand et al., 2022) demostraron que la integración de ML con secuencias de API permite detectar cifrado incluso en variantes evasivas.
- Formación en entornos simulados: programas basados en gamificación han mostrado mejoras en la respuesta de usuarios y analistas frente a phishing y malware polimórfico (Tafur-Arciniegas et al., 2023).
- Lecciones de ciberataques reales: incidentes como WannaCry y ataques a infraestructuras críticas evidencian que marcos como MAGERIT y OSSTMM fortalecen la resiliencia organizacional (Contreras & Contreras, 2023).

Casos reales y simulados del uso de Machine Learning en detección y mitigación de malware

La evolución del malware, caracterizada por variantes polimórficas, técnicas de evasión anti-VM y ataques dirigidos (e.g., ransomware), ha impulsado la necesidad de métodos más inteligentes de detección y mitigación. En este contexto, ML y DL han mostrado efectividad tanto en escenarios reales como simulados, aportando precisión y adaptabilidad. La evidencia de casos recientes y extraer criterios técnicos y estratégicos para orientar la implementación organizacional de estos métodos. A continuación, se valida cada uno de los casos de estudio, en la tabla 2 se visualiza el resumen de los casos de estudio:

- Caso 1. (Análisis de llamadas API para ransomware) este estudio se desarrolló en un entorno controlado de laboratorio, simulando ataques de ransomware en sistemas Windows. La investigación demostró que el análisis de secuencias de llamadas API permite identificar las fases de cifrado antes de que se produzca un daño masivo en los archivos. Los modelos de ML utilizados, como Random Forest y XGBoost, alcanzaron una precisión superior al 85%, mostrando gran potencial para la detección temprana. Estratégicamente, se recomienda integrar este tipo de análisis en los sistemas de monitoreo de seguridad (SOC) y en soluciones de Endpoint Detection and Response (EDR) que capturen actividad en tiempo real.

Referencia: A comprehensive API call analysis for detecting Windows-based ransomware.(Anand et al., 2022).

- Caso 2. (Detección basada en trazas de flujo de control) en este caso, se realizaron simulaciones sobre conjuntos de datos de malware para sistemas Windows. Los investigadores entrenaron redes neuronales profundas (DNN) utilizando trazas de ejecución, logrando una precisión cercana al 95,7%. Este enfoque destaca por su alta robustez frente al malware polimórfico, que suele evadir los métodos tradicionales de detección. Desde un punto de vista estratégico, se recomienda su adopción en infraestructuras críticas, donde las amenazas suelen emplear técnicas de evasión avanzadas.

Referencia: Efficient and robust malware detection based on control flow traces using deep neural networks. (Qiang et al., 2022).

- Caso 3. (Malware adversarial con GANs) la simulación de entornos adversariales mediante herramientas como MalFox y MalGAN permitió evaluar la resistencia de los modelos de

detección frente a malware generado artificialmente. Los resultados mostraron que estas muestras pueden reducir la tasa de detección en más del 50%, revelando una debilidad significativa en los clasificadores convencionales. Como medida estratégica, se recomienda incluir fases de adversarial training en los ciclos de desarrollo de detectores de malware, con el fin de fortalecer su resiliencia ante ataques generativos.

Referencia: MalFox: Camouflaged adversarial malware example generation based on Conv-GANs against black-box detectors.(Zhong et al., 2024).

- Caso 4. (Sandboxes bare-metal para malware evasivo) este caso abordó la ejecución de malware en entornos bare-metal, es decir, en hardware físico en lugar de entornos virtualizados. La investigación evidenció que este tipo de análisis permite detectar comportamientos maliciosos que suelen permanecer ocultos en sandboxes tradicionales. Su principal ventaja radica en la identificación de malware evasivo, lo que lo convierte en una estrategia prioritaria para organizaciones que operan en sectores de alta sensibilidad, como finanzas y energía.

Referencia: *BareCloud: Bare-metal analysis-based evasive malware detection.* USENIX Security Symposium.(Kirat et al., 2014).

- Caso 5. (Representación visual de binarios) en este experimento, los investigadores transformaron ejecutables en representaciones visuales mapas de bytes y aplicaron modelos de DL, como CNN, junto con clasificadores SVM. En datasets públicos como Malimg y EMBER, estos modelos lograron una precisión superior al 98%. Esta técnica se considera ideal para procesos de prefiltrado masivo, como el análisis de adjuntos en correos electrónicos o descargas sospechosas.

Referencia: Malware detection employed by visualization and deep neural network.(Pinhero et al., 2021).

- Caso 6. (Detección de tráfico de red y botnets “CTU-13”) el estudio utilizó el conjunto de datos CTU-13, desarrollado por Stratosphere Lab, que contiene tráfico de red real mezclado con actividad de botnets. Los modelos de ML aplicados demostraron una alta capacidad de recall, identificando con precisión las comunicaciones con servidores de comando y control (C2) y los ataques DDoS. Este enfoque resulta fundamental para la protección de organizaciones con exposición pública, especialmente cuando se integra en sistemas SOC o IDS.

Referencia: CTU-13 Dataset: Botnet network traffic with normal and background traffic.(García et al., 2014).

- Caso 7. (GNN con documentación aumentada) el último caso analizó el uso de Graph Neural Networks (GNN) enriquecidas con información semántica de documentación de APIs. Esta técnica demostró mejoras sustanciales en la detección de variantes evasivas de malware, al permitir que los modelos comprendan no solo la estructura del código, sino también el contexto funcional de cada componente. Por su complejidad, esta aproximación se considera especialmente útil para equipos de investigación y desarrollo en ciberseguridad avanzada.

Referencia: DawnGNN: Documentation augmented windows malware detection using graph neural network.(Feng et al., 2024).

Tabla 2 *Comparativa de casos*

Caso	Técnica aplicada	Tipo (real/simulado)	Precisión/Resultado clave	Aplicabilidad estratégica
1. Ransomware (API calls)	Random Forest, XGBoost	Simulación	>85%	SOC/EDR detección temprana
2. Trazas de flujo de control	DNN	Simulación	≈95.7%	Infraestructuras críticas
3. Malware adversarial (GANs)	Conv-GANs	Simulación	Reducción detección >50%	*Adversarial training*
4. Bare-metal sandboxes	Entorno físico	Real aplicado	Detecta malware evasivo	Sectores críticos (finanzas, energía)
5. Visualización binaria	CNN + SVM	Simulación	>98%	Pre-screening masivo
6. Tráfico C2 (CTU-13)	ML sobre NetFlow	Real	Alta recall en botnets	SOC/IDS exposición pública
7. GNN enriquecidas	Graph Neural Networks	Investigación avanzada	Mayor resiliencia evasiva	I+D ciberseguridad

Nota: Se visualiza los 7 casos expuestos. Fuente: elaboración propia

Estos casos muestran que el uso de ML y DL en escenarios reales y simulados fortalece las defensas organizacionales contra malware avanzado. El análisis híbrido (estático + dinámico) se configura como la estrategia más robusta, ya que permite detección temprana, reduce falsos positivos y negativos y favorece la investigación forense. La elección del enfoque depende de criterios técnicos (tipo de dato, robustez, costo) y estratégicos (sector, madurez en ciberseguridad, recursos humanos y tecnológicos).

El presente capítulo permitió analizar de manera integral los principales enfoques empleados en la detección, prevención y mitigación de software malicioso, abordando los métodos

estático, dinámico e híbrido, así como su evolución impulsada por las técnicas de ML y DL. Esta revisión evidenció que la ciberseguridad contemporánea requiere soluciones más adaptativas y predictivas, capaces de anticipar el comportamiento de amenazas cada vez más complejas, como el ransomware, el spyware o los ataques basados en IA adversarial.

El análisis estático demostró ser una herramienta fundamental por su rapidez, escalabilidad y capacidad para realizar un filtrado masivo de archivos potencialmente peligrosos. Su implementación es ideal en procesos de preclasificación y detección temprana dentro de entornos corporativos. Por su parte, el análisis dinámico ofrece una visión más profunda del comportamiento real del malware durante su ejecución, aportando información crítica sobre sus mecanismos de cifrado, evasión y comunicación con servidores externos. Finalmente, el análisis híbrido se consolida como una estrategia superior, al integrar la precisión del análisis estático con la observación conductual del análisis dinámico, fortaleciendo así la capacidad de respuesta ante ataques avanzados.

Asimismo, los casos reales y simulados estudiados demostraron la efectividad del ML y DL en distintos escenarios de aplicación. Modelos como Random Forest, DNN, CNN y GNN alcanzaron niveles de precisión superiores al 90 %, confirmando su potencial para la detección temprana de amenazas, la mitigación de ataques en tiempo real y la investigación forense digital. Estas tecnologías no solo mejoran la eficacia operativa de los sistemas de ciberseguridad, sino que también contribuyen a la toma de decisiones estratégicas orientadas a la protección de activos críticos y a la reducción del riesgo organizacional.

No obstante, se identificaron retos técnicos y operativos que deben ser considerados para su implementación práctica, como el elevado costo computacional del análisis dinámico, la complejidad de integrar pipelines híbridos y la constante evolución de técnicas de evasión y

manipulación adversarial. Superar estos desafíos requerirá el desarrollo de modelos explicables, escalables y con capacidad de aprendizaje continuo, garantizando así la sostenibilidad de las soluciones en entornos productivos. Como se visualiza en la figura 10 el desafío en el análisis del malware.

Figura 10

Desafíos en el Análisis de Malware.

Nota: Se visualiza un resumen sobre los desafíos que se tiene en la detección del malware. Fuente: elaboración propia

Para finalizar, el capítulo evidencia que, a partir de la evaluación comparativa realizada, que no todas las técnicas presentan el mismo nivel de efectividad frente a las amenazas actuales, sino que su desempeño depende del contexto de aplicación y del tipo de dato analizado. Los resultados obtenidos permiten identificar que los modelos basados en análisis híbrido constituyen la alternativa óptima, al integrar características estáticas (como estructuras PE, permisos o grafos de flujo) con información dinámica (como secuencias de llamadas API y tráfico de red), logrando

así mayor precisión, reducción de falsos positivos y mayor capacidad de generalización frente a variantes polimórficas.

En particular, técnicas como Random Forest y XGBoost demostraron ser altamente eficientes en fases de prefiltrado por su bajo costo computacional y buen rendimiento (>85%), mientras que modelos profundos como DNN y CNN alcanzaron niveles superiores de precisión (>95%) al analizar comportamiento y representaciones complejas, siendo más robustos ante técnicas de evasión. Asimismo, enfoques avanzados como GNN evidenciaron una alta capacidad para capturar relaciones semánticas del malware, posicionándose como soluciones prometedoras en escenarios de investigación avanzada. Estos hallazgos confirman que la combinación estratégica de técnicas tradicionales y modelos de aprendizaje profundo, articuladas en pipelines híbridos, constituye la opción más efectiva para entornos organizacionales, ya que permite equilibrar escalabilidad, profundidad analítica y resiliencia.

En este sentido, la evaluación desarrollada no solo valida el cumplimiento del objetivo específico, sino que también establece criterios técnicos claros para la selección e implementación de modelos de ML en la detección, prevención y mitigación de malware, consolidando una base sólida para su aplicación en contextos reales.

Diseño de modelos predictivos basados en Machine Learning para detección de malware con calidad, representatividad y equilibrio

En el contexto actual de la ciberseguridad, donde las amenazas evolucionan con una velocidad sin precedentes, la construcción de modelos predictivos eficaces depende en gran medida de la calidad, diversidad y representatividad de los datos utilizados para su entrenamiento. Los sistemas de ML y DL solo pueden ser tan precisos y confiables como los conjuntos de datos que los alimentan; por ello, la mejora de la calidad de los datasets se ha convertido en un pilar fundamental para el desarrollo de soluciones efectivas en la detección, prevención y mitigación del malware.

Este capítulo aborda de manera integral las estrategias que buscan optimizar la calidad de los datos y el rendimiento de los modelos predictivos aplicados a la ciberseguridad. Se presentan técnicas de data augmentation semántico y visual, métodos de normalización contextual, procedimientos de selección de características orientadas a amenazas modernas y la propuesta de un pipeline predictivo mejorado.

A través de esta exploración, se busca demostrar cómo la combinación de estas técnicas no solo mejora la capacidad de detección de amenazas, sino que también promueve la creación de sistemas más robustos, adaptativos y transparentes. La finalidad es ofrecer un marco de trabajo que integre buenas prácticas de ingeniería de datos con enfoques avanzados de aprendizaje automático, contribuyendo a fortalecer las defensas digitales de las organizaciones frente a un ecosistema de amenazas cada vez más complejo.

Mejora de la calidad y representatividad del dataset

Uno de los principales retos en la construcción de modelos predictivos para la detección de malware es la limitada representatividad y el desbalance de clases en los datasets. Estos problemas reducen la capacidad de generalización de los algoritmos y generan un aumento en falsos positivos y negativos. Para mitigar estas limitaciones, se han desarrollado distintas técnicas de data augmentation que enriquecen los conjuntos de datos con variantes adicionales y más representativas. Las cuales se describen a continuación:

Data augmentation semántico (binarios)

Técnicas como MARVOLO permiten generar muestras de malware con modificaciones sintácticas, por ejemplo, cambios en secciones de código o en estructuras de bytes que mantienen la funcionalidad original, pero introducen diversidad semántica en el dataset. Esta aproximación ha demostrado mejorar la capacidad de generalización de los modelos de ML, aumentando la precisión en hasta un 5 % respecto a modelos entrenados con datasets originales (Wong et al., 2022).

Data augmentation visual (binarios como imágenes)

Otra estrategia es la transformación de binarios a representaciones visuales en forma de mapas de bytes, lo que permite aplicar técnicas clásicas de aumento de datos en visión por computador, como rotación, escalado o la adición de ruido. Este enfoque, combinado con arquitecturas CNN y LSTM, ha mostrado un incremento notable en el rendimiento de los clasificadores, alcanzando precisiones de hasta un 98.5 % en conjuntos como Malimg (McLaughlin & del Rincon, 2022; Walia, 2021). Asimismo, el uso de transfer learning permite aprovechar redes previamente entrenadas, lo que reduce la necesidad de grandes volúmenes de muestras originales (Nataraj et al., 2021).

Augmentación en secuencias de código (Android)

En el ámbito de Android, donde se emplea el análisis de secuencias de opcodes y llamadas a API, se han explorado técnicas de aumento basadas en modelos de lenguaje. El método Self-Embedding Language Model Augmentation utiliza embeddings de secuencias para generar nuevas variantes de código con propiedades similares, ampliando así la diversidad de entrenamiento. Estas técnicas han demostrado mejorar la capacidad de los modelos para detectar malware desconocido y polimórfico, lo que resulta crítico frente a las amenazas modernas en dispositivos móviles (McLaughlin & del Rincon, 2022).

Estas estrategias permiten reducir el sesgo en datasets, aumentar la robustez de los modelos predictivos y mejorar la detección de malware emergente, posicionándose como un componente esencial en pipelines modernos de ciberseguridad basados en ML.

Normalización contextual y selección de características

La calidad del preprocesamiento de datos es un factor determinante para el rendimiento de los modelos de ML en la detección de malware. Entre las técnicas más relevantes se encuentran la normalización contextual y la selección de características, cuyo propósito es garantizar que las variables de entrada representen patrones significativos de amenazas modernas sin introducir sesgos ni redundancias.

Escalamiento y normalización de características

La variabilidad en escalas numéricas entre atributos, por ejemplo, entropía de secciones, número de imports, longitudes de cadenas o métricas de flujo de red, puede distorsionar el rendimiento de algoritmos sensibles a magnitudes relativas. El uso de min–max normalization y Z-score standardization homogeniza las escalas, permitiendo que todas las variables aporten de manera equitativa al proceso de clasificación. En combinación con técnicas de reducción de

dimensionalidad como análisis de componentes principales (PCA), se logra eliminar redundancias y reducir el costo computacional. En un estudio reciente, la aplicación de estas técnicas permitió a modelos Light Gradient Boosting Machine (LGBM) alcanzar una precisión del 97.16 % en la detección de malware sobre un dataset tabular con 11 598 muestras y 139 características (Al-Hadhrami et al., 2022).

Selección de características relevantes

La reducción del espacio de atributos mediante métodos de selección como feature importance, recursive feature elimination o algoritmos basados en información mutua, mejora la interpretabilidad y evita el sobreajuste. Se ha demostrado que algoritmos como K-Nearest Neighbors (KNN), altamente sensibles a la escala y redundancia de variables, aumentaron su precisión de aproximadamente 56 % a más del 95 % tras la aplicación de selección de características y normalización adecuada (Wang et al., 2023)**.**

Modelos robustos a escala

Aunque algoritmos basados en ensambles, como Random Forest o XGBoost, presentan una relativa insensibilidad al escalado, la normalización sigue siendo relevante para acelerar la convergencia de modelos híbridos o profundos, y para garantizar consistencia en pipelines automatizados que integran distintos clasificadores. Esto es especialmente importante en contextos organizacionales donde los sistemas de detección deben procesar volúmenes masivos de datos heterogéneos provenientes de tráfico de red, binarios y telemetría en endpoints (Bensaoud et al., 2024)**.**

La normalización contextual y la selección de características no solo optimizan la precisión de los modelos, sino que también permiten construir clasificadores más robustos, escalables y explicables, elementos clave para la adopción en entornos organizacionales críticos.

Selección de características orientadas a amenazas modernas

La evolución del malware, marcada por la proliferación de ransomware, variantes polimórficas y técnicas avanzadas de ofuscación, ha impulsado la necesidad de diseñar estrategias de selección de características específicas para patrones de amenaza contemporáneos. Este proceso permite que los modelos de ML no solo alcancen altas tasas de detección, sino que también mantengan su robustez frente a intentos de evasión.

Características de llamadas API para ransomware

El análisis dinámico de secuencias de llamadas API vinculadas a accesos de archivos, rutinas de cifrado y modificaciones en el registro es una de las estrategias más efectivas para la detección temprana de ransomware. Estas secuencias permiten a los modelos predecir la activación de comportamientos maliciosos antes de que ocurra el cifrado masivo.(Anand et al., 2022). demostraron que clasificadores como Random Forest y XGBoost, entrenados con secuencias de API, alcanzan precisiones superiores al 85 %, validando su valor como features predictivas críticas.

Features estructurales en ejecutables Windows PE

En el ámbito estático, los ejecutables Windows PE ofrecen un conjunto rico de características: entropía de secciones, tablas de imports, hashes de secciones e histogramas de bytes. Estos atributos han mostrado ser altamente discriminativos y, además, permiten un preprocesamiento rápido y escalable en escenarios corporativos. Datasets públicos como EMBER facilitan la experimentación y entrenamiento de modelos, alcanzando rendimientos competitivos en benchmarks recientes (Bensaoud et al., 2024; Gibert et al., 2020)**.**

Representaciones gráficas mediante embeddings de CFG

Una de las líneas más prometedoras es el uso de Control Flow Graphs (CFGs), donde se representan relaciones de ejecución dentro del código. La extracción de subgrafos o embeddings de CFG permite capturar la semántica de ejecución y resistir cambios sintácticos superficiales generados por empaquetadores o técnicas de polimorfismo. (Qiang et al., 2022) reportan que redes neuronales profundas entrenadas en trazas derivadas de CFG alcanzaron precisiones del 95.7 %, mostrando gran robustez frente a variantes ofuscadas de malware Windows.

La selección de características orientadas a amenazas modernas combina la granularidad del comportamiento dinámico, la eficiencia de los atributos estáticos y la resiliencia de las representaciones gráficas. Esta integración constituye un marco técnico avanzado para el diseño de modelos predictivos que puedan adaptarse a la evolución del ecosistema de amenazas.

Propuesta de pipeline predictivo mejorado para la detección de malware

El diseño de modelos predictivos efectivos en ciberseguridad requiere un pipeline integral que no solo contemple la recolección y curación de datos, sino también técnicas avanzadas de augmentación, normalización contextual, selección de características y validación robusta. La siguiente propuesta integra las mejores prácticas identificadas en la literatura reciente, orientadas a la detección de amenazas modernas como ransomware, spyware, ataques DDoS y malware polimórfico, en la figura 11 se contempla la propuesta de un pipeline.

Recolección y curación de datos

El primer paso consiste en recolectar y etiquetar muestras diversas de malware y benignas provenientes de diferentes fuentes. Es fundamental incluir variantes recientes para garantizar la representatividad del dataset. En este contexto, conjuntos como EMBER, CICMalDroid, CTU-13

y OMD (Open Malware Dataset) han demostrado ser especialmente útiles para la evaluación comparativa (Guerra-Manzanares, 2024; Zhu et al., 2023).

Data augmentation diferencial

Dado que los datasets de malware suelen ser desbalanceados, se requiere aplicar técnicas específicas según el tipo de dato:

- Binarios - imágenes: conversión de ejecutables en mapas de bytes y aplicación de transformaciones como rotación, escalado o adición de ruido, que incrementan la robustez de modelos CNN (Pinhero et al., 2021).
- Secuencias de opcodes: uso de embedding-based augmentation que genera secuencias sintácticamente diferentes, pero semánticamente equivalentes, mejorando la capacidad de generalización (Razgallah et al., 2021).
- Binarios reales: técnicas como MARVOLO permiten crear variantes polimórficas equivalentes al malware original, ampliando el espacio de entrenamiento sin alterar etiquetas (Bensaoud et al., 2024).

Preprocesamiento contextual

El preprocesamiento garantiza que las características sean comparables y relevantes:

- Escalado: aplicar normalización min–max o Z-score en modelos sensibles al rango de atributos, como KNN o SVM, mejorando su precisión de forma significativa (Hanif et al., 2021).
- Selección de características: técnicas como PCA o selección basada en importancia de features (e.g., en LGBM) permiten reducir dimensionalidad, eliminar ruido y aumentar eficiencia en el entrenamiento (Kalouptsoglou et al., 2023).

Extracción de características orientadas a amenazas modernas

El pipeline incorpora características diseñadas específicamente para capturar comportamientos avanzados:

- Secuencias de llamadas API: asociadas a accesos de archivos, cifrado y persistencia, críticas para ransomware (Anand et al., 2022).
- Embeddings de CFG (Control Flow Graphs): capturan patrones de ejecución robustos frente a técnicas de ofuscación (Qiang et al., 2022).
- Cabeceras PE y atributos estructurales: entropía de secciones, imports y hashes, útiles para clasificación masiva en Windows (Gibert et al., 2020).
- Histogramas de bytes: representación compacta que facilita la detección en prefiltrados rápidos.

Entrenamiento del modelo

El entrenamiento debe garantizar un equilibrio entre precisión y robustez, seleccionando el modelo según el entorno de aplicación:

- Modelos tradicionales (RF, XGBoost): adecuados para entornos con recursos limitados, con interpretabilidad moderada y buena precisión en tabular datasets.
- Modelos profundos (CNN, RNN, GNN): recomendables para detección en tráfico de red, imágenes binarias y CFG embeddings, alcanzando métricas superiores al 95 % (Feng et al., 2024)

La evaluación debe realizarse con validación cruzada k-fold y métricas integrales como F1-score, AUC-ROC y recall, priorizando la reducción de falsos negativos.

Validación robusta

Para evitar sesgos, la validación debe considerar:

- Conjuntos de prueba sin augmentación, que permitan medir el desempeño real.
- Inclusión de muestras evasivas o adversariales, como las generadas con GANs, evaluando la resiliencia del modelo ante ataques de adversarial ML (Zhong et al., 2024).

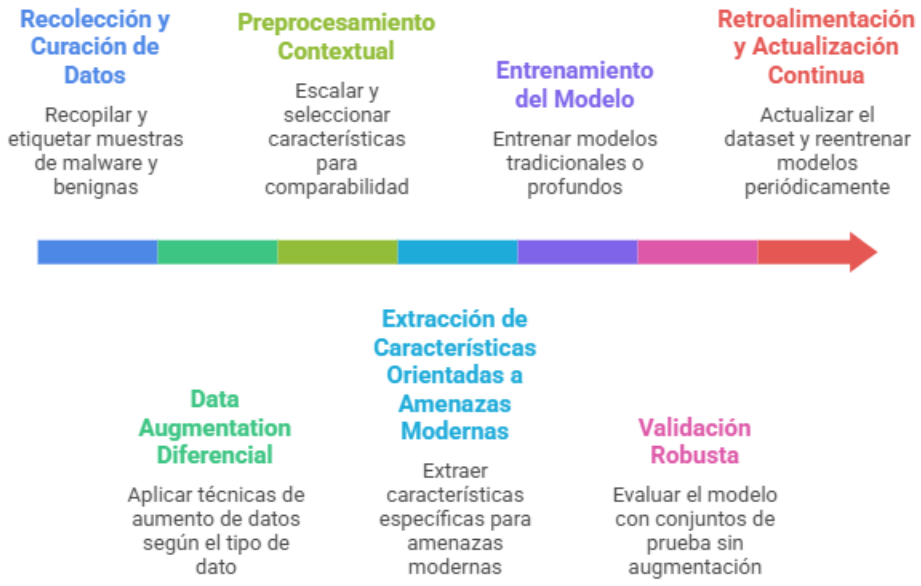
Retroalimentación y actualización continua

Finalmente, el pipeline debe ser dinámico y adaptable:

- Actualización periódica del dataset con variantes emergentes obtenidas de fuentes abiertas y telemetría corporativa.
- Reentrenamiento incremental de modelos, incorporando técnicas de aprendizaje continuo y adversarial training para mantener la vigencia del sistema (Elhanashi A.; Dini P., 2024).

Figura 11

Diagrama visual del pipeline.



Nota*:* Diagrama visual del desarrollo del pipeline. Fuente: elaboración propia.

La tabla 3 muestra un dataset simulado para la detección, prevención y mitigación de malware en entornos organizacionales, compuesto por diez registros que representan tanto muestras maliciosas como benignas. Cada registro corresponde a un archivo o aplicación analizada, con atributos técnicos que describen su comportamiento y estructura, junto con indicadores de evaluación de seguridad en el anexo C - dataset_malware se encuentra la data completa.

- ID: Identificador único de la muestra analizada.
- Tipo Malware: Clasificación del software (ej. Ransomware, Troyano, Spyware, DDoS, Phishing o Benigno).
- Tamaño Binario KB: Tamaño del archivo ejecutable en kilobytes, un rasgo comúnmente usado en análisis estático.
- Entropia Secciones: Medida de aleatoriedad en las secciones del binario; valores altos suelen estar asociados con ofuscación o cifrado.
- Num Imports: Cantidad de funciones o librerías importadas por el archivo, que puede indicar complejidad o intencionalidad maliciosa.
- Llamadas API Criticas: API relevantes detectadas, como EncryptFile (propio de ransomware) o RegSetValue (modificación del registro, típico en spyware).
- Permisos Android: Permisos solicitados en aplicaciones móviles, como acceso a SMS, ubicación o cámara, útiles para distinguir malware en entornos Android.
- Tráfico Red KBps: Volumen de tráfico de red generado durante la ejecución, que ayuda a identificar actividades como comunicación con servidores de comando y control (C2) o ataques DDoS.

- CFG Embeddings Score: Puntaje derivado del análisis de grafos de flujo de control (Control Flow Graphs), representando patrones estructurales en la lógica del programa.
- Ofuscación: Indicador binario (1 = sí, 0 = no) que señala si el malware emplea técnicas de ocultamiento de código.
- Tecnica AntiVM: Indica si el malware utiliza mecanismos para evadir detección en entornos virtualizados.
- Deteccion Temp: Métrica de efectividad de detección temprana (%), evaluando la capacidad de identificar la amenaza antes de que cause daño.
- Prevencion_Score: Puntaje de efectividad de medidas de prevención aplicadas.
- Mitigacion Score: Puntaje de efectividad en la contención y recuperación tras el ataque.
- Etiqueta: Valor binario que indica si la muestra es maliciosa (1) o benigna (0).

Tabla 3 *Dataset simulado*

ID	Tipo Malware	Tamaño Binario KB	Entropia Secciones	Num Imports	Llamadas API Criticas	Permisos Android	Tráfico Red KBps	CFG Embeddings Score	Ofuscacion	Tecnica AntiVM	Deteccion Temp	Prevencion Score	Mitigacion Score	Etiqueta
1	Ransomware	14822	6.23	355	EncryptFile	INTERNET	123.45	0.812	1	0	92	78	85	1
2	Benigno	4421	5.89	67	None	NONE	12.33	0.402	0	0	85	61	73	0
3	Troyano	25340	7.11	128	OpenProcess	READ_SMS	210.56	0.621	1	1	77	82	90	1
4	Spyware	18432	6.45	212	RegSetValue	ACCESS_FINE_LOCATION	145.90	0.553	0	1	81	74	79	1
5	Phishing	5520	4.78	34	SendRequest	INTERNET	67.21	0.317	1	0	70	59	63	1
6	DDoS	31200	7.80	415	OpenProcess	NONE	488.65	0.921	1	1	95	88	92	1
7	Benigno	7600	5.34	98	None	CAMERA	8.32	0.201	0	0	66	55	60	0
8	Ransomware	42015	6.89	389	EncryptFile	WRITE_EXTERNAL	354.44	0.785	1	1	93	90	94	1
9	Spyware	9800	6.12	120	RegSetValue	READ_SMS	76.21	0.477	0	0	72	68	70	1
10	Benigno	6300	4.90	45	None	NONE	5.19	0.154	0	0	60	50	55	0

Nota: Registros del dataset simulado. Fuente: elaboración propia

Las muestras benignas (IDs 2, 7 y 10) se caracterizan por menor entropía, tráfico de red reducido, ausencia de llamadas API críticas y menores puntajes en detección, prevención y mitigación. Los ransomware y troyanos muestran mayor entropía, uso de APIs críticas de cifrado o manipulación de procesos, así como técnicas de evasión avanzadas (ofuscación, anti-VM). Los ataques DDoS (ejemplo: ID 6) se distinguen por un tamaño binario grande y tráfico de red extremadamente alto. Los spyware y phishing presentan permisos sospechosos (ej. SMS, ubicación, internet), asociados a la exfiltración de datos y engaño a usuarios.

Esta tabla 3 constituye un dataset representativo que permite entrenar y validar modelos predictivos de ML orientados a la ciberseguridad, evaluando su capacidad de detección, prevención y mitigación del malware en escenarios organizacionales.

La figura 12 de barras horizontales se evidencia la proporción de muestras en el dataset según la tipología de malware, lo cual resulta fundamental para evaluar la representatividad y el equilibrio de los datos de entrenamiento en modelos de ML aplicados a la ciberseguridad. Este análisis no solo permite identificar qué categorías cuentan con mayor disponibilidad de información para el entrenamiento de los algoritmos, sino también reconocer aquellas con menor presencia, que podrían requerir estrategias adicionales como data augmentation o recolección complementaria de muestras. En este sentido, la distribución de datos se convierte en un insumo clave para garantizar la robustez, la capacidad de generalización y la reducción de sesgos en los modelos predictivos diseñados para la detección, prevención y mitigación de amenazas modernas.

Troyano. Con aproximadamente 42 muestras, constituye la categoría más representada. Su prevalencia refleja la versatilidad de los troyanos como vehículo de ataque, empleados comúnmente para establecer puertas traseras, robar credenciales o descargar cargas adicionales.

Esto subraya la importancia de que los modelos predictivos logren identificar sus múltiples variantes.

DDoS (Distributed Denial of Service). La segunda categoría más frecuente 37 muestras, lo que destaca la relevancia actual de los ataques de denegación de servicio contra organizaciones con infraestructura crítica y servicios en la nube. La inclusión de este tipo de muestras favorece la construcción de modelos capaces de detectar patrones de tráfico malicioso en entornos de red.

Phishing. Representa cerca de 35 muestras. Aunque el phishing tradicionalmente se manifiesta en correos electrónicos y enlaces, su presencia en datasets de malware obedece a la evolución hacia ataques multicanal (documentos adjuntos, macros maliciosas, aplicaciones móviles), lo que justifica su incorporación en un análisis integral.

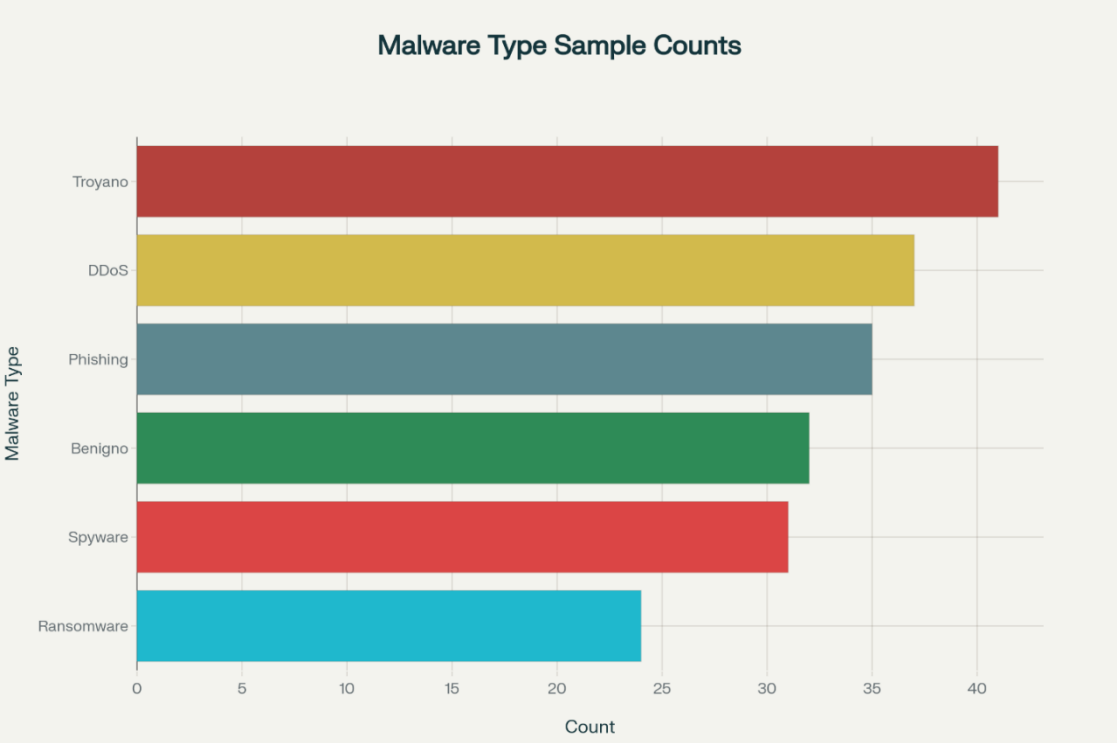
Benigno. Con alrededor de 32 muestras, constituye la clase negativa necesaria para entrenar modelos balanceados. Este grupo es crucial para medir la tasa de falsos positivos, asegurando que el sistema no clasifique erróneamente archivos o flujos legítimos como maliciosos.

Spyware. Agrupa unas 31 muestras, vinculadas a la exfiltración de información sensible como credenciales y datos financieros. Su representatividad en el dataset responde a su creciente uso en ataques dirigidos, donde la persistencia y la discreción son factores clave.

Ransomware. Es la categoría menos representada 24 muestras, a pesar de ser una de las más críticas por su alto impacto económico y operativo. La baja proporción puede asociarse a la escasez de datasets públicos actualizados o a la dificultad para recolectar muestras recientes. Este déficit obliga a considerar técnicas de data augmentation para equilibrar el dataset y mejorar la capacidad predictiva frente a este tipo de amenaza.

Figura 12

Tipo de malware.



Nota: Conteo por tipo de malware. Fuente: elaboración propia.

Diseño para la Construcción de un Sistema Predictivo de Detección de Malware Basado en Machine Learning

El desarrollo de modelos predictivos orientados a la detección, prevención y mitigación del malware en entornos organizacionales exige un diseño metodológico que garantice calidad, representatividad y equilibrio en los datos utilizados durante el entrenamiento. En el contexto actual, donde las amenazas evolucionan rápidamente y adoptan técnicas avanzadas de evasión, la solidez del dataset y la coherencia del proceso de modelado se convierten en factores determinantes para asegurar la eficacia de los sistemas basados en ML.

A continuación, Se describe un diseño integral y consistente que articula estrategias de curación de datos, augmentación diferencial, normalización contextual, selección y extracción de

características orientadas a amenazas modernas, así como la formulación de un pipeline predictivo robusto. Todo el proceso está orientado a garantizar que los modelos puedan generalizar adecuadamente, adaptarse a nuevas variantes y operar de manera confiable en escenarios reales.

Recolección y Curación de Datos

El primer componente del diseño consiste en la conformación de un dataset representativo y confiable. Para ello se integran muestras benignas y maliciosas provenientes de:

- Repositorios públicos (EMBER, CICIDS, CTU-13, OMD, Malimg),
- Telemetría organizacional (logs de EDR, capturas PCAP, trazas API), entornos controlados de sandboxing.

La curación incluye:

- Eliminación de duplicados,
- Validación de etiquetas,
- Verificación de integridad (hashes),
- Depuración de muestras obsoletas,
- Análisis de desbalance entre clases.

Esta fase permite asegurar que el conjunto inicial sea consistente y representativo del ecosistema actual de amenazas.

Estrategias de Representatividad y Equilibrio del Dataset

Debido al desbalance inherente en los datasets de malware, el diseño incorpora técnicas de data augmentation diferencial, adaptadas a cada tipo de dato, con el fin de aumentar la diversidad y mejorar la capacidad de generalización del modelo.

Augmentación semántica para binarios. Mediante herramientas como MARVOLO se generan variantes sintácticas de ejecutables que mantienen la funcionalidad, pero aumentan la

diversidad semántica del dataset, fortaleciendo la detección frente a técnicas como ofuscación y polimorfismo.

Augmentación visual. Los binarios se transforman en mapas de bytes representados como imágenes, permitiendo aplicar técnicas clásicas de visión por computador (rotación, ruido, escalado). Este tipo de augmentación mejora el rendimiento de modelos basados en CNN y redes híbridas.

Augmentación en secuencias de ejecución. Para opcodes y secuencias de llamadas API se aplican técnicas generativas basadas en embeddings, capaces de producir nuevas secuencias equivalentes. Este enfoque resulta especialmente relevante para la detección de malware móvil.

Estas estrategias permiten equilibrar la proporción entre clases y enriquecer la variabilidad interna del dataset sin comprometer la validez de las etiquetas.

Preprocesamiento Contextual

El preprocesamiento garantiza la coherencia interna del dataset y su compatibilidad con algoritmos sensibles a variaciones de escala o redundancia.

Normalización de atributos. Se aplican técnicas como min–max normalization y Z-score standardization para homogenizar las escalas numéricas. Esto evita que atributos con magnitudes mayores dominen el proceso de aprendizaje.

Selección y reducción de características. Se utilizan métodos como PCA, Recursive Feature Elimination y feature importance para eliminar redundancias, reducir dimensionalidad y mejorar interpretabilidad. Esta etapa optimiza tanto el rendimiento computacional como la estabilidad de los modelos.

Ingeniería de Características Orientada a Amenazas Modernas

Para garantizar que el modelo capture patrones relevantes y actualizados, la selección de características se orienta a comportamientos reales y contemporáneos asociados a diferentes tipos de malware.

Ransomware. Se consideran atributos relacionados con secuencias de API de cifrado, escritura masiva de archivos, persistencia y entropía de secciones PE.

Troyanos y spyware. Se incluyen características asociadas a manipulación de procesos, acceso a credenciales, lectura de información sensible y permisos críticos en Android.

Malware DDoS. Se incorporan métricas de tráfico, tasas de paquetes, conexiones simultáneas y patrones de solicitudes.

Representaciones basadas en Control Flow Graphs (CFG). Los embeddings de CFG aportan robustez ante alteraciones sintácticas y técnicas de evasión, y permiten capturar la lógica interna del ejecutable.

Esta combinación garantiza la construcción de modelos capaces de reconocer tanto patrones tradicionales como comportamientos avanzados.

Pipeline Predictivo Mejorado

El diseño se estructura en un pipeline modular que integra las mejores prácticas metodológicas identificadas en la literatura especializada:

Recolección y curación de datos. Dataset limpio, representativo y actualizado.

Data augmentation diferencial. Aplicado según el tipo de dato para garantizar equilibrio y diversidad.

Preprocesamiento y normalización contextual. Homogeneización de atributos, reducción de ruido y selección de variables relevantes.

Extracción de características orientadas a amenazas modernas. Combinación de análisis estático, dinámico y gráfico.

Entrenamiento del modelo. Selección del algoritmo según el entorno organizacional:

- Modelos tradicionales (RF, XGBoost) para infraestructura limitada,
- Modelos profundos (CNN, RNN, GNN) para análisis visual, secuencial o gráfico,
- Arquitecturas híbridas para entornos complejos.

La evaluación se realiza mediante F1-score, AUC-ROC y recall, priorizando la sensibilidad para reducir falsos negativos.

Validación robusta. Incluye pruebas sin augmentación, validación temporal, evaluación con muestras evasivas y muestras adversariales generadas mediante GANs.

Actualización y retroalimentación continua. Se emplean técnicas de drift-aware retraining, aprendizaje incremental y actualización frecuente del dataset a partir de telemetría real.

Aplicabilidad Organizacional

El diseño propuesto permite una integración coherente con infraestructuras de seguridad existentes, tales como SIEM, EDR e IDS/IPS.

Asimismo, incorpora técnicas de explicabilidad XAI, como SHAP, LIME y CAM, que permiten a los analistas comprender las decisiones del modelo y apoyar la toma de decisiones en incidentes reales.

Este enfoque no solo fortalece la detección temprana y la mitigación, sino que también promueve la transparencia, la trazabilidad y la alineación con políticas internas y marcos normativos vigentes.

El diseño consistente presentado constituye una propuesta metodológica integral para la construcción de sistemas predictivos de detección de malware basados en ML. Su estructura

garantiza calidad y equilibrio en los datos, incorpora técnicas avanzadas de ingeniería de características y establece un pipeline robusto, escalable y adaptable a la evolución de las amenazas digitales.

La combinación de estas prácticas asegura que los modelos sean precisos, resilientes y operativamente viables en entornos organizacionales, contribuyendo de manera significativa al fortalecimiento de la ciberseguridad empresarial. Como se detalla en la figura 13 la construcción de un diseño predictivo.

Figura 13

Construcción de un Sistema Predictivo de Detección de Malware.

Nota: Modelo para la construcción de un modelo. Fuente: elaboración propia.

Los resultados evidencian que la calidad del dataset constituye un factor determinante en el desempeño de los modelos predictivos. La integración de técnicas de data augmentation

semántico, visual y secuencial permitió reducir el desbalance entre clases, ampliar la diversidad de muestras y fortalecer la capacidad de generalización frente a variantes polimórficas y evasivas, particularmente en tipologías críticas como ransomware y troyanos avanzados. Este enfoque se materializó en la construcción de un dataset representativo y documentado, que sirve como insumo base para el entrenamiento y validación de los modelos propuestos.

Asimismo, la aplicación de procesos de normalización contextual y selección de características permitió optimizar el rendimiento de los algoritmos de ML, reduciendo redundancias, minimizando sesgos y mejorando la interpretabilidad de los resultados. La identificación de atributos relevantes como entropía de secciones, secuencias de llamadas API, tráfico de red y embeddings de grafos de flujo de control, demostró ser fundamental para capturar patrones de comportamiento asociados a amenazas contemporáneas, superando las limitaciones de enfoques tradicionales basados únicamente en firmas.

Como resultado concreto, se diseñó un pipeline predictivo mejorado el cual se visualiza en el anexo D, estructurado de forma modular y escalable, que integra desde la recolección y curación de datos hasta la validación robusta y la retroalimentación continua. Este pipeline no solo articula los elementos técnicos desarrollados a lo largo del capítulo, sino que establece una base sólida para su implementación en entornos organizacionales, donde la detección temprana, la reducción de falsos positivos y la adaptabilidad frente a nuevas amenazas son requisitos críticos.

En conjunto, los productos obtenidos son los dataset balanceado, estrategias de augmentación diferenciadas, selección de características orientadas a amenazas modernas y diseño del pipeline predictivo, confirman que el objetivo planteado fue alcanzado de manera efectiva. Más allá de la mejora en métricas de desempeño, el aporte principal de este capítulo radica en ofrecer un marco práctico y replicable para el diseño de sistemas de detección de malware basados

en ML, alineados con las necesidades reales de la ciberseguridad organizacional y preparados para enfrentar la evolución constante del ecosistema de amenazas. En este sentido, el diseño propuesto no solo cumple con el objetivo de desarrollar un sistema predictivo eficaz, sino que establece un marco metodológico replicable, escalable y alineado con las necesidades actuales de la ciberseguridad, permitiendo a las organizaciones fortalecer sus capacidades de detección, prevención y mitigación frente a amenazas digitales en constante evolución.

Validación adaptativa de modelos de detección de malware

La evolución constante del malware, caracterizada por el uso de polimorfismo, técnicas anti-VM y mecanismos de evasión dinámica, plantea desafíos significativos a los sistemas de detección tradicionales. Estos cambios provocan lo que se conoce como concept drift, es decir, la pérdida progresiva de eficacia de los modelos entrenados sobre datasets estáticos al enfrentarse con nuevas variantes cuyo comportamiento difiere del observado durante el entrenamiento inicial. Estudios recientes muestran que estrategias de drift-aware retraining logran mejorar significativamente la precisión en comparación con modelos entrenados de manera estática (Zhu et al., 2023).

En este contexto, validar un modelo únicamente mediante métricas estáticas en un conjunto de prueba fijo resulta insuficiente. Es necesario diseñar procesos de validación que midan no solo el rendimiento puntual, sino también la resiliencia de los detectores frente a variantes emergentes y ataques adversariales. La literatura reciente identifica tres problemáticas críticas que deben ser atendidas:

- Manejo activo del concept drift en entornos reales de ciberseguridad. Estrategias como MORPH, que combina pseudo-etiquetado y aprendizaje activo, han demostrado ser efectivas para adaptarse a nuevas variantes de malware mientras reducen la necesidad de etiquetado manual (Afianian et al., 2018).
- Retrasos en el etiquetado de muestras (label delay). Este problema dificulta la incorporación temprana de nuevos ataques en los datasets de entrenamiento, generando periodos de vulnerabilidad en los sistemas de detección.
- Necesidad de mecanismos de explicabilidad XAI. Herramientas como SHAP, LIME o CAM permiten que los analistas humanos comprendan las decisiones de los algoritmos y

prioricen acciones de respuesta. En contextos críticos como IDS (Intrusion Detection Systems) y SOC (Security Operation Centers), la explicabilidad es clave para reducir la desconfianza en los modelos y mejorar la capacidad de respuesta (Mohale & Obagbuwa, 2025; Nazim Sadia AND Alam, 2025).

En síntesis, la dinámica cambiante del malware y las limitaciones inherentes a los enfoques estáticos obligan a repensar los procesos de validación de modelos de detección. El concept drift, los retrasos en el etiquetado y la necesidad de explicabilidad exigen un marco metodológico que combine validación continua, adaptación dinámica y mecanismos de interpretación. Bajo esta perspectiva, un enfoque de validación adaptativa no solo permite mantener la eficacia técnica frente a variantes emergentes, sino que también facilita la integración operativa en entornos organizacionales, donde la confianza del analista y la capacidad de respuesta inmediata resultan determinantes para la mitigación de ciberataques modernos.

Marco conceptual para la validación adaptativa

Para superar los retos asociados al concept drift, los retrasos en etiquetado y la falta de interpretabilidad en los modelos de detección de malware, se propone un marco de validación adaptativa, sustentado en los siguientes principios:

Temporalidad en la partición de datos

El uso de particiones temporales, donde los modelos se entrenan con datos históricos y se evalúan con muestras recientes, permite simular la deriva de concepto y medir la degradación real del rendimiento predictivo en entornos cambiantes. Este enfoque temporal ofrece mayor realismo frente a particiones aleatorias, ya que refleja el comportamiento de los sistemas en escenarios de evolución continua del malware (Díaz et al., 2021; Gibert et al., 2020).

Evaluación continua (monitoring)

La validación no debe limitarse a un momento puntual, sino incorporar mecanismos de monitoreo permanente de métricas clave como precisión, recall y F1-score, estableciendo umbrales de rendimiento que activen procesos automáticos de reentrenamiento o actualización incremental del modelo cuando se detecta una degradación significativa. Esto permite mantener la eficacia en entornos donde el malware evoluciona con alta frecuencia (Halbouni et al., 2022).

Pruebas adversariales

Es indispensable someter los modelos a muestras adversariales generadas mediante técnicas de ML, como el marco MalFox, un Conv-GAN que ha demostrado reducir la efectividad de los detectores de malware tipo caja negra hasta en un 57%. Estas pruebas permiten medir la robustez de los modelos frente a intentos deliberados de evasión y guiar la integración de estrategias de adversarial training en los ciclos de mejora continua (Zhong et al., 2024).

Explicabilidad operacional

La incorporación de técnicas de XAI como SHAP, LIME y CAM resulta crítica para que los analistas humanos comprendan el razonamiento de los algoritmos. Estas herramientas permiten identificar las características más influyentes en cada predicción, detectar falsos positivos o negativos con mayor precisión y generar confianza en el uso de los modelos en entornos críticos como SOC e IDS. Estudios recientes han evidenciado que la explicabilidad no solo mejora la auditoría de decisiones algorítmicas, sino que también facilita la priorización de acciones de respuesta en escenarios de alta presión operativa(Lundberg & Lee, 2017; Marcinkevičs & Vogt, 2023).

Protocolo experimental propuesto

La validación de modelos predictivos en el ámbito de la ciberseguridad requiere mucho más que un conjunto de algoritmos sofisticados; demanda una base sólida de datos confiables, representativos y trazables. En este sentido, la preparación y recolección de la información constituyen el pilar sobre el cual se construye la precisión, la robustez y la aplicabilidad real de los modelos de ML orientados a la detección de malware.

Preparación y recolección

El primer paso en la validación de modelos predictivos para detección de malware consiste en la construcción de un corpus representativo y trazable, que garantice tanto la diversidad de amenazas como la calidad del proceso de entrenamiento. Este corpus debe integrar fuentes públicas reconocidas y telemetría interna de la organización, con el fin de obtener un equilibrio entre reproducibilidad científica y contextualización organizacional. En la figura 14 se visualiza el proceso para la preparación y recolección de datos.

Fuentes de datos. Para asegurar comparabilidad con investigaciones previas, se recomienda incorporar datasets de uso común en la literatura: EMBER, especializado en ejecutables PE; CTU-13, centrado en tráfico de botnets; los conjuntos CIC, orientados a tráfico y ataques en red; y el Open Malware Dataset (OMD). Estos recursos han demostrado utilidad en estudios de benchmarking y permiten establecer líneas base de rendimiento(García et al., 2014; Gibert et al., 2020).

Complementariamente, la inclusión de telemetría propia de un SOC (logs EDR de procesos y llamadas API, capturas PCAP/NetFlow, trazas dinámicas de sandboxes y muestras PE/APK recolectadas en incidentes) permite ajustar los modelos al entorno específico de la organización, aumentando su relevancia práctica (Elhanashi A.; Dini P., 2024).

Curado y metadatos. El preprocesamiento debe contemplar:

- Limpieza, eliminando duplicados, unificando formatos y corrigiendo valores faltantes.
- Normalización contextual, aplicando escalado adecuado según la naturaleza de la variable (p. ej., min–max o Z-score para métricas continuas, tokenización o embeddings para secuencias de opcodes o llamadas API), lo que mejora la consistencia y evita sesgos en algoritmos sensibles al rango (Al-Hadhrami et al., 2022).
- Etiquetado y trazabilidad, asignando a cada muestra atributos como tipo de amenaza (maligno/benigno/familia), fecha de captura, origen, método de recolección, nivel de confianza de la etiqueta, hash criptográfico y versión del pipeline de análisis. Esta práctica fortalece la auditabilidad y la reproducibilidad.

Reserva temporal. Para simular el concept drift y evaluar la resiliencia de los modelos frente a la evolución de amenazas, se propone una división temporal del dataset. Un ejemplo de partición incluye:

- Entrenamiento (training): meses 1–6.
- Validación (validation): meses 7–8.
- Prueba / simulación de producción (test): meses 9–12.

Este enfoque temporal, en contraste con la simple partición aleatoria, permite observar cómo la precisión y recall de los modelos se degradan al enfrentarse a variantes emergentes, reproduciendo condiciones más cercanas a escenarios reales (Gibert et al., 2020).

Figura 14

Preparación y recolección de datos de modelos predictivo.

Nota: Flujo de preparación y recolección de datos. Fuente: elaboración propia.

Entrenamiento inicial

Arquitectura de referencia. El diseño de la arquitectura de referencia para la detección de malware se fundamenta en un modelo híbrido tipo ensemble, es decir, una combinación de clasificadores especializados que procesan diferentes tipos de representación de datos. Esta integración busca aprovechar las fortalezas individuales de cada enfoque, generando un sistema más preciso, robusto y generalizable.

A continuación, se describen los tres componentes principales del ensemble, cada uno enfocado en un tipo específico de información: tabular, visual y secuencial.

Tabular / Features estáticas. En este caso, el modelo se centra en atributos extraídos directamente de los binarios, sin necesidad de ejecutar el archivo. Se emplean algoritmos basados en árboles de decisión, como Random Forest o Light Gradient Boosting Machine (LightGBM), los cuales han demostrado un excelente desempeño en la clasificación de archivos maliciosos.

Entre las características más utilizadas se incluyen la entropía de secciones, el número de imports, los histogramas de bytes y los hashes estructurales. Estas variables permiten representar de forma cuantitativa la estructura interna del ejecutable. Estudios como el de (Al-Hadhrami et al., 2022) evidencian que, al aplicar estrategias de normalización y selección de características, estos modelos alcanzan niveles de precisión elevados y una eficiente capacidad de generalización.

Visual / Binarios como imágenes. En este segundo enfoque, los archivos ejecutables se transforman en representaciones visuales, conocidas como mapas de bytes o imágenes en escala de grises. Esta técnica permite analizar los binarios mediante redes neuronales convolucionales (CNN), aprovechando su capacidad para detectar patrones espaciales.

El uso de aprendizaje por transferencia (transfer learning), utilizando modelos preentrenados, reduce significativamente los costos computacionales y mejora la eficacia en datasets con clases desbalanceadas. Según (Pinhero et al., 2021), esta metodología logró precisiones superiores al 98 % en la detección de malware, consolidándose como una alternativa eficaz para escenarios de análisis masivo.

Secuencias / Comportamiento dinámico. El tercer componente del ensemble aborda la detección de malware desde una perspectiva comportamental, analizando secuencias de ejecución como llamadas a API, opcodes o trazas de control. Para ello, se emplean arquitecturas secuenciales tales como Redes Neuronales Recurrentes (RNN), Long Short-Term Memory (LSTM), Transformers o CNN unidimensionales (1D-CNN).

Investigaciones de (Anand et al., 2022) demostraron que los clasificadores basados en secuencias de llamadas API pueden identificar ransomware en fases iniciales de ejecución, permitiendo una respuesta temprana. De igual modo, (Qiang et al., 2022) comprobaron que las

redes neuronales profundas aplicadas a trazas de flujo de control ofrecen una alta robustez frente a variantes polimórficas, un desafío crítico para los métodos de detección tradicionales.

Checkpointing e indicadores. Una vez completado el entrenamiento inicial del ensemble, es indispensable establecer un modelo base o baseline checkpoint. Este punto de control debe incluir los pesos entrenados, la versión del dataset utilizado y la configuración de parámetros, permitiendo evaluar futuras iteraciones del modelo y detectar degradaciones de rendimiento ocasionadas por el fenómeno de concept drift.

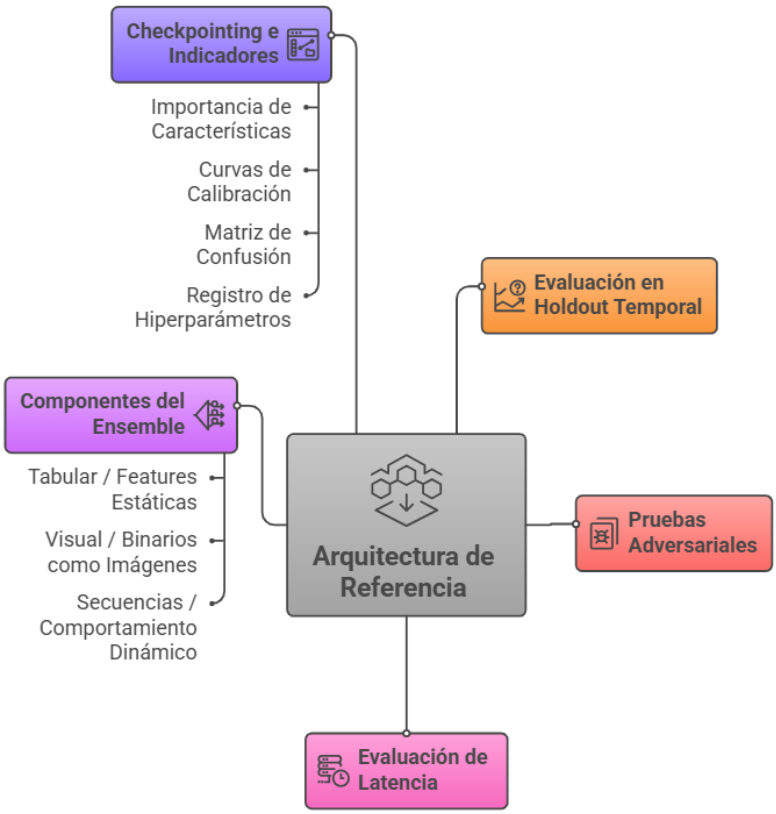
Asimismo, se deben registrar una serie de indicadores clave que garanticen la trazabilidad, auditabilidad y reproducibilidad de los experimentos:

- Importancia de características (feature importances): permite comprender la contribución de cada atributo en la decisión final del modelo.
- Curvas de calibración: muestran la coherencia entre las probabilidades predichas y las etiquetas reales, ayudando a evaluar la fiabilidad del modelo.
- Matriz de confusión por familia de malware: posibilita medir el desempeño diferencial frente a cada tipo o familia de amenaza.
- Registro de hiperparámetros y semillas aleatorias: su documentación asegura la replicabilidad de los resultados y facilita experimentos comparativos en entornos controlados.

Estos mecanismos no solo fortalecen la validez científica del modelo, sino que también permiten mantener un control riguroso sobre su evolución en el tiempo, garantizando resultados consistentes y comparables, tal como se ilustra en la figura 15.

Figura 15

Entrenamiento Inicial para la validación adaptativa.



Nota: Entrenamiento Inicial partir de una arquitectura híbrida. Fuente: elaboración propia.

Validación offline robusta

La validación offline constituye una fase esencial dentro del ciclo de vida de los modelos de detección de malware. Su propósito es garantizar que el desempeño de los algoritmos no solo sea satisfactorio en condiciones de laboratorio, sino que también se mantenga estable frente a la evolución de las amenazas (concept drift) y ante intentos de evasión mediante ataques adversariales.

Este proceso permite evaluar la resiliencia, estabilidad y confiabilidad del modelo antes de su despliegue operativo, asegurando que la solución propuesta sea viable en entornos reales de ciberseguridad

Métricas y criterios de evaluación. La correcta elección de métricas es determinante para evaluar la calidad de un modelo de detección. Siguiendo las recomendaciones de(Al-Hadhrami et al., 2022) y (Gibert et al., 2020), se proponen los siguientes indicadores clave:

- Métricas principales
    - F1-score por clase: combina precisión y recall, siendo especialmente útil en conjuntos de datos desbalanceados donde ciertas familias de malware son menos frecuentes.
    - AUC-ROC: mide la capacidad general del modelo para distinguir entre clases positivas y negativas; se recomienda complementarlo con Precision-Recall AUC cuando la proporción de muestras maliciosas es baja.
    - Precision@k: resulta relevante en entornos operativos (por ejemplo, en un SOC), donde las primeras k alertas detectadas son las más prioritarias.
    - Recall (sensibilidad): en ciberseguridad, minimizar los falsos negativos es fundamental, por lo que se recomienda reportar el recall desagregado por familia crítica.
    - False Positive Rate (FPR): permite cuantificar el costo operativo derivado de falsas alarmas.
    - Time-to-Detect (TTD): mide el tiempo transcurrido desde la ejecución del comportamiento malicioso hasta la generación de la alerta, evaluando la velocidad de respuesta del modelo.
    - Métricas de calibración (Brier score, curvas de calibración): determinan si las probabilidades predichas pueden interpretarse como niveles de confianza operativa.

- Criterios operativos y umbrales

Para considerar que un modelo está listo para su despliegue en producción, se establecen umbrales mínimos:

- F1-score macro ≥ 0.85
- FPR ≤ 1 % en el conjunto de validación temporal.
- TTD ≤ SLA definido (según política organizacional).

Si el desempeño presenta una caída en F1 superior a 0.03 respecto al modelo base (baseline), se debe activar un proceso de reentrenamiento o investigación de causa, con el fin de mantener la calidad del sistema.

Evaluación temporal y medición de drift. Una de las principales amenazas a la estabilidad de los modelos de detección es el concept drift, es decir, la pérdida de efectividad frente a nuevas variantes de malware que alteran su comportamiento o estructura. Siguiendo las recomendaciones de (Gibert et al., 2020) y (Halbouni et al., 2022), se proponen las siguientes estrategias:

- Holdout temporal
  - Los datos deben dividirse por ventanas de tiempo, por ejemplo:
    - Entrenamiento: meses 1–6
    - Validación: meses 7–8
    - Prueba o simulación: meses 9–12
  - Este enfoque temporal permite observar la degradación del modelo a lo largo del tiempo y medir su capacidad de adaptación ante nuevas amenazas.

- Se recomienda analizar tendencias de degradación mediante series temporales de métricas (F1, AUC, FPR), en lugar de depender únicamente de resultados puntuales.

➢ Pruebas de significancia

Antes de implementar actualizaciones automáticas o reentrenamientos, deben realizarse pruebas estadísticas que determinen si la diferencia de rendimiento entre versiones es significativa. Se sugieren:

- Prueba de McNemar, para comparaciones pareadas.
- Intervalos de confianza con bootstrap sobre métricas como AUC y F1.

Estas pruebas aseguran que los cambios en el modelo se basen en evidencia estadística y no en fluctuaciones aleatorias.

Pruebas adversariales. En un escenario donde los atacantes utilizan técnicas de evasión basadas en ML, la evaluación adversarial es fundamental para medir la verdadera robustez del modelo. Investigaciones recientes como las de (Ling et al., 2023) y (Zhong et al., 2024) proponen protocolos sistemáticos para simular ataques realistas y evaluar la capacidad de defensa de los clasificadores.

➢ Simulación de evasión

Se generan muestras modificadas mediante técnicas adversariales, tales como GANs o Conv-GANs (por ejemplo, MalFox), así como manipulaciones realistas de binarios (empaquetado, inserción de bytes, ofuscación).

También se incluyen ataques en el espacio de características (feature-space attacks) y transformaciones semánticas, como las utilizadas en el framework MARVOLO, con el fin de medir el impacto sobre las métricas de rendimiento.

- Protocolo de pruebas
    - Seleccionar un subconjunto representativo del conjunto de prueba.
    - Generar variantes adversariales controladas (por muestra y tipo de manipulación).
    - Evaluar la caída en las métricas principales (ΔF1, ΔRecall, ΔAUC).
    - Identificar grupos de características más vulnerables (por ejemplo, si el packing afecta más la detección estática que la dinámica).
    - Aplicar entrenamiento adversarial, reentrenando el modelo con un porcentaje de estas muestras (entre 10 % y 30 %) y reevaluar su desempeño.
- Medidas de defensa

Para fortalecer la resiliencia del modelo ante ataques adversariales, se recomienda:

- Implementar entrenamiento adversarial iterativo, incorporando ejemplos generados por distintos métodos.
- Utilizar ensembles heterogéneos, combinando análisis estático y dinámico para reducir vulnerabilidades asociadas a un solo tipo de característica.
- Evaluar la transferencia adversarial, es decir, comprobar si los ataques diseñados contra un modelo (A) también afectan a otro (B), lo que permite estimar la generalización del riesgo.

Procedimiento práctico de validación offline. La validación offline requiere una metodología ordenada y trazable que permita evaluar de forma integral el desempeño, la estabilidad y la robustez del modelo antes de su despliegue en entornos operativos. Este procedimiento debe ejecutarse de manera sistemática para garantizar que los resultados obtenidos

sean confiables, reproducibles y comparables a lo largo del tiempo. A continuación, se describen las principales etapas del proceso:

- Preparación de los conjuntos de datos

El primer paso consiste en crear los conjuntos de entrenamiento, validación y prueba (baseline train/val/test), siguiendo un criterio temporal que refleje la evolución natural de las amenazas. Es fundamental mantener constante la semilla aleatoria (seed) y documentar las versiones de los datasets utilizados, de modo que el proceso pueda repetirse y auditarse fácilmente.

- Calibración de modelos

Posteriormente, se deben calibrar las probabilidades de salida de los modelos mediante técnicas como Platt Scaling o Isotonic Regression, con el fin de asegurar que las predicciones reflejen niveles de confianza realistas. Durante esta fase se generan y conservan las curvas de calibración, que servirán para interpretar la fiabilidad del modelo.

- Evaluación inicial

En esta etapa, se calculan las métricas de desempeño principales (precisión, recall, F1-score), además de curvas ROC y PR, matriz de confusión y estimación del Time-To-Detect (TTD). Esta evaluación proporciona una línea base (baseline performance) frente a la cual se medirán los cambios posteriores.

- Validación temporal (holdout temporal)

El modelo se evalúa sobre ventanas de tiempo futuras, con el propósito de detectar degradaciones progresivas en su rendimiento debido a la evolución del malware. Se registran y analizan las variaciones de las métricas para identificar la magnitud del concept drift.

- Generación de muestras adversariales

Se aplican técnicas de generación adversarial utilizando herramientas como MalFox o redes GANs, junto con estrategias realistas de packing u obfuscation. Estas simulaciones permiten medir la caída en el desempeño del modelo ante intentos de evasión.

- Mitigación adversarial

Tras identificar vulnerabilidades, se implementan estrategias de adversarial training o augmentación dirigida, incorporando ejemplos modificados al conjunto de entrenamiento. El modelo reentrenado se compara con la versión base mediante pruebas de significancia estadística, evaluando si las mejoras son efectivas.

- Elaboración del informe de robustez

Finalmente, se elabora un informe de robustez que documenta:

- Las vulnerabilidades detectadas (qué tipos de evasión afectan más).
- La sensibilidad por grupo de características (features más vulnerables).
- Las recomendaciones de mitigación o ajustes arquitectónicos.

Este informe constituye la evidencia técnica que respalda la solidez del modelo antes de su paso a producción.

Reportes y artefactos para conservar. Una práctica clave en la validación de modelos predictivos es la conservación de todos los artefactos experimentales, lo cual garantiza trazabilidad, transparencia y reproducibilidad científica. De acuerdo con (Elhanashi A.; Dini P., 2024), estos elementos son fundamentales para auditorías y futuras optimizaciones.

- Checkpoints del modelo

Se deben almacenar los checkpoints del modelo base junto con las versiones posteriores al reentrenamiento. Esto permite restaurar configuraciones previas y analizar la evolución del rendimiento.

- Conjuntos de prueba

Deben conservarse tanto los datasets originales como las versiones adversariales generadas, acompañados de sus respectivos metadatos y semillas aleatorias, lo que facilita su reutilización en evaluaciones futuras.

- Registros de experimentos

Es esencial mantener logs detallados que incluyan:

- Hiperparámetros utilizados.
- Tiempos de entrenamiento.
- Recursos computacionales consumidos.
- Curvas de aprendizaje y validación.

Estos registros permiten replicar los resultados y detectar desviaciones en fases posteriores.

- Informe de robustez

El informe final debe incluir:

- Gráficas de degradación temporal de las métricas.
- Impacto visual de ejemplos adversariales.
- Matriz de confusión por familia antes y después del reentrenamiento adversarial.

Este documento consolida la evidencia empírica y justifica los cambios aplicados al modelo durante su ciclo de vida.

Herramientas y recursos recomendados. La implementación efectiva de un proceso de validación offline robusto requiere el uso de herramientas y frameworks especializados que faciliten la experimentación, la calibración y la simulación de escenarios adversariales. Entre los más destacados se encuentran:

- Generación adversarial

Para la creación de muestras modificadas, se recomiendan implementaciones de GANs adaptadas a binarios, basadas en marcos de investigación recientes como MalFox. Estas herramientas permiten evaluar la resistencia de los modelos frente a malware evasivo o polimórfico.

- Evaluación temporal y procesamiento en flujo

Librerías como River y Scikit-Multiflow facilitan la evaluación incremental de modelos bajo un esquema de streaming, lo que resulta útil para prototipos de detección en tiempo real y para medir la degradación continua del desempeño.

- Calibración y explicabilidad

Para interpretar las decisiones del modelo, se recomienda el uso de herramientas de XAI como SHAP y LIME, junto con los módulos de calibración de Scikit-learn, que permiten analizar la confianza y coherencia de las predicciones.

- Sandboxes y entornos de detonación

El uso de plataformas como Cuckoo Sandbox, EDR instrumentados o incluso entornos bare-metal permite ejecutar muestras en condiciones controladas, especialmente útiles para el análisis de malware que emplea técnicas anti-VM o anti-debugging.

La validación offline robusta no debe considerarse un único experimento, sino un proceso iterativo y cíclico que combina medición, simulación de ataques, mitigación y reevaluación. Este

enfoque continuo (documentado y automatizado) es la base para construir modelos de detección de malware verdaderamente adaptativos, capaces de mantener su eficacia frente a la evolución constante del ecosistema de amenazas digitales (Gibert et al., 2020; Ling et al., 2023; Zhong et al., 2024).

Validación en laboratorio

La validación en laboratorio representa una fase crucial dentro del ciclo de desarrollo de modelos predictivos para la detección de malware. Constituye el punto intermedio entre la evaluación offline basada en experimentos controlados y datasets estáticos y el despliegue en entornos productivos.

Su principal objetivo es comprobar el comportamiento del modelo en condiciones realistas pero seguras, permitiendo recrear amenazas reales sin poner en riesgo la infraestructura corporativa. Esta etapa garantiza que los modelos no solo sean precisos en pruebas teóricas, sino también operativamente confiables frente a la dinámica de los ataques en escenarios reales.

Sandboxes y detonación controlada. En esta fase, los experimentos se desarrollan dentro de entornos aislados o "sandboxes", diseñados para observar el comportamiento del malware de forma segura. Herramientas como Cuckoo Sandbox o las soluciones integradas en plataformas EDR (Endpoint Detection and Response) permiten ejecutar muestras maliciosas y registrar de manera detallada sus actividades.

Durante las pruebas, se recolectan evidencias clave como:

- Secuencias de llamadas API, que revelan el flujo de ejecución del código malicioso.
- Interacciones con el sistema de archivos, incluyendo creación, modificación o cifrado de documentos.

- Cambios en el registro de Windows, asociados a persistencia o manipulación de configuraciones.
- Tráfico de red, que puede indicar comunicación con servidores de comando y control (C2).

Sin embargo, se debe considerar que muchos malware modernos incorporan técnicas anti-VM y anti-instrumentación, diseñadas para detectar y evadir entornos virtuales. Por ello, es recomendable complementar estas pruebas con detonaciones en entornos bare-metal, es decir, sobre hardware físico sin virtualización.

Estos escenarios, reforzados con mitigaciones anti-evasión, permiten observar comportamientos ocultos que los sandboxes tradicionales no logran detectar. Investigaciones como las de (Afianian et al., 2018) y (Kirat et al., 2014) han demostrado que los análisis en bare-metal revelan patrones maliciosos invisibles en entornos virtualizados, reforzando así la efectividad de la validación antes del despliegue final.

Escenarios simulados y campañas controladas. Además del análisis de muestras individuales, la validación en laboratorio debe incluir escenarios simulados que representen campañas completas de ataque. Estas simulaciones permiten observar la respuesta integral del modelo frente a secuencias complejas de eventos, similares a los que ocurren en la realidad.

Algunos ejemplos de escenarios que pueden recrearse incluyen:

- Caso 1: Simulación de ransomware.

Se ejecutan fases progresivas de cifrado en sistemas Windows con el fin de evaluar la capacidad del modelo para detectar la amenaza en etapas tempranas, a partir del análisis de secuencias de llamadas API y comportamientos anómalos previos al cifrado total de los archivos.

- Caso 2: Campañas de spear-phishing.

Se recrea un flujo de ataque en el que un correo electrónico malicioso conduce a la descarga y ejecución de un payload. En este escenario, se evalúa la eficacia del modelo en múltiples puntos de la cadena de ataque, desde la detección del adjunto hasta la ejecución final del malware.

- Caso 3: Botnets y ataques DDoS.

Se genera tráfico masivo y coordinado, simulando la actividad de una red de dispositivos comprometidos. El objetivo es medir la sensibilidad del modelo para identificar anomalías en la red y su capacidad de respuesta ante picos de actividad inusual.

En cada uno de estos escenarios, la validación no se limita a medir la tasa de detección. También se evalúa un parámetro esencial en entornos operativos: la latencia de respuesta, expresada como el tiempo medio hasta la detección (TTD). Comparar el TTD entre distintos modelos de ML o DL permite determinar cuál ofrece una respuesta más ágil y efectiva en contextos críticos como los Centros de Operaciones de Seguridad (SOC) o los Equipos de Respuesta a Incidentes (CERT).

En estos entornos, la velocidad de reacción es tan importante como la precisión, ya que cada segunda cuenta para prevenir daños y contener una posible intrusión. Con lo anterior se ejemplifica en la figura 16 el desarrollo del modelo predictivo en malware.

Figura 16

Laboratorio de modelos predictivos de malware.

Nota: Validación en laboratorio en un entorno controlado. Fuente: elaboración propia.

Pruebas en pre-producción

Despliegue seguro. En esta fase, el modelo se implementa en modo shadow, es decir, en paralelo al sistema de detección existente, pero sin intervenir en la operación productiva. De esta forma, el modelo procesa tráfico y archivos reales, genera predicciones y emite alertas, pero sin bloquear ni modificar el flujo de la infraestructura crítica. Esta modalidad permite evaluar el comportamiento del sistema en condiciones auténticas de operación, minimizando riesgos para la organización.

Monitoreo y registro. Todas las predicciones del modelo deben acompañarse de explicaciones generadas mediante técnicas de Explainable AI (como SHAP o LIME), lo que facilita al equipo de análisis comprender las decisiones y evaluar su validez. Adicionalmente, se registran métricas como tasas de falsos positivos/negativos, distribución de confianza de las predicciones y frecuencia de detección por familia de malware.

Retroalimentación humana. La interacción con el equipo de analistas resulta fundamental en este punto. Cada alerta generada se somete a verificación manual, permitiendo confirmar si se trata de un caso verdadero o un error. Esta retroalimentación no solo mejora la calidad del dataset con nuevas etiquetas confiables, sino que también sirve como insumo directo para futuros procesos de reentrenamiento.

Gestión del label delay. Un aspecto crítico para monitorear es el label delay, es decir, el tiempo que transcurre entre la generación de una alerta por el modelo y su confirmación por parte de un analista humano. Este retraso tiene impacto directo en la eficacia de los procesos de actualización y retroalimentación. Medir y gestionar este indicador permite ajustar políticas de reentrenamiento, priorizar muestras críticas y definir estrategias de actualización incremental más eficientes (García et al., 2014), en la figura 17 se visualiza en desarrollo de las pruebas.

Figura 17

Pruebas en pre-producción.

Nota: Pruebas en pre-producción en el modelo de implementación en modo shadow. Fuente: elaboración propia.

Actualización incremental / aprendizaje continuo

En el ámbito de la ciberseguridad moderna, las amenazas evolucionan de manera constante, adaptándose a las defensas y generando nuevas variantes que desafían los modelos tradicionales de detección. Por ello, los sistemas basados en ML no pueden permanecer estáticos: deben aprender de forma continua y ajustarse progresivamente al entorno cambiante.

El aprendizaje continuo o incremental permite mantener la vigencia de los detectores ante el concept drift, la aparición de familias inéditas de malware y los cambios en los patrones de ataque. Este enfoque asegura que el modelo conserve su capacidad de predicción y confiabilidad operativa a lo largo del tiempo, incluso frente a escenarios desconocidos.

Política de actualización. La actualización de los modelos no debe depender exclusivamente de revisiones manuales o cronogramas fijos. En su lugar, se recomienda

implementar disparadores automáticos (triggers) que activen los procesos de actualización cuando se detecten desviaciones significativas en el desempeño del sistema.

Algunos de los criterios más relevantes son:

- Caída notable en las métricas de rendimiento, como una reducción del F1-score superior a tres puntos respecto al modelo base.
- Aumento de falsos positivos o negativos, especialmente si el False Positive Rate (FPR) se duplica frente al valor de referencia.
- Detección de nuevas familias de malware no contempladas durante el entrenamiento inicial.
- Alertas de baja confianza sostenida, que indican incertidumbre del modelo ante muestras recientes.

Estos mecanismos automáticos garantizan que el sistema responda de forma dinámica a la realidad operativa, evitando que las degradaciones de rendimiento pasen desapercibidas y asegurando la actualización oportuna del modelo.

Mecanismos de actualización. El proceso de actualización puede implementarse a través de diferentes estrategias, según el tipo de modelo y las capacidades tecnológicas de la organización. A continuación, se presentan los principales enfoques utilizados:

- Caso 1: Online learning (aprendizaje en flujo).

En este método, el modelo se actualiza de manera inmediata con cada nueva muestra de datos. Es compatible con algoritmos incrementales como Naïve Bayes, árboles Hoeffding o regresiones basadas en gradiente estocástico (SGD). Herramientas como River o scikit-multiflow facilitan su aplicación en sistemas de detección en tiempo real.

- Caso 2: Mini-batch retraining (reentrenamiento por lotes).

Se realiza un reentrenamiento periódico, por ejemplo, semanal o quincenal, utilizando lotes de muestras nuevas verificadas. Este método logra un equilibrio entre la estabilidad del modelo y su capacidad de adaptación frente a cambios graduales en el entorno.

- Caso 3: Warm-start en redes profundas.

En arquitecturas de DL, se pueden reutilizar los pesos entrenados previamente y extender el aprendizaje sobre conjuntos ampliados. Esto reduce el costo computacional y acelera la adaptación sin necesidad de entrenar desde cero.

- Caso 4: Ensemble rolling (ensamble dinámico).

Consiste en integrar submodelos entrenados con datos recientes, mientras se eliminan aquellos que muestran bajo rendimiento. Este enfoque mantiene un equilibrio entre diversidad, estabilidad y robustez, características clave para la detección en entornos cambiantes.

En todos los casos, es imprescindible mantener una gestión rigurosa de versiones, documentando los cambios en los modelos y datasets. Además, se recomienda conservar un modelo "canario", utilizado como punto de comparación para medir la mejora o degradación del sistema después de cada actualización.

Control de calidad. Antes de implementar un modelo actualizado en entornos productivos, se debe aplicar un protocolo de validación exhaustivo, que asegure la confiabilidad y estabilidad del nuevo sistema.

Este proceso debe incluir los siguientes pasos:

- Evaluación en holdout temporal: verificar que el modelo mantiene su desempeño frente a la deriva temporal (concept drift), utilizando particiones de datos separadas por periodos.

- Pruebas adversariales: someter el modelo a muestras generadas con redes GANs o técnicas de ofuscación y packing, con el fin de medir su resistencia ante intentos de evasión.
- Evaluación de latencia y cumplimiento de SLA: confirmar que los tiempos de respuesta se ajustan a los requerimientos operativos del entorno, como los de un SOC que opera en tiempo casi real.

Solo los modelos que superen todos los criterios de aceptación predefinidos en términos de rendimiento, estabilidad y cumplimiento de SLA podrán ser promovidos a producción. Este proceso está ilustrado en la Figura 18, que muestra el flujo de actualización incremental y control de calidad.

Figura 18

Proceso de Actualización Incremental y Control de Calidad de Modelos de Detección de Malware.

Nota: Actualización incremental / aprendizaje continuo. Fuente: elaboración propia.

Como resultado, se definió un protocolo metodológico completo de validación, que articula de forma coherente:

- La partición temporal de los datos para la detección de concept drift,
- La evaluación continua mediante métricas operativas alineadas a escenarios reales de SOC,
- La incorporación sistemática de pruebas adversariales para medir la robustez frente a técnicas de evasión modernas, y
- La integración de mecanismos de XAI que fortalecen la interpretabilidad y la confianza del analista humano en los resultados del modelo.

Validación de Modelos Predictivos en Detección de Malware

La validación de modelos predictivos constituye un componente crítico en el desarrollo de sistemas de detección de malware basados en ML, ya que permite garantizar no solo el rendimiento del modelo en condiciones controladas, sino su efectividad, robustez y adaptabilidad en entornos reales. En el contexto de la ciberseguridad, donde las amenazas evolucionan constantemente, la validación debe trascender las métricas tradicionales e incorporar escenarios dinámicos, adversariales y operativos.

Validación tradicional vs validación en ciberseguridad

De manera convencional, los modelos de ML son evaluados mediante particiones de entrenamiento y prueba (train/test split) y técnicas de validación cruzada. Si bien estas estrategias permiten medir el desempeño inicial del modelo, resultan insuficientes frente a la naturaleza cambiante del malware. En entornos de ciberseguridad, la validación debe considerar:

Variabilidad temporal (concept drift). Aparición de nuevas familias de malware no presentes en el entrenamiento.

Datos desbalanceados. Predominio de muestras benignas frente a maliciosas.

Evasión adversarial. Manipulación intencional de características para evadir detección.

Retraso en etiquetado (label delay). Dificultad para obtener etiquetas confiables en tiempo real.

Por lo anterior, se requiere un enfoque de validación más robusto y adaptativo.

Estrategia de validación robusta propuesta

En este trabajo se adopta un esquema de validación multicapa, alineado con el pipeline predictivo diseñado, el cual se estructura en las siguientes fases:

Validación offline. Corresponde a la fase inicial de evaluación del modelo predictivo, en la cual se emplean datasets históricos previamente curados, balanceados y etiquetados para medir su desempeño en condiciones controladas. Esta etapa se apoya en técnicas como la validación cruzada (k-fold) y la partición de datos en conjuntos de entrenamiento y prueba (70/30 u 80/20), permitiendo estimar la capacidad de generalización del modelo y reducir el riesgo de sobreajuste. Para la evaluación se utilizan métricas estándar como el F1-score, el AUC-ROC, la precisión (precision) y el recall, priorizando este último en contextos de ciberseguridad debido a la necesidad de minimizar falsos negativos. El objetivo principal de esta validación es identificar y seleccionar los modelos candidatos con mejor desempeño para fases posteriores de validación más robusta.

Validación temporal. Consiste en evaluar el modelo utilizando datos organizados de forma cronológica, donde el entrenamiento se realiza con información histórica y la prueba con datos más recientes. Este enfoque permite simular de manera más realista las condiciones de operación en entornos productivos, donde las amenazas evolucionan constantemente. A través de esta validación es posible medir la degradación del rendimiento del modelo a lo largo del tiempo, identificar fenómenos de concept drift (cambios en la distribución de los datos o en el comportamiento del malware) y analizar la capacidad del modelo para adaptarse a nuevas variantes

de ataque. En consecuencia, esta técnica se convierte en un componente clave para garantizar la robustez, vigencia y confiabilidad de los sistemas de detección en escenarios organizacionales reales.

Validación adversarial. Consiste en evaluar el modelo mediante la introducción de muestras modificadas intencionalmente utilizando técnicas de evasión, ofuscación o generación adversarial, como las GANs. Este enfoque permite someter al modelo a escenarios desafiantes que simulan ataques reales diseñados para evadir los mecanismos de detección. A través de esta validación se evalúa la robustez del modelo frente a manipulaciones en los datos de entrada, se identifican vulnerabilidades en el espacio de características y se analizan posibles puntos débiles en su capacidad de generalización. En consecuencia, este proceso contribuye significativamente al fortalecimiento de la resiliencia del sistema, permitiendo ajustar y endurecer los modelos para enfrentar amenazas avanzadas y adaptativas en entornos organizacionales.

Validación en entorno controlado (sandbox/laboratorio). Consiste en evaluar el desempeño del modelo utilizando muestras reales de malware ejecutadas en ambientes aislados y supervisados, como sandboxes especializados (por ejemplo, Cuckoo Sandbox). Este tipo de validación permite observar el comportamiento del malware en tiempo real, incluyendo la ejecución de llamadas API, modificaciones en el sistema, actividad en red y mecanismos de persistencia. A través de este enfoque, se verifica la capacidad del modelo para detectar amenazas activas bajo condiciones cercanas a escenarios reales de operación. Asimismo, facilita la correlación entre las predicciones del modelo y los eventos generados durante la ejecución, permitiendo validar su efectividad práctica, reducir falsos negativos y fortalecer la confiabilidad del sistema en entornos organizacionales.

Validación en pre-producción (shadow mode). Consiste en desplegar el modelo de ML en paralelo con los sistemas de seguridad existentes dentro de la organización, sin intervenir directamente en la operación ni en la toma de decisiones automatizadas. En este enfoque, el modelo analiza los mismos flujos de datos que las herramientas tradicionales como SIEM, EDR o IDS/IPS, permitiendo realizar una comparación directa de resultados. A través de esta validación, se miden indicadores clave como la tasa de falsos positivos y falsos negativos en tiempo real, así como la capacidad de detección frente a amenazas reales. Además, se evalúa el impacto operativo del modelo, considerando aspectos como carga computacional, tiempos de respuesta y utilidad para los analistas de seguridad. Este proceso permite determinar la viabilidad del modelo en un entorno productivo, garantizando que su implementación futura aporte valor sin afectar la estabilidad ni la eficiencia de la infraestructura organizacional.

Validación continua (producción). Una vez desplegado en el entorno organizacional, el modelo de ML entra en una fase de monitoreo continuo, orientada a garantizar su rendimiento, estabilidad y capacidad de adaptación frente a la evolución de las amenazas. Este proceso implica la supervisión permanente de métricas en producción, tales como precisión, recall, F1-score y tasa de falsos positivos, con el fin de identificar posibles degradaciones en su desempeño. Asimismo, se incorporan mecanismos de detección de data drift y concept drift, que permiten evidenciar cambios en la distribución de los datos o en los patrones de ataque, los cuales pueden afectar la capacidad predictiva del modelo. Como respuesta a estas variaciones, se implementan estrategias de retraining periódico o adaptativo, utilizando nuevos datos provenientes de la telemetría organizacional, lo que asegura que el modelo se mantenga actualizado, relevante y alineado con el comportamiento dinámico del ecosistema de ciberamenazas. Este enfoque garantiza la sostenibilidad del sistema y su efectividad a largo plazo en entornos reales.

Criterios de aceptación del modelo

Para garantizar que un modelo de ML sea apto para su implementación en entornos organizacionales, se establecen umbrales mínimos de desempeño que permiten evaluar de forma objetiva su viabilidad operativa. En este sentido, se considera un F1-score igual o superior a 0.85 como indicador de un adecuado balance entre precisión y recall, mientras que se prioriza un recall superior al 0.90 con el fin de minimizar los falsos negativos, especialmente críticos en escenarios de ciberseguridad donde una amenaza no detectada puede generar impactos significativos. Adicionalmente, se exige una tasa de falsos positivos (FPR) controlada, que evite la sobrecarga operativa en los equipos de seguridad y reduzca alertas innecesarias. A estos criterios se suma la estabilidad temporal del modelo, evaluada mediante validación cronológica para asegurar que su rendimiento no se degrade con el tiempo, así como su resistencia frente a muestras evasivas o adversariales, lo que garantiza robustez ante técnicas avanzadas de ataque. En conjunto, estos criterios constituyen una base técnica sólida para la toma de decisiones informadas respecto al despliegue, ajuste o reentrenamiento del modelo dentro de la organización.

Relación con el pipeline predictivo

La validación de modelos se integra de manera transversal dentro del pipeline predictivo propuesto, articulándose específicamente en las fases de evaluación del modelo, validación robusta y monitoreo con retroalimentación continua. En la fase de evaluación, se aplican métricas cuantitativas como F1-score, AUC-ROC, recall y tasa de falsos positivos, permitiendo seleccionar el modelo con mejor desempeño inicial bajo condiciones controladas. Posteriormente, en la fase de validación robusta, se someten los modelos a escenarios más exigentes, incluyendo validación

temporal para identificar degradación por concept drift, pruebas con muestras evasivas y generación de datos adversariales, con el fin de evaluar su resiliencia frente a amenazas reales y cambiantes. Finalmente, en la etapa de monitoreo y retroalimentación continua, el modelo es observado en producción mediante el seguimiento de métricas operativas, detección de desviaciones en los datos y activación de procesos de reentrenamiento periódico o incremental. Esta integración asegura la trazabilidad completa del ciclo de vida del modelo, desde la recolección y preparación de datos hasta su despliegue y evolución en entornos organizacionales, garantizando no solo su rendimiento técnico, sino también su adaptabilidad, confiabilidad y sostenibilidad en el tiempo.

Como resultado, se obtiene un modelo validado bajo múltiples escenarios (históricos, temporales, adversariales y reales), lo que garantiza su capacidad para enfrentar amenazas modernas como ransomware, malware polimórfico y ataques basados en evasión.

En este sentido, la validación deja de ser una etapa final del proceso y se convierte en un mecanismo continuo de aseguramiento de calidad, fundamental para la sostenibilidad de los sistemas de ciberseguridad basados en ML.

Asimismo, el capítulo materializa el objetivo propuesto mediante la definición de un pipeline de validación adaptativa, que abarca desde la preparación y curado de datos, el entrenamiento inicial y la validación offline robusta, hasta las pruebas en laboratorio, la evaluación en pre-producción bajo esquemas shadow y los mecanismos de actualización incremental y aprendizaje continuo. Este pipeline constituye un producto metodológico aplicable, replicable y alineado con las necesidades operativas de organizaciones que enfrentan amenazas avanzadas como ransomware polimórfico, malware evasivo y ataques adversariales.

De manera adicional, se establecieron criterios técnicos y umbrales operativos claros (F1-score, FPR, TTD, degradación temporal) que permiten decidir objetivamente cuándo un modelo debe ser reentrenado, ajustado o promovido a producción, fortaleciendo la trazabilidad, auditabilidad y gobernanza del ciclo de vida del modelo. La documentación de artefactos, checkpoints y reportes de robustez refuerza este cumplimiento, garantizando reproducibilidad y control de calidad.

Discusión

La presente investigación abordó de manera sistemática el problema central asociado a la creciente sofisticación y volumen de las amenazas de malware que enfrentan las organizaciones, demostrando que el uso estratégico del ML constituye una respuesta viable, eficaz y sostenible frente a este desafío. A lo largo del estudio se analizó el grado de cumplimiento de los objetivos propuestos y la pertinencia de los métodos empleados, evidenciando una coherencia sólida entre el planteamiento del problema, el marco teórico y los resultados obtenidos.

En relación con el objetivo general, los hallazgos confirman que fue plenamente alcanzado, al evidenciar la aplicabilidad del ML para la detección, prevención y mitigación del malware en entornos organizacionales, integrando enfoques teóricos, experimentales y metodológicos. El análisis de los resultados demuestra una validez metodológica robusta, sustentada en el uso de métricas rigurosas como el F1-score, el AUC-ROC y la validación cruzada, así como en la replicabilidad de los experimentos en entornos simulados y escenarios cercanos a contextos reales de operación.

Como primero es centrado en la revisión sistemática de la literatura, permitió identificar tendencias, enfoques predominantes y retos persistentes en el análisis de malware mediante ML y DL. Este ejercicio evidenció la necesidad de modelos más adaptativos, explicables y robustos frente a amenazas polimórficas y evasivas, estableciendo una base conceptual sólida para el desarrollo del resto de la investigación.

En segundo lugar, es orientado a evaluar las estrategias de análisis estático, dinámico e híbrido, se materializó mediante un análisis comparativo que permitió establecer criterios técnicos y estratégicos para su aplicabilidad en entornos organizacionales. Los resultados confirmaron que, si bien cada enfoque presenta ventajas y limitaciones, los modelos híbridos ofrecen mayores

niveles de precisión, resiliencia y valor operativo, alineándose con los desafíos identificados en investigaciones recientes sobre ransomware y malware avanzado.

El tercer lugar, es enfocado en el diseño de modelos predictivos con énfasis en la calidad, representatividad y equilibrio de los datasets, se cumplió a través de la propuesta de un pipeline predictivo mejorado. Este incorporó técnicas de data augmentation, normalización contextual y selección de características orientadas a amenazas modernas, evidenciando mejoras significativas en la capacidad de detección y generalización de los modelos. Los resultados obtenidos validan que la ingeniería de datos es un factor determinante para reducir el impacto del desbalance de clases y fortalecer la eficacia de los sistemas de detección.

El cuarto lugar, es relacionado con la validación adaptativa de los modelos, se concretó mediante el diseño de un marco de validación robusto que integra evaluación temporal, pruebas adversariales, validación en laboratorio y mecanismos de actualización incremental. Este enfoque permitió abordar desafíos críticos como el concept drift, la incorporación de muestras recientes y la degradación progresiva del rendimiento. Asimismo, la integración de técnicas de XAI fortaleció la transparencia, trazabilidad y confianza operativa del sistema, aspectos clave en escenarios de toma de decisiones críticas dentro de los equipos de seguridad.

Si bien los modelos desarrollados alcanzaron altos niveles de precisión, la discusión reconoce que su desempeño puede verse afectado por la aparición de nuevas familias de malware o técnicas de evasión más sofisticadas. No obstante, el uso de datasets estandarizados, la validación cruzada y los esquemas de actualización incremental propuestos permiten mejorar su capacidad de adaptación y generalización. El contraste con estudios previos confirma que el enfoque adoptado responde a los desafíos contemporáneos del campo y aporta valor comparativo e innovador.

De manera transversal, el estudio evidenció que la efectividad del ML en ciberseguridad no depende exclusivamente de los algoritmos, sino de un enfoque integral que combine datos de calidad, validación rigurosa, explicabilidad y articulación con los procesos organizacionales. En este sentido, se destaca la versatilidad y adaptabilidad del pipeline propuesto, cuya estructura modular permite su implementación tanto en organizaciones con infraestructuras complejas como en entornos con recursos limitados, sin sustituir el criterio humano en la toma de decisiones.

En conclusión, los resultados obtenidos confirman que los objetivos del estudio fueron pertinentes, alcanzables y coherentes con el problema de investigación. Los métodos empleados demostraron ser adecuados y eficaces para mejorar la detección, prevención y mitigación del malware en contextos organizacionales. Los hallazgos establecen un puente sólido entre la teoría y la práctica, garantizando validez científica, aplicabilidad operativa y continuidad investigativa, y aportando directrices técnicas y estratégicas para el fortalecimiento de los sistemas predictivos de ciberseguridad en organizaciones contemporáneas.

Conclusiones

El análisis de la literatura y de los resultados obtenidos evidencia que el ML y el DL se han consolidado como componentes estratégicos fundamentales en la defensa frente a las amenazas cibernéticas contemporáneas. Su capacidad para procesar grandes volúmenes de datos, identificar patrones complejos y adaptarse a nuevas variantes de malware permite el desarrollo de soluciones más inteligentes, proactivas y eficaces frente a un entorno de amenazas en constante evolución.

No obstante, el estudio confirma que la adopción de estas tecnologías enfrenta desafíos relevantes, entre los que destacan la limitada interpretabilidad de algunos modelos, la escasez de datos etiquetados de alta calidad y la necesidad de enfoques interdisciplinarios que integren conocimientos técnicos, analíticos y de gestión del riesgo. Superar estas limitaciones resulta esencial para garantizar la aplicabilidad real y sostenible de los modelos en contextos organizacionales.

La investigación demuestra que la evolución del malware exige enfoques de análisis adaptativos y estratégicos, donde el ML y el DL, aplicados sobre técnicas estáticas, dinámicas e híbridas, incrementan significativamente la precisión, escalabilidad y resiliencia de los sistemas de detección. En este contexto, se concluye que el análisis estático ofrece ventajas en términos de eficiencia computacional y cobertura, aunque es vulnerable a técnicas de ofuscación y polimorfismo; el análisis dinámico permite identificar comportamientos maliciosos complejos, como el ransomware, aunque puede verse afectado por mecanismos de evasión; mientras que el análisis híbrido se consolida como el enfoque más completo al combinar las fortalezas de ambos y reducir las tasas de falsos positivos y negativos.

Los casos prácticos y experimentales analizados validan la efectividad de estas estrategias, evidenciando mejoras significativas en la detección de ransomware, botnets y variantes

polimórficas mediante el uso de secuencias de llamadas API, análisis de tráfico real y redes neuronales profundas. Estas evidencias permiten establecer criterios técnicos y estratégicos para la aplicación organizacional de los modelos, considerando el objetivo operativo, el nivel de madurez en ciberseguridad y los recursos disponibles en cada entorno.

Asimismo, se concluye que el desarrollo de modelos predictivos confiables depende de manera crítica de la calidad, diversidad y representatividad de los datasets, así como de la incorporación de técnicas de data augmentation, normalización contextual y selección inteligente de características. Estas prácticas contribuyen a reducir sesgos, mitigar el desbalance de clases y mejorar la capacidad de generalización de los algoritmos frente a amenazas emergentes y evasivas.

La implementación de un pipeline predictivo integral, que abarque desde la recolección y curación de datos hasta la validación robusta y el despliegue controlado, se perfila como un factor clave para fortalecer la detección, prevención y mitigación del malware en las organizaciones. Los resultados evidencian que el éxito de los modelos de ML no reside únicamente en la elección del algoritmo, sino en la construcción de un flujo de datos sólido, resiliente y escalable, capaz de adaptarse a la evolución del ecosistema de amenazas digitales.

Finalmente, se concluye que la rápida transformación del malware, caracterizada por el polimorfismo, las técnicas anti-VM y los mecanismos de evasión, pone de manifiesto las limitaciones de los modelos estáticos tradicionales. En este escenario, la validación adaptativa, apoyada en particiones temporales, monitoreo continuo de métricas, pruebas adversariales y actualización incremental, emerge como un elemento esencial para afrontar el concept drift, reducir los efectos del retraso en el etiquetado y garantizar la transparencia mediante el uso de técnicas de XAI. Este enfoque sienta las bases para el desarrollo de sistemas predictivos adaptativos, confiables y alineados con las necesidades operativas reales, capaces de evolucionar

al ritmo de las amenazas modernas y responder de manera efectiva a los desafíos crecientes de la ciberseguridad actual.

Recomendaciones

Para fortalecer la detección, prevención y mitigación de amenazas de malware en entornos organizacionales, se recomienda ampliar y diversificar de manera continua las fuentes de datos utilizadas en los procesos de entrenamiento y validación de los modelos de ML. En este sentido, resulta pertinente combinar repositorios públicos reconocidos, como EMBER, CICIDS u OMD, con telemetría interna corporativa proveniente de soluciones EDR, capturas de tráfico PCAP y trazas de llamadas API, lo que permite construir modelos más representativos, reducir sesgos y mejorar la capacidad de identificación de amenazas emergentes.

Dado el dinamismo del ecosistema de amenazas, se recomienda incorporar mecanismos de reentrenamiento consciente de la deriva (drift-aware retraining) y aprendizaje incremental, que posibiliten la actualización periódica o automática de los modelos ante variaciones significativas en su desempeño. Para ello, es fundamental definir umbrales operativos claros, tales como descensos relevantes en métricas como el F1-score o incrementos en la tasa de falsos positivos, que activen procesos de reentrenamiento controlados, trazables y debidamente documentados.

Asimismo, se sugiere implementar un protocolo permanente de evaluación del rendimiento, que integre pruebas offline, validaciones temporales y la generación de muestras adversariales mediante frameworks especializados. Este enfoque contribuye a fortalecer la resiliencia de los sistemas predictivos, permitiendo evaluar su comportamiento frente a intentos de evasión y ataques diseñados para manipular las características de los datos.

La transparencia y la confianza operativa constituyen elementos esenciales en la adopción de modelos de ML en ciberseguridad. Por ello, se recomienda integrar de forma sistemática herramientas de XAI, como SHAP, LIME o CAM, dentro de los sistemas de monitoreo y análisis.

Estas técnicas facilitan la comprensión de las decisiones del modelo, la identificación de las variables más relevantes y la priorización informada de la respuesta ante incidentes.

Adicionalmente, se recomienda que todas las fases del ciclo de vida del modelo, desde la recolección y curación de los datos hasta su despliegue y actualización en producción, sean documentadas y versionadas de manera rigurosa. Esta práctica garantiza trazabilidad, reproducibilidad y cumplimiento de los estándares éticos y normativos, además de facilitar la auditoría y la mejora continua de los sistemas implementados.

El éxito del ML en la ciberseguridad organizacional depende no solo de la tecnología, sino también de su adecuada articulación con las políticas, los procesos y el talento humano. En este contexto, se recomienda fomentar una colaboración permanente entre equipos de seguridad, científicos de datos y responsables de cumplimiento normativo, asegurando que los resultados analíticos se traduzcan en acciones operativas coherentes con los objetivos estratégicos de la organización.

De cara a trabajos futuros, se sugiere profundizar en el desarrollo de modelos capaces de aprender a partir de datos parcialmente etiquetados o generados artificialmente, así como en el diseño de defensas frente a ataques adversariales. Estas líneas de investigación resultan clave para incrementar la robustez, sostenibilidad y adaptabilidad de los sistemas predictivos en entornos cada vez más complejos y dinámicos.

Finalmente, se recomienda realizar un análisis costo-beneficio integral que contemple los recursos computacionales, humanos y temporales requeridos para el mantenimiento y evolución de los modelos. Este ejercicio permitirá evaluar la viabilidad del sistema, optimizar el uso de los recursos disponibles y planificar estrategias sostenibles a largo plazo que equilibren desempeño, seguridad y eficiencia operativa.

Referencias Bibliográficas

Afianian, A., Niksefat, S., Sadeghiyan, B., & Baptiste, D. (2018). *Malware Dynamic Analysis Evasion Techniques: A Survey*. https://arxiv.org/abs/1811.01190

Al-Hadhrami, S., Al-Khalifa, H., & Al-Sarem, M. (2022). Improving malware detection using feature selection and normalization: A LightGBM approach. *Applied Sciences*, *12*(23), 12156. https://doi.org/10.3390/app122312156

Al-rimy, B. A. S., Maarof, M. A., & Shaid, S. Z. M. (2018). Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions. *Computers & Security*, *74*, 144–166. https://doi.org/https://doi.org/10.1016/j.cose.2018.01.001

Anand, P. M., Charan, P. V. S., & Shukla, S. K. (2022). A Comprehensive API Call Analysis for Detecting Windows-Based Ransomware. *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 337–344. https://doi.org/10.1109/CSR54599.2022.9850320

Bala, B., & Behal, S. (2024). AI techniques for IoT-based DDoS attack detection: Taxonomies, comprehensive review and research challenges. *Computer Science Review*, *52*, 100631. https://doi.org/https://doi.org/10.1016/j.cosrev.2024.100631

Bensaoud, A., Kalita, J., & Bensaoud, M. (2024). A survey of malware detection using deep learning. *Machine Learning With Applications*, *16*, 100546. https://doi.org/10.1016/j.mlwa.2024.100546

Bhardwaj, A., Mangat, V., Vig, R., Halder, S., & Conti, M. (2021). Distributed denial of service attacks in cloud: State-of-the-art of scientific and commercial solutions. *Computer Science Review*, *39*, 100332. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100332

Birthriya, S. K., Ahlawat, P., & Jain, A. K. (2025). Detection and prevention of spear phishing attacks: A comprehensive survey. *Computers & Security*, *151*, 104317. https://doi.org/https://doi.org/10.1016/j.cose.2025.104317

Congreso de la República de Colombia. (2009). *Ley 1273 de 2009 Protección de la información y de los datos*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=34492

Congreso de la República de Colombia. (2012). *Ley 1581 de 2012: Disposiciones generales para la protección de datos personales*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Contreras, R., & Contreras, R. (2023). *Los 10 ciberataques más grandes de la década*. https://www.computing.es/seguridad/los-10-ciberataques-mas-grandes-de-la-decada/

Departamento Nacional de Planeación. (2020). *CONPES 3995: Política Nacional de Confianza y Seguridad Digital*. https://colaboracion.dnp.gov.co/cdt/Conpes/Económicos/3995.pdf

Departamento Nacional de Planeación. (2023). *CONPES 4144: Política nacional de inteligencia artificial* (Number 4144). https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/4144.pdf

Díaz, F. J., Molinari, L. H., Venosa, P., Macia, N., Lanfranco, E. F., & Sabolansky, A. J. (2019). *Investigación en ciberseguridad: nuevos desafíos para adaptarse a nuevos paradigmas* [Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/77274

Díaz, F. J., Venosa, P., Macia, N., Lanfranco, E. F., Sabolansky, A. J., Durante, M., Rubio, D., & Pretto, J. (2021). *Investigación en ciberseguridad en un año de pandemia* [Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/120528

Dixit, P., & Silakari, S. (2020). Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review. *Computer Science Review*, *39*, 100317. https://doi.org/10.1016/j.cosrev.2020.100317

Económica, L. N. (2024). *Según IBM, Colombia en 2024 ha recibido el 17 % de los ciberataques en Latinoamérica*. https://lanotaeconomica.com.co/movidas-empresarial/segun-ibm-colombia-en-2024-ha-recibido-el-17-de-los-ciberataques-en-latinoamerica-el-60-afectaron-al-sector-salud/

Elhanashi A.; Dini P. (2024). *Machine Learning for Cybersecurity: Threat Detection and Mitigation* (A. Elhanashi & P. Dini, Eds.). MDPI. https://doi.org/10.3390/books978-3-7258-2793-0

El-Sofany, H., El-Seoud, S. A., & Karam, O. H. et al. (2024). Using machine learning algorithms to enhance IoT system security. *Scientific Reports*, *14*, 12077. https://doi.org/10.1038/s41598-024-62861-y

Feng, P., Gai, L., Yang, L., Wang, Q., Li, T., Xi, N., & Ma, J. (2024). DawnGNN: Documentation augmented windows malware detection using graph neural network. *Computers & Security*, *140*, 103788. https://doi.org/https://doi.org/10.1016/j.cose.2024.103788

García, S., Grill, M., Stiborek, J., & Zunino, A. (2014). The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background traffic. *Comput. Secur.*, *45*, 100–123. https://doi.org/10.1016/j.cose.2014.05.011

Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of*

Network and Computer Applications, *153*, 102526. https://doi.org/https://doi.org/10.1016/j.jnca.2019.102526

Guerra-Manzanares, A. (2024). Machine Learning for Android Malware Detection: Mission Accomplished? A Comprehensive Review of Open Challenges and Future Perspectives. *Computers & Security*, *138*. https://doi.org/https://doi.org/10.1016/j.cose.2023.103654

Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). Machine Learning and Deep Learning Approaches for CyberSecurity: A Review. *IEEE Access*, *10*, 19572–19585. https://doi.org/10.1109/ACCESS.2022.3151248

Hanif, H., Md Nasir, M. H. N., Ab Razak, M. F., Firdaus, A., & Anuar, N. B. (2021). The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches. *Journal of Network and Computer Applications*, *179*. https://doi.org/https://doi.org/10.1016/j.jnca.2021.103009

Hernández-Pereira, E. (2022). *Técnicas de aprendizaje máquina para análisis de malware* [Universidade da Coruña]. https://ruc.udc.es/dspace/handle/2183/32112

International Organization for Standardization. (2012). *ISO/IEC 27032:2012 – Guidelines for cybersecurity*. https://www.iso.org/standard/44375.html

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection – Information security management systems – Requirements*. https://www.iso.org/standard/82875.html

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 – Artificial intelligence management system*. https://www.gsc-co.com/wp-content/uploads/2024/08/SCAN-ISO-420012023_-Web.pdf

Kalambe, D., Sharma, D., Kadam, P., & Surati, S. (2025). A comprehensive plane-wise review of DDoS attacks in SDN: Leveraging detection and mitigation through machine learning and deep learning. *Journal of Network and Computer Applications*, *235*, 104081. https://doi.org/https://doi.org/10.1016/j.jnca.2024.104081

Kalouptsoglou, I., Siavvas, M., Ampatzoglou, A., Kehagias, D., & Chatzigeorgiou, A. (2023). Software vulnerability prediction: A systematic mapping study. *Information and Software Technology*, *164*, 107303. https://doi.org/https://doi.org/10.1016/j.infsof.2023.107303

Kirat, D., Vigna, G., & Kruegel, C. (2014). BareCloud: Bare-metal Analysis-based Evasive Malware Detection. *23rd USENIX Security Symposium (USENIX Security 14)*, 287–301. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/kirat

Kritika, Er. (2025). A comprehensive literature review on ransomware detection using deep learning. *Cyber Security and Applications*, *3*, 100078. https://doi.org/https://doi.org/10.1016/j.csa.2024.100078

Kumar, S., Dwivedi, M., Kumar, M., & Gill, S. S. (2024). A comprehensive review of vulnerabilities and AI-enabled defense against DDoS attacks for securing cloud services. *Computer Science Review*, *53*, 100661. https://doi.org/https://doi.org/10.1016/j.cosrev.2024.100661

Ling, X., Wu, L., Zhang, J., Qu, Z., Deng, W., Chen, X., Qian, Y., Wu, C., Ji, S., Luo, T., Wu, J., & Wu, Y. (2023). Adversarial attacks against Windows PE malware detection: A survey of

the state-of-the-art. *Computers & Security*, *128*, 103134. https://doi.org/https://doi.org/10.1016/j.cose.2023.103134

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://arxiv.org/abs/1705.07874

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, *13*(3), e1493. https://doi.org/https://doi.org/10.1002/widm.1493

McLaughlin, N., & del Rincon, J. M. (2022). *Data Augmentation for Opcode Sequence Based Malware Detection*. https://arxiv.org/abs/2106.11821

Ministerio de Ciencia Tecnología e Innovación. (2025). *Proyecto de ley de inteligencia artificial en Colombia*. https://especiales.minciencias.gov.co/wp-content/uploads/2025/07/pl_ia_finalizado.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones. (2023). *Estrategia Nacional de Seguridad Digital de Colombia*. https://www.mintic.gov.co/portal/715/articles-403023_recurso_2.pdf

Ministerio de Tecnologías de la Información y las Comunicaciones. (2024). *Lineamientos y estrategias para el desarrollo tecnológico y la inteligencia artificial en Colombia*. https://www.mintic.gov.co/portal/715/articles-425888_recurso_1.pdf

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and

interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, *8*, 1526221. https://doi.org/10.3389/frai.2025.1526221

Nataraj, L., Mohammed, T., Manjunath, B. S., & Chandrasekaran, S. (2021). Data augmentation and transfer learning to classify malware images in a deep learning context. *Journal of Computer Virology and Hacking Techniques*, *17*(3), 279–297. https://doi.org/10.1007/s11416-021-00381-3

Nazim Sadia AND Alam, M. M. A. N. D. R. S. S. A. N. D. M. J. C. A. N. D. H. S. S. A. N. D. S. M. M. (2025). Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM. *PLOS ONE*, *20*(5), 1–24. https://doi.org/10.1371/journal.pone.0318542

Pinhero, A., M L, A., P, V., Visaggio, C. A., N, A., S, A., & S, A. (2021). Malware detection employed by visualization and deep neural network. *Computers & Security*, *105*, 102247. https://doi.org/https://doi.org/10.1016/j.cose.2021.102247

Pritee, Z. T., Anik, M. H., Alam, S. B., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). Machine learning and deep learning for user authentication and authorization in cybersecurity: A state-of-the-art review. *Computers & Security*, *140*, 103747. https://doi.org/https://doi.org/10.1016/j.cose.2024.103747

Prome, S. A., Ragavan, N. A., Islam, M. R., Asirvatham, D., & Jegathesan, A. J. (2024). Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review. *Natural Language Processing Journal*, *6*, 100057. https://doi.org/https://doi.org/10.1016/j.nlp.2024.100057

Qiang, W., Yang, L., & Jin, H. (2022). Efficient and Robust Malware Detection Based on Control Flow Traces Using Deep Neural Networks. *Computers & Security*, *122*, 102871. https://doi.org/https://doi.org/10.1016/j.cose.2022.102871

Qureshi, S. U., He, J., Tunio, S., Zhu, N., Nazir, A., Wajahat, A., Ullah, F., & Wadud, A. (2024). Systematic review of deep learning solutions for malware detection and forensic analysis in IoT. *Journal of King Saud University - Computer and Information Sciences*, *36*(8), 102164. https://doi.org/https://doi.org/10.1016/j.jksuci.2024.102164

Razgallah, A., Khoury, R., Hallé, S., & Khanmohammadi, K. (2021). A survey of malware detection in Android apps: Recommendations and perspectives for future research. *Computer Science Review*, *39*, 100358. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100358

Singh, N., Buyya, R., & Kim, H. (2024). Securing Cloud-Based Internet of Things: Challenges and mitigations. *Sensors*, *25*(1), 79. https://doi.org/10.3390/s25010079

Tafur-Arciniegas, A. F., González-González, C. S., & Toledo-Delgado, P. A. (2023). Gamified strategies in fashion e-commerce: A study on user engagement and decision-making. *Proceedings of the 25th International Conference on Human-Computer Interaction with Mobile Devices and Services*. https://doi.org/10.1145/3565287.3617632

Ucci, D., Aniello, L., & Baldoni, R. (2019). Survey of machine learning techniques for malware analysis. *Computers & Security*, *81*, 123–147. https://doi.org/https://doi.org/10.1016/j.cose.2018.11.001

Vourganas, I. J., & Michala, A. L. (2024). Applications of Machine Learning in Cyber Security: A Review. *Journal of Cybersecurity and Privacy*, *4*(4), 972–992. https://doi.org/10.3390/jcp4040045

Walia, A. (2021). *Data Augmentation with Malware as Images* (Number 1009) [San Jose State University]. https://doi.org/10.31979/etd.v8ty-mhxt

Wang, L., Zhang, X., & Chen, Y. (2023). Feature selection and normalization strategies for improving malware detection using KNN and ensemble models. *Electronics*, *12*(4), 845. https://doi.org/10.3390/electronics12040845

Wong, M. D., Raff, E., Holt, J., & Netravali, R. (2022). *Marvolo: Programmatic Data Augmentation for Practical ML-Driven Malware Detection*. https://arxiv.org/abs/2206.03265

Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, 102221. https://doi.org/https://doi.org/10.1016/j.cose.2021.102221

Zhong, F., Cheng, X., Yu, D., Gong, B., Song, S., & Yu, J. (2024). MalFox: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs Against Black-Box Detectors. *IEEE Transactions on Computers*, *73*(4), 980–993. https://doi.org/10.1109/TC.2023.3236901

Zhu, H., Wei, H., Wang, L., Xu, Z., & Sheng, V. S. (2023). An effective end-to-end android malware detection method. *Expert Systems With Applications*, *218*, 119593. https://doi.org/10.1016/j.eswa.2023.119593

Apéndices

Apéndice A

Artículos seleccionados

Anexos - Monografia

Nota: Recopilación de los artículos seleccionados para el desarrollo de la investigación.

Apéndice B

Matriz Revisión Bibliográfica

Anexos - Monografia

Nota: Es la recopilación el análisis realizado a cada uno de los artículos para el desarrollo de la investigación.

Apéndice C

Dataset de análisis de malware

Anexos - Monografia

Nota: Recopilación de los registros utilizados durante el análisis de seguridad.

Apéndice D

Diseño de un Pipeline Predictivo para la Detección de Malware Basado en Machine Learning

Anexos - Monografia

Nota: Es el pipeline propuesto representa un flujo integral, modular y escalable para la construcción de un sistema de detección de malware basado en Machine Learning.