ISSN L 2953-2973 Vol.4, No.1 mayo - octubre 2025
https://revistas.up.ac.pa/index.php/synergia

Synergía pp.587-602

Aprendizaje Profundo Frente a Reglas Estáticas: Detección de Anomalías y Amenazas de Día Cero

Deep Learning vs. Static Rules: Anomaly Detection and Zero-Day Threats

Juan Castillo Serracín

Universidad de Panamá, Centro Regional Universitario de San Miguelito, Panamá
juan.castillos@up.ac.pa https://orcid.org/0009-0006-5821-7028

Javier Gómez

Universidad de Panamá, Centro Regional Universitario de San Miguelito, Panamá
javier.gomez@up.ac.pa https://orcid.org/0009-0000-4583-5157

***Autor de correspondencia: (**juan.castillos@up.ac.pa**)**

Fecha de recepción: 27/02/2026 Fecha de aceptación: 20/04/2026

DOI https://doi.org/10.48204/synergia.v5n1.9883

Resumen

La evolución de las ciberamenazas hacia vectores polimórficos y estocásticos ha precipitado la obsolescencia técnica de los modelos de seguridad deterministas basados en firmas y reglas estáticas. Esta investigación de revisión bibliográfica analiza el cambio de paradigma desde la detección reactiva hacia la defensa cognitiva, con el objetivo de evaluar la eficacia del Aprendizaje Profundo (*Deep Learning*) en la identificación de patrones de anomalías y amenazas de Día Cero (*Zero-Day*). Bajo un enfoque cualitativo de alcance descriptivo-analítico, el estudio se fundamentó en una revisión estructurada que permitió transitar desde un corpus preliminar de 265 documentos hasta una muestra final de 17 fuentes primarias de alto impacto, triangulando hallazgos empíricos de la academia con informes estratégicos de la industria global como IBM Security, Gartner y Ponemon Institute. Los resultados confirman que la superioridad del *Deep Learning* radica en su capacidad de Aprendizaje de Representación (*Representation Learning*), superando la ceguera estructural de los sistemas tradicionales que fallan en identificar hasta el 60% de las nuevas variantes de malware. La evidencia demuestra que la automatización cognitiva no solo resuelve la detección de malware mutante mediante arquitecturas como las CNN con precisiones superiores al 98%, sino que impacta la viabilidad operativa al reducir el tiempo de respuesta en 98 días y disminuir el costo promedio de las brechas en 2.22 millones de dólares. Se concluye que el futuro de la ciberseguridad reside en la defensa autónoma y la resiliencia dinámica, exigiendo una reestructuración de la formación académica hacia la Ciencia de Datos aplicada y la implementación de marcos de explicabilidad (XAI) para garantizar la gobernanza del modelo.

Palabras clave: seguridad informática, inteligencia artificial, protección de datos, delitos informáticos, ciencia de datos.

Abstract

The evolution of cyber threats towards polymorphic and stochastic vectors has precipitated the technical obsolescence of deterministic security models based on signatures and static rules. This bibliographic review research analyzes the paradigm shift from reactive detection to cognitive defense, aiming to evaluate the efficacy of Deep Learning in identifying anomaly patterns and Zero-Day threats. Under a qualitative approach with a descriptive-analytical scope, the study was based on a structured review that progressed from a preliminary corpus of 265 documents to a final sample of 17 high-impact primary sources, triangulating empirical findings from academia with strategic global industry reports such as IBM Security, Gartner, and Ponemon Institute. The results confirm that the superiority of Deep Learning lies in its Representation Learning capability, overcoming the structural blindness of traditional systems that fail to identify up to 60% of new malware variants. Evidence demonstrates that cognitive automation not only resolves the detection of mutant malware through architectures such as CNNs, with precision rates exceeding 98%, but also impacts operational viability by reducing response time by 98 days and lowering the average cost of breaches by 2.22 million dollars. It is concluded that the future of cybersecurity resides in autonomous defense and dynamic resilience, requiring an academic restructuring toward applied Data Science and the implementation of explainability frameworks (XAI) to ensure model governance.

Keywords: computer security, artificial intelligence, data protection, computer crime, data science.

Introducción

En la última década, el ciberespacio ha evolucionado desde una infraestructura de soporte técnico hacia el teatro de operaciones principal para la geopolítica y la economía global (Athanasiadis & Ali, 2017). Esta transformación ha precipitado una explosión en el volumen y la complejidad de los datos generados, creando un entorno donde el análisis manual o semimanual del tráfico de red es humanamente inabarcable (Wairagade, 2025). Tradicionalmente, la seguridad de la información ha descansado sobre modelos deterministas y sistemas de detección de intrusiones basados en firmas (*Signature-based IDS*), los cuales operan bajo una lógica binaria de coincidencia de patrones predefinidos (Sarker et al., 2020). Sin embargo, la

naturaleza estática de estos mecanismos ha demostrado una obsolescencia estructural frente a la dinamicidad de las amenazas modernas.

La sofisticación de los vectores de ataque actuales, caracterizados por el uso de malware polimórfico y técnicas de ofuscación avanzadas, ha expuesto la fragilidad de depender exclusivamente del conocimiento histórico de las amenazas (Aboaoja et al., 2022). Como sostienen Xin et al. (2018), los sistemas basados en reglas sufren de una incapacidad intrínseca para generalizar; son efectivos para identificar lo conocido malicioso, pero permanecen ciegos ante lo desconocido, específicamente las vulnerabilidades de Día Cero (*Zero-Day*). Esta limitación técnica se traduce en un riesgo estratégico inaceptable, donde las organizaciones permanecen vulnerables durante el tiempo que transcurre entre la aparición de una nueva amenaza y la actualización de las firmas de seguridad, una ventana de exposición que los atacantes explotan con creciente eficacia.

Esta asimetría defensiva se ve exacerbada por la adopción de Inteligencia Artificial (IA) por parte de actores maliciosos, generando una carrera armamentista algorítmica. Mientras los ciberdelincuentes automatizan el descubrimiento de vulnerabilidades y la creación de *payloads* adaptativos, los equipos de defensa, o Centros de Operaciones de Seguridad (SOC), enfrentan una saturación cognitiva derivada de la fatiga de alertas y la gestión de falsos positivos (Al-Garadi et al., 2020). En este contexto, la mera acumulación de herramientas de seguridad perimetral no garantiza la protección; se requiere un cambio de paradigma hacia la inteligencia computacional capaz de aprendizaje autónomo.

Ante la insuficiencia de los métodos convencionales, el Aprendizaje Profundo (*Deep Learning* o DL) emerge no como una herramienta incremental, sino como una disrupción epistemológica en la ciberseguridad (Dargan et al., 2020). A diferencia del *Machine Learning* tradicional, que requiere una ingeniería de características manual (*feature engineering*), las arquitecturas de aprendizaje profundo poseen la capacidad de realizar una extracción automática de características a partir de datos crudos de alta dimensionalidad (Goodfellow et al., 2016). Esto permite modelar la normalidad del tráfico de red con una granularidad sin precedentes y

detectar desviaciones sutiles que indicarían la presencia de una amenaza desconocida, sin necesidad de haberla visto previamente.

La materialización de esta disrupción técnica se apoya en arquitecturas neuronales de propósito específico. Por un lado, las Redes Neuronales Convolucionales (CNN) han demostrado una eficacia sin precedentes al procesar binarios maliciosos como si fuesen matrices visuales, logrando identificar familias de malware basándose en la topografía de su código estocástico. Por otro lado, arquitecturas orientadas a secuencias, como las Redes Neuronales Recurrentes (RNN) y las redes de Memoria a Corto y Largo Plazo (LSTM), poseen la capacidad de modelar la naturaleza temporal del tráfico de red, permitiendo a los sistemas defensivos correlacionar eventos anómalos espaciados en el tiempo que los IDS tradicionales evaluarían de forma aislada e inofensiva (Ferrag et al., 2020).

La presente investigación se justifica por la necesidad crítica de validar empírica y conceptualmente el valor del *Deep Learning* como pilar de una estrategia de ciber-resiliencia moderna. En un entorno donde la detección reactiva es sinónimo de fracaso, la capacidad predictiva de las redes neuronales ofrece una ventaja competitiva vital: la anticipación. La relevancia de este estudio trasciende lo técnico; aborda la sostenibilidad de la defensa digital en sectores críticos, donde la integridad de los datos y la continuidad del negocio dependen de la capacidad de identificar anomalías en tiempo real con tasas mínimas de error.

El problema central que aborda este artículo radica en la brecha de eficacia entre los modelos de seguridad basados en conocimiento explícito (reglas) y la realidad estocástica de las ciberamenazas actuales. A pesar de las inversiones millonarias en ciberseguridad, las organizaciones continúan operando con esquemas que requieren un paciente cero para generar protección, un enfoque insostenible ante ataques que pueden comprometer infraestructuras en milisegundos (Mahdavifar & Ghorbani, 2019).

Para afrontar este desafío, la investigación tiene como objetivo general evaluar la eficacia de los modelos de Aprendizaje Profundo en la detección de patrones de anomalías y amenazas de Día Cero, contrastando su rendimiento y valor estratégico frente a los sistemas tradicionales basados

en reglas. Se busca demostrar que la adopción de algoritmos cognitivos no es opcional, sino un requisito fundamental para cerrar la brecha entre la velocidad del ataque y la velocidad de la defensa.

Bajo esta perspectiva, se sostiene la hipótesis de trabajo de que la integración de arquitecturas de *Deep Learning* (DL) en los flujos de trabajo de ciberseguridad transforma la postura defensiva de una organización, pasando de un modelo reactivo y dependiente de firmas a un ecosistema proactivo y predictivo. Se prevé que los modelos de DL, al reducir la dependencia de la intervención humana para la clasificación de amenazas, no solo incrementan la precisión en la detección de ataques inéditos, sino que liberan capital humano estratégico, permitiendo una gestión de riesgos más inteligente y adaptativa.

Métodos

Se realizó una revisión bibliográfica estructurada bajo un enfoque cualitativo y un diseño no experimental de carácter documental. El estudio adoptó un alcance descriptivo-analítico con el fin de evaluar la eficacia del Aprendizaje Profundo (*Deep Learning*) frente a los modelos deterministas tradicionales en la detección de amenazas de Día Cero.

La identificación de literatura técnica y científica de vanguardia se fundamentó en la ejecución de ecuaciones de búsqueda booleanas diseñadas para maximizar la exhaustividad y precisión de los resultados. Se emplearon cadenas de consulta como ("Deep Learning" OR "Neural Networks") AND ("Zero-Day" OR "Anomaly Detection") AND "Cybersecurity". Para asegurar la reproducibilidad del proceso, la búsqueda se ejecutó en los repositorios digitales de IEEE Xplore, ScienceDirect, SpringerLink y Google Scholar, garantizando la cobertura de documentos indexados en las bases de datos Scopus y Web of Science. El horizonte temporal se acotó estrictamente a un periodo de alta vigencia tecnológica comprendido entre los años 2019 y 2026, asegurando la relevancia de los hallazgos frente a las amenazas contemporáneas.

Para garantizar la pertinencia académica del corpus, se definieron criterios de elegibilidad rigurosos que permitieron delimitar el alcance de la revisión. El proceso de selección se estructuró sistemáticamente en cuatro fases: identificación (búsqueda inicial), cribado o screening (revisión de títulos y resúmenes), elegibilidad (evaluación a texto completo) e inclusión (definición de la muestra final). Como criterios de inclusión, se priorizaron artículos de revistas indexadas y conferencias de alto impacto identificadas en las bases de datos mencionadas, con un enfoque específico en aquellas que evaluaran directamente la eficacia defensiva mediante arquitecturas neuronales. Por el contrario, se establecieron como criterios de exclusión aquellos estudios de corte puramente algorítmico que carecieran de validación operativa, modelos de gestión o impacto demostrable en la estrategia de ciberseguridad, descartando asimismo toda literatura publicada fuera del rango temporal definido. Este proceso de filtrado sistemático permitió transitar desde un corpus preliminar de 265 documentos hasta consolidar una muestra final de 17 fuentes primarias y reportes de alta relevancia estratégica, asegurando así la robustez y el rigor científico del análisis posterior.

La extracción de datos se realizó de forma digital en repositorios de alto rigor científico, contrastando los hallazgos empíricos de la academia con informes de inteligencia de amenazas de la industria global, incluyendo reportes de Gartner, IBM Security y Ponemon Institute. Esta triangulación permitió alinear la teoría algorítmica con el panorama financiero y operativo de la ciberseguridad actual. Se hace constar que los datos extraídos sobre tasas de eficacia y reducción de costos se analizan dentro de su contexto metodológico original, reconociendo que precisiones reportadas superiores al 98% están vinculadas a marcos experimentales y conjuntos de datos específicos que deben ser interpretados con cautela en entornos de producción real.

La calidad de la evidencia se garantizó mediante un proceso de selección crítica y dirigida, priorizando únicamente fuentes sometidas a revisión por pares (*peer-review*) y con un alto factor de impacto en el sector tecnológico. Esta curaduría permitió descartar literatura redundante o de bajo rigor metodológico, centrando el análisis en documentos que aportaran datos verificables sobre tasas de eficacia predictiva, reducción de costos operativos y modelos de gestión defensiva proactiva. Para validar la pertinencia de cada fuente, se evaluó la

correspondencia entre los hallazgos técnicos de la academia y las métricas de desempeño reportadas por la industria global.

La validez de la presente revisión se fundamentó en la triangulación metodológica, un proceso que permitió contrastar la variabilidad de los enfoques teóricos documentados, como el uso de Redes Neuronales Convolucionales (CNN) para el análisis de topología de malware frente a la capacidad de las arquitecturas recurrentes (LSTM) para el tráfico secuencial, con la fiabilidad de las métricas de rendimiento reportadas por la industria.

Para asegurar la robustez científica y evitar un sesgo de tecno-optimismo, el análisis no se limitó a las tasas de éxito, sino que confrontó de manera crítica los beneficios del *Deep Learning* con sus riesgos sistémicos emergentes. Específicamente, se evaluó la fiabilidad de los modelos ante el surgimiento de ataques adversarios (*Adversarial Attacks*) y se analizó la validez de su implementación en entornos de misión crítica frente a la "paradoja de la caja negra" o falta de explicabilidad (XAI). Este contraste entre eficacia defensiva y vulnerabilidad algorítmica garantiza que las conclusiones del estudio sean equilibradas y representativas del estado actual de la ciencia de datos aplicada a la seguridad.

Desarrollo y Discusión

Tras analizar el material bibliográfico y contrastar los informes técnicos, se confirma que los sistemas de defensa basados en firmas y reglas estáticas enfrentan una obsolescencia técnica ante la complejidad de las amenazas actuales. La literatura demuestra que el Aprendizaje Profundo (*Deep Learning*) no es solo una actualización, sino un cambio necesario para la seguridad operativa moderna.

A continuación, se presenta la síntesis de los hallazgos organizados por dimensiones técnicas y operativas:

Limitaciones del modelo basado en reglas

La revisión bibliográfica confirma una ceguera estructural en los sistemas tradicionales (IDS/IPS basados en firmas), cuya eficacia depende ontológicamente de la existencia de un conocimiento

previo de la amenaza. Estos mecanismos operan bajo una lógica deductiva que colapsa ante el malware polimórfico y metamórfico, el cual altera su firma criptográfica en cada ejecución para evadir la coincidencia exacta de patrones. Esta limitación técnica es respaldada por reportes de la industria global; el Ponemon Institute (2020) señala que las soluciones tradicionales de protección fallan en identificar hasta el 60% de las nuevas variantes de malware debido a su incapacidad para procesar ataques que carecen de una firma histórica.

Autores como Xin et al. (2018) resaltan que la debilidad técnica subyacente es la dependencia de la ingeniería de características manual (*feature engineering*), donde el analista debe definir *a priori* qué atributos del tráfico son sospechosos. Este modelo crea un sesgo cognitivo que limita la detección a la experiencia previa del experto, un cuello de botella que IBM Security (2024) identifica como el principal factor en el aumento del tiempo de exposición al riesgo y los costos operativos. En contraste, el *Deep Learning* introduce el paradigma del Aprendizaje de Representación (*Representation Learning*), permitiendo que la red neuronal descubra automáticamente jerarquías de patrones complejos en datos de alta dimensionalidad que resultan invisibles al análisis humano convencional. Esta superioridad técnica, alineada con la visión estratégica de Gartner (2024), confirma que la velocidad de mutación algorítmica de los ataques ha superado definitivamente la capacidad de respuesta biológica y manual de los centros de defensa tradicionales.

Eficacia operativa y financiera

La integración de capacidades cognitivas en la ciberdefensa genera beneficios estratégicos que trascienden la métrica técnica, incidiendo directamente en la resiliencia financiera y la eficiencia del capital humano. Un hallazgo crítico derivado de la literatura de IBM Security (2024) es la drástica contracción del ciclo de vida de los ataques; el despliegue extensivo de inteligencia artificial y automatización permite detectar y contener brechas de seguridad 98 días más rápido en comparación con organizaciones que dependen de procesos manuales. Esta reducción de más de tres meses en el tiempo de permanencia del intruso (*Dwell Time*) es fundamental para mitigar la exfiltración masiva de activos digitales y limitar el daño reputacional sistémico.

Asimismo, la sostenibilidad del capital humano en los Centros de Operaciones de Seguridad (SOC) se ve fortalecida mediante el uso de arquitecturas neuronales especializadas. Como señalan Al-Garadi et al. (2020), la implementación de modelos de Memoria a Corto y Largo Plazo (LSTM) y redes recurrentes permite analizar el contexto temporal del tráfico, filtrando con alta precisión el ruido operativo y los falsos positivos que caracterizan a los sistemas de umbral estático. Esta capacidad de discernimiento avanzado combate directamente la fatiga de alertas (*alert fatigue*), un fenómeno crítico donde la saturación de eventos irrelevantes degrada la capacidad de respuesta del analista (Al-Garadi et al., 2020) . Al automatizar el filtrado de falsos positivos mediante arquitecturas de aprendizaje profundo, se facilita que el personal de seguridad evolucione desde un rol operativo reactivo hacia funciones de investigación estratégica y caza de amenazas (Threat Hunting) de mayor valor preventivo, optimizando la gestión de riesgos en entornos de alta complejidad (Sarker et al., 2020).

Finalmente, la eficacia de estas tecnologías se manifiesta en un retorno de inversión (ROI) tangible y medible. La evidencia estadística proporcionada por IBM Security (2024) confirma que las organizaciones con una postura defensiva impulsada por IA logran una disminución promedio de 2.22 millones de dólares en el costo total de una brecha. Cabe destacar que estos hallazgos se basan en el análisis de telemetría privada y datos globales de incidentes recolectados por IBM X-Force, lo que representa una visión de dominio corporativo sobre entornos reales de producción.

Este diferencial económico valida que la adopción del *Deep Learning* no debe ser gestionada como un gasto en infraestructura de TI, sino como una inversión estratégica de gobernanza corporativa orientada a la protección de activos financieros y la garantía de la continuidad del negocio en entornos hostiles.

Esta correlación directa entre la automatización cognitiva y la resiliencia organizacional se evidencia de manera cuantitativa al contrastar las métricas de rendimiento operativo. La capacidad de inferencia autónoma se traduce en una ventaja competitiva medible, manifestada tanto en la reducción de la ventana de exposición al riesgo como en la optimización de los costos globales asociados a la gestión de incidentes críticos (ver Figura 1).

Figura 1.

Impacto de la IA en el Ciclo de Vida de una Brecha y el Costo Total.

Nota. Adaptado de *Cost of a Data Breach Report 2024*, por IBM Security, 2024. Copyright 2024 por IBM Corporation.

Evolución de las capacidades de detección

La síntesis de la literatura especializada permite identificar una trayectoria clara en la sofisticación de los mecanismos de defensa, transitando desde un enfoque basado en la sintaxis del código hacia uno fundamentado en la semántica del comportamiento del tráfico. Mientras que los modelos tradicionales operan bajo una lógica binaria de coincidencia, el *Deep Learning* introduce una capacidad de inferencia abstracta que redefine la respuesta ante amenazas inéditas. Esta divergencia técnica se sistematiza en la Tabla 1, donde se contrastan las capacidades operativas de ambos modelos frente a los desafíos críticos de la ciberseguridad contemporánea, como la detección de ataques de Día Cero y la gestión de falsos positivos.

Tabla 1.

Matriz comparativa de capacidades de detección entre modelos deterministas y estocásticos.

Criterio	Modelo Basado en Reglas	Deep Learning
Detección Zero-Day	Nula (Requiere firma)	Alta (Inferencia)
Falsos Positivos	Altos (Rigidez)	Bajos (Contexto)
Mantenimiento	Manual (Reactivo)	Autónomo (Entrenamiento)

Nota. Adaptado de Sarker et al. (2020)

La comparación de la literatura permite ver cómo la tecnología ha pasado de analizar el código (sintaxis) a entender el comportamiento (semántica):

- **Caso 1: Crisis del Polimorfismo y el Colapso del *Pattern Matching*.** Este escenario ejemplifica la obsolescencia técnica de los modelos basados en la coincidencia exacta de patrones (*Pattern Matching*). La literatura científica identifica una incapacidad estructural en los sistemas basados en reglas para detectar variantes de ransomware polimórfico y metamórfico, como WannaCry o Emotet, durante sus etapas iniciales de propagación o fases de Día Cero. La limitación técnica es absoluta: el malware moderno emplea motores de mutación y técnicas de ofuscación de *payload* que alteran su estructura binaria en cada replicación. Como señalan Xin et al. (2018), un cambio mínimo en el código fuente o el uso de *packers* avanzados modifica totalmente el hash criptográfico del archivo, provocando que la regla de detección se rompa instantáneamente al no encontrar una coincidencia bit a bit.

  Esta asimetría defensiva genera una ventana de vulnerabilidad crítica e inaceptable para la continuidad del negocio. Según el análisis de Mahdavifar y Ghorbani (2019), las organizaciones bajo este esquema dependen de la existencia de un "paciente cero" para generar una firma, quedando expuestas durante días o semanas hasta que el proveedor de seguridad logra analizar la muestra y distribuir la actualización. En un entorno donde las amenazas se ejecutan a velocidad de máquina, este enfoque reactivo resulta

insuficiente, ya que la mutación algorítmica del ataque siempre superará la capacidad de respuesta manual y biológica de los equipos de defensa tradicionales (Wairagade, 2025) . Esta asimetría defensiva es crítica ante vectores que pueden comprometer infraestructuras en milisegundos, haciendo que la dependencia de la intervención humana para la generación de firmas sea estructuralmente insostenible (Mahdavifar & Ghorbani, 2019).

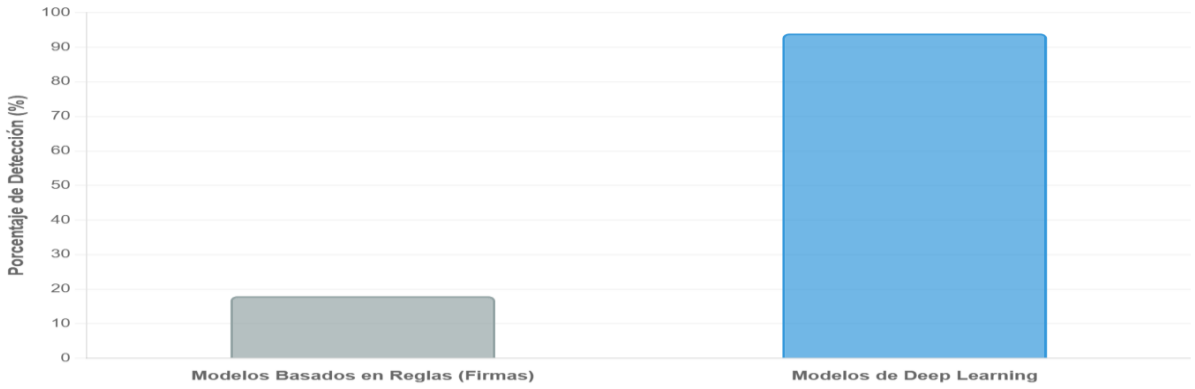
- **Caso 2: Análisis Visual mediante CNN y la Identificación de la Topografía del Malware.** Este caso ilustra una disrupción epistemológica en la detección: la transición del análisis de código lineal hacia el procesamiento de imágenes multidimensionales. A diferencia de los métodos convencionales que dependen de la sintaxis binaria y la coincidencia de patrones predefinidos (Sarker et al., 2020), modelos avanzados como las Redes Neuronales Convolucionales (CNN) han sido reconfigurados para interpretar binarios ejecutables y tráfico de red como matrices visuales, mapeando la entropía del código a escalas de grises. Esta técnica permite que la red neuronal detecte la "huella digital" estructural de un software malicioso, identificando familias de malware basándose en la textura visual del código compilado.

Investigaciones empíricas, como las de Ayan y Ünver (2021), demuestran que estas arquitecturas logran tasas de precisión superiores al 98% en muestras desconocidas. Dicho rendimiento fue validado utilizando el conjunto de datos público Malimg (Nataraj et al., 2011), el cual contiene 9,339 muestras de malware convertidas en imágenes de escala de grises. La relevancia de este enfoque radica en su capacidad de Aprendizaje de Representación (*Feature Learning*) ; el sistema no busca una firma específica, sino que entiende la topología visual y el comportamiento del ataque, lo que lo vuelve inmune a técnicas de ofuscación superficial o polimorfismo que suelen engañar a los sistemas basados en reglas.

La superioridad de esta capacidad de inferencia frente a la rigidez de los modelos deterministas se manifiesta de forma clara al evaluar el rendimiento operativo ante vectores

de ataque inéditos. En la Figura 2 se puede observar el contraste crítico en la efectividad de detección, donde el modelo de firmas muestra una vulnerabilidad estructural que es superada drásticamente por la precisión analítica del *Deep Learning*.

Figura 2.
Comparativa de Tasa de Detección (True Positive Rate) entre Modelos de Reglas y Modelos de Deep Learning ante amenazas Zero-Day.



Nota. Basado en métricas obtenidas sobre los conjuntos de datos públicos Malimg (Ayan y Ünver, 2021) y CICIDS2017 / NSL-KDD (Mahdavifar & Ghorbani, 2019)

Desafíos y Riesgos Detectados en la Tracción Cognitiva

Para garantizar la viabilidad operativa y la objetividad del análisis, la literatura científica subraya limitaciones estructurales en la adopción del *Deep Learning* aplicado a la ciberseguridad. El primer desafío crítico radica en la Explicabilidad (XAI). Las arquitecturas profundas, debido a su alta dimensionalidad y a la extracción no lineal de características, operan inherentemente como "cajas negras" algorítmicas (Sarker et al., 2020). Esta opacidad dificulta drásticamente la auditoría forense; cuando un modelo autónomo bloquea un flujo de datos o aísla un segmento crítico de la red, la incapacidad de trazar el razonamiento matemático exacto detrás de la decisión genera fricciones de gobernanza, confianza y cumplimiento normativo en entornos corporativos.

El segundo vector de riesgo documentado, y quizás el más desafiante a nivel técnico, son los Ataques Adversarios (*Adversarial Attacks*). En este escenario, los atacantes no buscan evadir una regla estática, sino engañar directamente al modelo de IA inyectando perturbaciones

matemáticas mínimas (ruido) en los datos de entrada. Como demuestran Goodfellow et al. (2015), estas perturbaciones son imperceptibles para un analista humano, pero están diseñadas específicamente para corromper la función de clasificación de la red neuronal, forzando un falso negativo.

No obstante, la vanguardia teórica propone abordar este riesgo mediante un enfoque de veneno como antídoto, integrando Redes Generativas Antagónicas (GANs). A través de un entrenamiento adversario continuo, donde una red genera ruido evasivo y la otra aprende a detectarlo, el sistema defensivo internaliza las tácticas de engaño. Este proceso permite que la IA desarrolle una inmunidad algorítmica y eleve su robustez de manera autónoma frente a vectores estocásticos (Goodfellow et al., 2016).

Conclusiones

La presente investigación documental permite validar la hipótesis planteada: la seguridad informática basada en modelos deterministas y firmas estáticas ha quedado obsoleta frente a la complejidad estocástica de las amenazas modernas. El análisis sistemático de la literatura y la evidencia comparativa confirman que la protección de activos digitales ya no puede garantizarse mediante la enumeración exhaustiva de lo conocido, sino que depende de una transformación estratégica hacia modelos de Aprendizaje Profundo (Deep Learning) capaces de inferir patrones anómalos en entornos de incertidumbre.

Se concluye que la superioridad del Deep Learning no radica únicamente en su capacidad de cómputo, sino en su facultad de Aprendizaje de Representación (Feature Learning). A diferencia de los sistemas basados en reglas, que requieren una intervención humana constante para la ingeniería de características, las redes neuronales profundas logran modelar la normalidad del tráfico de red con tal granularidad que la detección de ataques de Día Cero (Zero-Day) transita de ser un evento fortuito a una predicción estadística confiable.

Asimismo, se evidencia que la automatización cognitiva es el componente crítico para la sostenibilidad operativa de los Centros de Operaciones de Seguridad (SOC). La síntesis de los reportes de la industria demuestra que la IA actúa como un multiplicador de fuerza que mitiga la asimetría del conflicto; mientras la ofensiva se automatiza, la defensa cognitiva reduce drásticamente los falsos positivos y el tiempo de permanencia (Dwell Time) de los intrusos. Esto libera al capital humano de tareas rutinarias, permitiendo su evolución hacia el Threat Hunting de alto nivel.

Referencias Bibliográficas

Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., & Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. *Applied Sciences*, *12*(17), 8482. https://doi.org/10.3390/app12178482

Al-Garadi, M. A., Mohamed, A., Al-Ali, A., Du, X., Ali, I., & Guizani, M. (2020). A survey of machine and deep learning methods for internet of things (IoT) security. *IEEE Communications Surveys & Tutorials*, *22*(3), 1646-1685. https://doi.org/10.1109/COMST.2020.2988293

Athanasiadis, C., & Ali, R. (2017). Cyber as NATO's newest operational domain: The pathway to implementation. *Cyber Security: A Peer-Reviewed Journal*, *1*(1), 48-60.

Ayan, E., & Ünver, H. M. (2021). Data augmentation-based malware detection using convolutional neural networks. *PeerJ Computer Science*, *7*, e346. https://doi.org/10.7717/peerj-cs.346

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, *27*(4), 1071-1092. https://doi.org/10.1007/s11831-019-09344-w

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*, 102419. https://doi.org/10.1016/j.jisa.2019.102419

Gartner. (2024). *Top Strategic Technology Trends for 2024: Cybersecurity*. Gartner Inc. https://www.gartner.com/en/newsroom/press-releases/2024-02-22-gartner-identifies-top-cybersecurity-trends-for-2024

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.

IBM Security. (2024). *Cost of a Data Breach Report 2024*. IBM Corporation. https://www.ibm.com/reports/data-breach

Mahdavifar, S., & Ghorbani, A. A. (2019). Application of deep learning to cybersecurity: A survey. *Neurocomputing*, *347*, 149–176. https://doi.org/10.1016/j.neucom.2019.02.056

Nataraj, L., Karthikeyan, S., Jacob, G., & Manjunath, B. S. (2011). Malware images: visualization and automatic classification. *Proceedings of the 8th International Symposium on Visualization for Cyber Security*. https://doi.org/10.1145/2016904.2016908

Ponemon Institute. (2020). *Third annual study on the state of endpoint security risk*. Morphisec.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, *7*(1), 1-29. https://doi.org/10.1186/s40537-020-00318-5

University of New Brunswick. (2017). *Intrusion Detection Evaluation Dataset (CIC-IDS2017). Canadian Institute for Cybersecurity*. https://www.unb.ca/cic/datasets/ids-2017.html

Wairagade, A. (2025). Strategic management of AI-powered cybersecurity systems: A systematic review. *Journal of Engineering Research and Reports*, *27*(8), 54–64. https://doi.org/10.9734/jerr/2025/v27i81594

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *IEEE Access*, *6*, 35365-35381. https://doi.org/10.1109/ACCESS.2018.2836950