Artículo Científico

ISSN: XXXX- XXXX

Ciberseguridad predictiva basada en inteligencia artificial contra ataques generativos

Predictive cybersecurity based on artificial intelligence against generative attacks

Gruezo-Realpe, Mariela Stephany [1]
https://orcid.org/0000-0002-5929-4336
mariela.gruezo.realpe@utelvt.edu.ec
Ecuador, Esmeraldas, Universidad Técnica Luis Vargas Torres de Esmeraldas.

Torres-Galves, Génesis Daniela [2]
https://orcid.org/0009-0004-7198-7057
gdtorresg@ube.edu.ec
Ecuador, Durán, Universidad Bolivariana del Ecuador.

Lascano-Rivera, Samuel Benjamín [3]
https://orcid.org/0000-0001-5967-6441
samuel.lascano@upec.edu.ec
Perú, Lima, Universidad Nacional Mayor de San Marcos.

Autor de correspondencia [1]

DOI / URL: https://doi.org/10.55813/gaea/revistacec/v3/n2/2

Resumen: La investigación analiza la ciberseguridad predictiva basada en inteligencia artificial como respuesta al incremento de ataques generativos capaces de producir phishing hiperpersonalizado, deepfakes, identidades sintéticas, malware adaptable y estrategias de evasión más sofisticadas. El objetivo fue examinar la literatura reciente sobre avances, riesgos, limitaciones y criterios de implementación defensiva frente a este tipo de amenazas. Metodológicamente, se desarrolló una revisión bibliográfica sustentada en artículos científicos, informes técnicos y marcos especializados, sin recolección de datos personales ni ejecución de pruebas ofensivas. Los resultados evidencian que la inteligencia artificial fortalece la detección anticipada al correlacionar señales débiles, reconocer anomalías multifuente y superar parcialmente la dependencia de firmas estáticas; sin embargo, también presenta limitaciones asociadas con sesgos de datos, falsos positivos, opacidad algorítmica, manipulación adversarial, inyección de instrucciones y dependencia excesiva de la automatización. Se concluye que la defensa predictiva debe integrarse en arquitecturas de seguridad por capas, con gobernanza de datos, explicabilidad operativa, evaluación adversarial permanente, supervisión humana y protocolos institucionales de respuesta, a fin de transitar desde una ciberseguridad reactiva hacia un modelo anticipatorio, resiliente y responsable.

Palabras clave: ciberseguridad predictiva; inteligencia artificial; ataques generativos; aprendizaje adversarial; detección de amenazas.

Check for updates

Received: 27/Feb/2026
Accepted: 22/Mar/2026
Published: 20/Abr/2026

Cita: Gruezo-Realpe, M. S., Torres-Galves, G. D., & Lascano-Rivera, S. B. (2026). Ciberseguridad predictiva basada en inteligencia artificial contra ataques generativos. *Revista Científica Enfoques Del Conocimiento*, *3*(2), 16-28. https://doi.org/10.55813/gaea/revistacec/v3/n2/2

Revista Científica Enfoques del Conocimiento (RCEC)
https://www.blez.edu.ec
https://revistacec.blez.edu.ec
revistacec@blez.edu.ec

Abstract:

The study analyzes predictive cybersecurity based on artificial intelligence as a response to the growing emergence of generative attacks capable of producing hyper-personalized phishing, deepfakes, synthetic identities, adaptive malware, and increasingly sophisticated evasion strategies. The objective was to examine recent literature regarding advances, risks, limitations, and defensive implementation criteria against these threats. Methodologically, a bibliographic review was conducted using scientific articles, technical reports, and specialized frameworks, without collecting personal data or performing offensive testing. The findings reveal that artificial intelligence strengthens early threat detection by correlating weak signals, identifying multi-source anomalies, and partially overcoming the limitations of static signature-based defenses. However, it also presents challenges related to data bias, false positives, algorithmic opacity, adversarial manipulation, prompt injection, and excessive reliance on automation. The study concludes that predictive defense should be integrated into layered security architectures supported by data governance, operational explainability, continuous adversarial evaluation, human supervision, and institutional response protocols in order to transition from reactive cybersecurity toward an anticipatory, resilient, and responsible model.

Keywords: predictive cybersecurity; artificial intelligence; generative attacks; adversarial learning; threat detection.

1. Introducción

La ciberseguridad predictiva basada en inteligencia artificial surge como respuesta a un entorno digital donde los ataques ya no dependen solo de firmas conocidas, sino de patrones cambiantes, automatización ofensiva y explotación acelerada de vulnerabilidades (Tirira-Chulde et al., 2026). En este escenario, la inteligencia artificial permite anticipar comportamientos anómalos, priorizar alertas y apoyar decisiones preventivas; sin embargo, la inteligencia artificial generativa también amplía la capacidad de los atacantes para producir señuelos, código, identidades sintéticas y campañas de ingeniería social más creíbles (Kaur et al., 2023; NCSC, 2024; World Economic Forum, 2025).

El problema central radica en que los mecanismos tradicionales de defensa, basados en reglas estáticas, listas negras o detección reactiva, resultan insuficientes frente a ataques generativos capaces de variar contenido, lenguaje, infraestructura y vectores de entrada (Boné-Andrade, 2023). La evidencia reciente muestra que los modelos generativos pueden fortalecer el phishing, el reconocimiento de objetivos, la producción de documentos señuelo, la automatización de interacciones maliciosas y la evasión de controles, lo cual incrementa la dificultad de distinguir comunicaciones legítimas de intentos de intrusión (Jabir et al., 2025; NCSC, 2024; OWASP, 2025).

A partir de ello, las afectaciones del problema se expresan en dimensiones técnicas, económicas, institucionales y sociales. Técnicamente, los ataques generativos pueden comprometer sistemas mediante inyección de prompts, envenenamiento de datos, manejo inseguro de salidas, filtración de información sensible y manipulación de cadenas de suministro de modelos; institucionalmente, erosionan la confianza en servicios digitales, decisiones automatizadas y canales de comunicación corporativa (Galarza-Sánchez, 2023). Socialmente, la creación de deepfakes y mensajes hiperpersonalizados intensifica la suplantación, el fraude y la desinformación, afectando tanto a usuarios como a organizaciones (Mirsky & Lee, 2021; NIST, 2024a; OWASP, 2025).

No obstante, la literatura también muestra que la inteligencia artificial posee un potencial defensivo relevante cuando se aplica a detección de intrusiones, análisis de malware, inteligencia de amenazas, clasificación de phishing, respuesta automatizada y aprendizaje de comportamientos anómalos (Erazo-Luzuriaga et al., 2023). La brecha aparece porque muchas investigaciones enfatizan la precisión de modelos, pero tratan de forma fragmentada la relación entre predicción, explicabilidad, robustez adversarial y ataques generativos. Por ello, se requiere una revisión que articule capacidades defensivas, riesgos de doble uso y criterios de implementación confiable (Ferrag et al., 2025; Salem et al., 2024; Sarker et al., 2024).

En consecuencia, la justificación de este artículo se sostiene en la necesidad de organizar críticamente un campo emergente donde la innovación defensiva y la amenaza ofensiva evolucionan de manera simultánea (Montalván-Vélez et al., 2024). Su relevancia social se vincula con la protección de datos, continuidad operativa y confianza digital; su valor teórico reside en integrar enfoques de ciberseguridad predictiva, aprendizaje automático, inteligencia generativa y aprendizaje adversarial; y su utilidad metodológica consiste en ofrecer una síntesis bibliográfica que oriente futuras investigaciones, taxonomías y criterios de evaluación (Kaur et al., 2023; NIST, 2024b; World Economic Forum, 2025).

Asimismo, el estudio resulta viable porque existe un volumen creciente de literatura científica, informes técnicos y marcos de referencia producidos por organismos especializados, lo que permite desarrollar una revisión bibliográfica sin recolectar datos personales ni ejecutar pruebas ofensivas sobre sistemas reales (Castelo-Vinueza, 2025). Esta viabilidad ética y documental es especialmente pertinente en un tema donde simular ataques, automatizar explotación o replicar malware podría generar riesgos operativos; por tanto, el análisis puede concentrarse en evidencia publicada, recomendaciones de mitigación y desafíos abiertos (ENISA, 2024; NIST, 2024a; OWASP, 2025).

En este marco, el objetivo general del artículo es analizar la literatura reciente sobre ciberseguridad predictiva basada en inteligencia artificial frente a ataques generativos, identificando avances, riesgos, limitaciones y líneas de investigación. De manera específica, se propone describir los principales vectores de ataque generativo,

comparar enfoques predictivos aplicados a detección y respuesta, examinar desafíos de explicabilidad y robustez, y determinar criterios para integrar modelos defensivos en arquitecturas de seguridad adaptativas (Ferrag et al., 2025; Salem et al., 2024; Sarker et al., 2024).

La contribución esperada consiste en ofrecer una lectura integradora de un problema que no puede abordarse solo desde la eficiencia algorítmica ni únicamente desde la gestión del riesgo (Boné-Andrade et al., 2025). La originalidad del trabajo se ubica en conectar la anticipación predictiva con la seguridad de modelos generativos, la protección contra aprendizaje adversarial y la necesidad de defensas explicables, actualizables y gobernadas. Así, la revisión busca aportar una base conceptual para que academia y práctica avancen desde una ciberseguridad reactiva hacia una ciberseguridad anticipatoria y resiliente (Kaur et al., 2023; NIST, 2024b; NCSC, 2024).

2. Materiales y métodos

La ciberseguridad predictiva basada en inteligencia artificial surge como respuesta a un entorno digital donde los ataques ya no dependen solo de firmas conocidas, sino de patrones cambiantes, automatización ofensiva y explotación acelerada de vulnerabilidades. En este escenario, la inteligencia artificial permite anticipar comportamientos anómalos, priorizar alertas y apoyar decisiones preventivas; sin embargo, la inteligencia artificial generativa también amplía la capacidad de los atacantes para producir señuelos, código, identidades sintéticas y campañas de ingeniería social más creíbles.

El problema central radica en que los mecanismos tradicionales de defensa, basados en reglas estáticas, listas negras o detección reactiva, resultan insuficientes frente a ataques generativos capaces de variar contenido, lenguaje, infraestructura y vectores de entrada. La evidencia reciente muestra que los modelos generativos pueden fortalecer el phishing, el reconocimiento de objetivos, la producción de documentos señuelo, la automatización de interacciones maliciosas y la evasión de controles, lo cual incrementa la dificultad de distinguir comunicaciones legítimas de intentos de intrusión.

A partir de ello, las afectaciones del problema se expresan en dimensiones técnicas, económicas, institucionales y sociales. Técnicamente, los ataques generativos pueden comprometer sistemas mediante inyección de prompts, envenenamiento de datos, manejo inseguro de salidas, filtración de información sensible y manipulación de cadenas de suministro de modelos; institucionalmente, erosionan la confianza en servicios digitales, decisiones automatizadas y canales de comunicación corporativa. Socialmente, la creación de deepfakes y mensajes hiperpersonalizados intensifica la suplantación, el fraude y la desinformación, afectando tanto a usuarios como a organizaciones.

No obstante, la literatura también muestra que la inteligencia artificial posee un potencial defensivo relevante cuando se aplica a detección de intrusiones, análisis de malware, inteligencia de amenazas, clasificación de phishing, respuesta automatizada y aprendizaje de comportamientos anómalos. La brecha aparece porque muchas investigaciones enfatizan la precisión de modelos, pero tratan de forma fragmentada la relación entre predicción, explicabilidad, robustez adversarial y ataques generativos. Por ello, se requiere una revisión que articule capacidades defensivas, riesgos de doble uso y criterios de implementación confiable.

En consecuencia, la justificación de este artículo se sostiene en la necesidad de organizar críticamente un campo emergente donde la innovación defensiva y la amenaza ofensiva evolucionan de manera simultánea. Su relevancia social se vincula con la protección de datos, continuidad operativa y confianza digital; su valor teórico reside en integrar enfoques de ciberseguridad predictiva, aprendizaje automático, inteligencia generativa y aprendizaje adversarial; y su utilidad metodológica consiste en ofrecer una síntesis bibliográfica que oriente futuras investigaciones, taxonomías y criterios de evaluación.

Asimismo, el estudio resulta viable porque existe un volumen creciente de literatura científica, informes técnicos y marcos de referencia producidos por organismos especializados, lo que permite desarrollar una revisión bibliográfica sin recolectar datos personales ni ejecutar pruebas ofensivas sobre sistemas reales. Esta viabilidad ética y documental es especialmente pertinente en un tema donde simular ataques, automatizar explotación o replicar malware podría generar riesgos operativos; por tanto, el análisis puede concentrarse en evidencia publicada, recomendaciones de mitigación y desafíos abiertos.

En este marco, el objetivo general del artículo es analizar la literatura reciente sobre ciberseguridad predictiva basada en inteligencia artificial frente a ataques generativos, identificando avances, riesgos, limitaciones y líneas de investigación. De manera específica, se propone describir los principales vectores de ataque generativo, comparar enfoques predictivos aplicados a detección y respuesta, examinar desafíos de explicabilidad y robustez, y determinar criterios para integrar modelos defensivos en arquitecturas de seguridad adaptativas.

La contribución esperada consiste en ofrecer una lectura integradora de un problema que no puede abordarse solo desde la eficiencia algorítmica ni únicamente desde la gestión del riesgo. La originalidad del trabajo se ubica en conectar la anticipación predictiva con la seguridad de modelos generativos, la protección contra aprendizaje adversarial y la necesidad de defensas explicables, actualizables y gobernadas. Así, la revisión busca aportar una base conceptual para que academia y práctica avancen desde una ciberseguridad reactiva hacia una ciberseguridad anticipatoria y resiliente.

3. Resultados

3.1. Alcances de la ciberseguridad predictiva basada en inteligencia artificial frente a ataques generativos

3.1.1. Detección anticipada de amenazas generativas

La ciberseguridad predictiva basada en inteligencia artificial desplaza la defensa digital desde la reacción tardía hacia la anticipación de señales tempranas de ataque. Frente a amenazas generativas, este enfoque resulta crucial, porque la IA puede producir phishing hiperpersonalizado, malware adaptable, deepfakes e identidades sintéticas con alta verosimilitud. Por ello, la detección no debe limitarse a firmas conocidas, sino integrar comportamiento, contexto y anomalías multifuente (Kaur et al., 2023; Ferrag et al., 2025).

Los modelos de aprendizaje automático permiten construir líneas base dinámicas sobre usuarios, redes y sistemas, identificando desviaciones que podrían anticipar una intrusión. Esta capacidad supera parcialmente las reglas estáticas, ya que los ataques generativos modifican lenguaje, estructura y apariencia para evadir controles tradicionales. En consecuencia, la predicción se apoya en patrones probabilísticos y no solo en indicadores previamente registrados (Salem et al., 2024; NCSC, 2024).

Además, la detección anticipada exige correlacionar señales débiles que, aisladas, podrían parecer irrelevantes. Un correo generado por IA puede ser gramaticalmente impecable, pero revelar inconsistencias en dominio, temporalidad, solicitud o relación histórica entre remitente y destinatario. Así, el valor defensivo de la IA reside en articular múltiples evidencias antes de que el ataque se materialice plenamente (Mirsky & Lee, 2021; NCSC, 2024).

La amenaza generativa también incrementa la escala y velocidad de los ataques, pues permite automatizar contenidos, adaptar mensajes al perfil de la víctima y ensayar múltiples variantes maliciosas. Por esta razón, los sistemas predictivos deben identificar campañas en formación y no únicamente incidentes consumados. Esta orientación anticipatoria fortalece la resiliencia organizacional frente a amenazas cambiantes (Ferrag et al., 2025; OWASP, 2025).

3.1.2. Limitaciones y riesgos de los modelos defensivos

Aunque la inteligencia artificial ofrece ventajas defensivas relevantes, sus modelos no son infalibles. Su desempeño depende de la calidad, representatividad y actualidad de los datos de entrenamiento, por lo que pueden reproducir sesgos o vacíos informativos. Si los datos no incorporan ataques emergentes, el sistema puede generar falsos positivos excesivos o falsos negativos críticos (Salem et al., 2024; Vassilev et al., 2024).

La transparencia interna constituye un eje fundamental para fortalecer la gestión contable y financiera dentro de las organizaciones, ya que permite identificar, corregir y prevenir errores antes de que afecten la toma de decisiones. La imagen presenta los principales caminos que contribuyen a este propósito, entre ellos la detección de errores, la eliminación de duplicidades, el control de omisiones, la resolución de inconsistencias y el fortalecimiento de la disciplina organizacional. Asimismo, destaca la importancia de la trazabilidad y la integración contable como mecanismos que permiten seguir el recorrido de los datos, comprender mejor los registros financieros y asegurar una información más precisa, confiable y útil para la gestión institucional.

Figura 1

¿Qué riesgo operativo priorizar al implementar inteligencia artificial defensiva?

Nota: (Autores, 2026).

Otra limitación importante es que los modelos defensivos también pueden convertirse en objetivos de ataque. Mediante ejemplos adversariales, envenenamiento de datos o manipulación de entradas, un adversario puede inducir decisiones erróneas en sistemas aparentemente robustos. En entornos con IA generativa, estos riesgos se amplían por la inyección de prompts, fuga de información y manejo inseguro de salidas (Vassilev et al., 2024; NIST, 2024).

La opacidad de muchos modelos de aprendizaje profundo representa un desafío adicional. Una alerta puede ser estadísticamente precisa, pero poco útil si el analista no comprende por qué fue emitida ni qué variables influyeron en la decisión. Por ello, la explicabilidad es una condición operativa para priorizar incidentes, justificar acciones y reconstruir trayectorias de ataque (Sarker et al., 2024; Salem et al., 2024).

Existe el riesgo de una dependencia excesiva de la automatización defensiva. Cuando las organizaciones delegan demasiado en modelos algorítmicos, pueden debilitar capacidades humanas esenciales como el análisis forense, la interpretación contextual y la evaluación crítica del riesgo. En ataques generativos, la defensa

requiere combinar automatización, supervisión experta y protocolos organizacionales de verificación (Kaur et al., 2023; NIST, 2024).

3.1.3. Criterios para una defensa predictiva robusta

Una defensa predictiva robusta debe concebirse como una arquitectura por capas, no como una solución única basada en IA. La inteligencia artificial debe complementar controles como autenticación reforzada, segmentación, monitoreo continuo, inteligencia de amenazas y respuesta automatizada supervisada. Esta integración resulta indispensable cuando los ataques generativos combinan texto, código, voz, imagen y manipulación social (Ferrag et al., 2025; Kaur et al., 2023).

El primer criterio de robustez es la gobernanza rigurosa de los datos. Los modelos predictivos requieren información trazable, actualizada, representativa y protegida contra manipulación, pues un conjunto contaminado puede degradar la defensa. Por tanto, la seguridad comienza antes del entrenamiento algorítmico, en la curación, documentación y validación continua del dato (Vassilev et al., 2024; OWASP, 2025).

El segundo criterio es la explicabilidad operativa. Un sistema defensivo debe indicar no solo que existe riesgo, sino por qué un evento fue clasificado como sospechoso y qué evidencias respaldan la alerta. Esto permite convertir la predicción en conocimiento accionable para analistas y responsables de respuesta. Sin esa trazabilidad, la IA puede aumentar la incertidumbre en lugar de reducirla (Sarker et al., 2024; Salem et al., 2024).

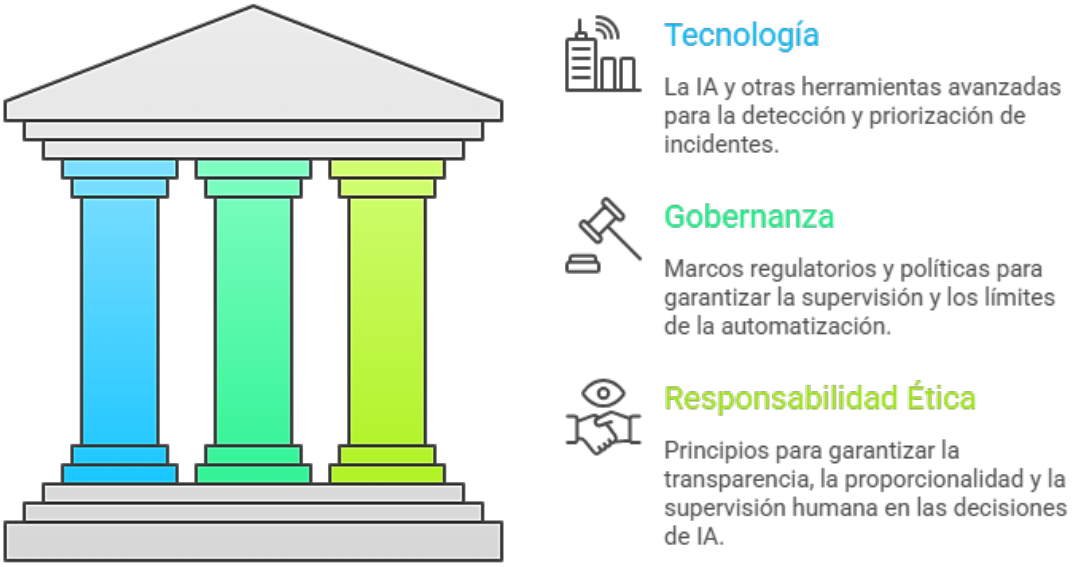
El tercer criterio es la evaluación adversarial permanente. Los modelos deben probarse frente a escenarios de evasión, prompts hostiles, datos contaminados y campañas generativas simuladas. Esta práctica permite detectar fragilidades antes de que sean explotadas y favorece la mejora continua. En un entorno de amenazas dinámicas, la robustez depende de la actualización constante (Vassilev et al., 2024; NIST, 2024).

Una defensa predictiva madura debe articular tecnología, gobernanza y responsabilidad ética. La IA puede acelerar la detección y priorización de incidentes, pero sus decisiones deben permanecer sujetas a supervisión humana, auditoría y límites claros de automatización. Así, la ciberseguridad predictiva frente a ataques generativos se consolida como un ecosistema sociotécnico orientado a anticipar amenazas sin sacrificar transparencia ni proporcionalidad (NIST, 2024; OWASP, 2025).

La defensa predictiva se fundamenta en la articulación equilibrada entre tecnología, gobernanza y responsabilidad ética, ya que la anticipación de amenazas no depende únicamente de herramientas automatizadas, sino también de criterios institucionales que orienten su uso responsable. En este sentido, la imagen sintetiza tres pilares esenciales: la tecnología, representada por la inteligencia artificial y los sistemas avanzados de detección y priorización de incidentes; la gobernanza, entendida como el conjunto de marcos regulatorios, políticas y mecanismos de control que delimitan la

automatización; y la responsabilidad ética, que asegura la transparencia, la proporcionalidad y la supervisión humana en las decisiones asistidas por IA. Estos fundamentos permiten comprender que una defensa predictiva robusta no solo busca prevenir riesgos, sino también garantizar que la seguridad digital se gestione con precisión, control y legitimidad.

Figura 2
Fundamentos de la Defensa Predictiva



Nota: (Autores, 2026).

4. Discusión

La revisión permite sostener que la ciberseguridad predictiva basada en inteligencia artificial no debe entenderse como una sustitución de los controles clásicos, sino como una ampliación estratégica de la capacidad defensiva (Galarza-Sánchez et al., 2023). Su principal aporte consiste en anticipar comportamientos anómalos, correlacionar señales dispersas y reducir la dependencia de firmas estáticas, especialmente frente a ataques generativos capaces de modificar lenguaje, código, identidad y modalidad de interacción (Kaur et al., 2023; Salem et al., 2024).

En este sentido, los ataques generativos introducen una ruptura significativa respecto de amenazas tradicionales, porque incrementan la velocidad, escala y personalización de las campañas maliciosas. El phishing asistido por IA, los deepfakes, la automatización de reconocimiento y la generación de malware adaptable evidencian que la amenaza ya no se limita al acceso técnico no autorizado, sino que combina manipulación algorítmica, ingeniería social y explotación contextual (Mirsky & Lee, 2021; Ferrag et al., 2025).

No obstante, los resultados también muestran que la inteligencia artificial defensiva posee una condición ambivalente: fortalece la detección, pero simultáneamente amplía la superficie de riesgo (Rodriguez-Vizuete et al., 2024). Los modelos pueden

ser vulnerados mediante envenenamiento de datos, evasión, extracción, inversión o manipulación adversarial, lo que obliga a abandonar cualquier lectura tecnodeterminista de la IA como solución autosuficiente (Vassilev et al., 2024; NIST, 2024).

La discusión central, por tanto, no radica únicamente en si la IA detecta más rápido, sino en bajo qué condiciones puede hacerlo de manera confiable, explicable y gobernable (Choez-Calderón & Aldo-Patricio, 2025). Un modelo con alta precisión estadística puede ser insuficiente si no permite comprender por qué emite una alerta, qué variables influyen en la decisión o cómo debe actuar el analista. Por ello, la explicabilidad constituye una exigencia operativa y no solo un atributo técnico deseable (Salem et al., 2024; Sarker et al., 2024).

Asimismo, la ciberseguridad predictiva requiere una arquitectura sociotécnica, donde los algoritmos se integren con políticas institucionales, supervisión humana, monitoreo continuo y protocolos de respuesta (Bonilla-Fierro & Boné-Andrade, 2025). Esta articulación es indispensable porque los ataques generativos explotan tanto debilidades técnicas como sesgos cognitivos, rutinas organizacionales y brechas de gobernanza. En consecuencia, la robustez defensiva depende de la interacción entre datos, modelos, personas y procesos (Kaur et al., 2023; NIST, 2024).

La principal contribución del enfoque revisado es proponer una transición desde una ciberseguridad reactiva hacia una ciberseguridad anticipatoria, adaptable y auditada. Sin embargo, esta transición solo será sostenible si se acompaña de evaluación adversarial permanente, curación rigurosa de datos, límites éticos a la automatización y mecanismos de trazabilidad (Villalva-Salguero & Toscano-Quispe, 2025). Así, la IA predictiva puede convertirse en un recurso decisivo contra ataques generativos, siempre que su adopción esté guiada por resiliencia, proporcionalidad y responsabilidad institucional (OWASP, 2025; Vassilev et al., 2024; NIST, 2024).

5. Conclusiones

La ciberseguridad predictiva basada en inteligencia artificial constituye una respuesta pertinente frente al crecimiento de ataques generativos, debido a su capacidad para anticipar patrones anómalos, correlacionar señales débiles y fortalecer la detección antes de que el incidente alcance mayor impacto. Su valor principal no reside únicamente en automatizar controles, sino en ampliar la capacidad analítica de las organizaciones ante amenazas más rápidas, personalizadas y difíciles de reconocer.

Los ataques generativos representan un desafío crítico porque combinan sofisticación técnica, manipulación social y variabilidad constante. El phishing hiperpersonalizado, los deepfakes, el malware adaptable y la suplantación automatizada evidencian que las defensas tradicionales resultan insuficientes cuando dependen solo de firmas, reglas fijas o indicadores previamente conocidos.

No obstante, la inteligencia artificial defensiva también introduce riesgos propios. Los modelos pueden fallar por sesgos, datos incompletos, opacidad, falsos positivos, falsos negativos o manipulación adversarial. Por ello, su implementación debe evitar una confianza absoluta en la automatización y considerar siempre supervisión humana, validación continua y criterios de explicabilidad.

En consecuencia, una defensa predictiva robusta requiere integrar tecnología, gobernanza, datos confiables, monitoreo permanente y evaluación adversarial. La inteligencia artificial debe operar como parte de una arquitectura de seguridad por capas, articulada con procesos institucionales, protocolos de respuesta, gestión del riesgo y formación de usuarios.

El artículo concluye que la ciberseguridad predictiva frente a ataques generativos debe avanzar hacia un modelo anticipatorio, explicable y resiliente. Su aporte principal consiste en orientar la transición desde una defensa reactiva hacia una estrategia preventiva, capaz de responder a amenazas emergentes sin descuidar la transparencia, la responsabilidad y la sostenibilidad operativa.

CONFLICTO DE INTERESES

“Los autores declaran no tener ningún conflicto de intereses”.

Referencias Bibliográficas

Boné-Andrade, M. F. (2023). Inclusión Digital y Acceso a Tecnologías de la Información en Zonas Rurales de Ecuador. *Revista Científica Zambos*, *2*(2), 1-16. https://doi.org/10.69484/rcz/v2/n2/40

Boné-Andrade, M. F., Mendoza-Loor, J. J. ., & Núñez-Freire, L. A. . (2025). Evaluación del Algoritmo F5 aplicado en la Esteganografía de imágenes JPEG: Un análisis basado en métricas de calidad. *Journal of Economic and Social Science Research, 5*(2), 144-158. https://doi.org/10.55813/gaea/jessr/v5/n2/194

Bonilla-Fierro, L. F., & Boné-Andrade, M. F. (2025). Desarrollo de plataformas de comunicación inclusivas mediante diseño universal. *Revista Científica Ciencia Y Método, 3*(2), 59-73. https://doi.org/10.55813/gaea/rcym/v3/n2/5

Castelo-Vinueza, E. M. (2025). Problemas de la investigación tecnológica y su aplicación en la generación de innovación. *Journal of Economic and Social Science Research*, *5*(1), 146–160. https://doi.org/10.55813/gaea/jessr/v5/n1/166

Choez-Calderón, C. J., & Aldo-Patricio, M. O. (2025). La ciberseguridad como prioridad empresarial dentro de marcos los regulatorios y normativos internacionales. *Revista Científica Ciencia Y Método, 3*(3), 14-27. https://doi.org/10.55813/gaea/rcym/v3/n3/38

Erazo-Luzuriaga, A. F., Ramos-Secaira, F. M., Galarza-Sánchez, P. C., & Boné-Andrade, M. F. (2023). La inteligencia artificial aplicada a la optimización de programas informáticos. *Journal of Economic and Social Science Research*, *3*(1), 48–63. https://doi.org/10.55813/gaea/jessr/v3/n1/61

European Union Agency for Cybersecurity. (2024). *ENISA threat landscape 2024*. https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*. https://doi.org/10.1016/j.iotcps.2025.01.001

Galarza-Sánchez, P. C. (2023). Adopción de Tecnologías de la Información en las PYMEs Ecuatorianas: Factores y Desafíos. *Revista Científica Zambos*, *2*(1), 21-40. https://doi.org/10.69484/rcz/v2/n1/36

Galarza-Sánchez, P. C., Erazo-Luzuriaga, A. F., & Boné-Andrade, M. F. (2023). Uso de computación cuántica en la mejora de algoritmos de aprendizaje automático. *Revista Científica Ciencia Y Método, 1*(4), 16-30. https://doi.org/10.55813/gaea/rcym/v1/n4/25

Jabir, R., Le, J., & Nguyen, C. (2025). Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors. *AI, 6*(8), 174. https://doi.org/10.3390/ai6080174

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion, 97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), Article 7. https://doi.org/10.1145/3425780

Montalván-Vélez, C. L., Mogrovejo-Zambrano, J. N., Romero-Vitte, I. J., & Pinargote-Carrera, M. L. D. C. (2024). Introducción a la Inteligencia Artificial: Conceptos Básicos y Aplicaciones Cotidianas . *Journal of Economic and Social Science Research*, *4*(1), 173–183. https://doi.org/10.55813/gaea/jessr/v4/n1/93

National Cyber Security Centre. (2024). *The near-term impact of AI on the cyber threat*. National Cyber Security Centre. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

OWASP Foundation. (2025). *OWASP Top 10 for Large Language Model Applications 2025*. https://genai.owasp.org/llm-top-10/

Rodriguez-Vizuete, J. D., Viteri-Ojeda, J. C., & Villa-Feijoó, A. L. (2024). Adopción de tecnologías sostenibles en infraestructuras de tecnologías de la información. *Revista Científica Ciencia Y Método, 2*(1), 55-67. https://doi.org/10.55813/gaea/rcym/v2/n1/31

Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data, 11*, Article 105. https://doi.org/10.1186/s40537-024-00957-y

Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. *ICT Express*. https://doi.org/10.1016/j.icte.2024.05.007

Tirira-Chulde, R. D., Rodríguez-Santillán, M. D., Taco-Cabrera, A. G., Merino-Villegas, L. R., & Tejada-Valencia, J. P. (2026). Efectos de los regímenes de conmutación sobre los parámetros eléctricos en lámparas led modulares, lámparas led compactas y lámparas fluorescentes compactas. *Revista Científica Zambos,* 5(1), 214-232. https://doi.org/10.69484/rcz/v5/n1/162

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2e2023). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2023

Villalva-Salguero, T., & Toscano-Quispe, S. Y. (2025). La brecha digital como obstáculo para la comunicación comunitaria en zonas rurales del Ecuador. *Revista Científica Ciencia Y Método, 3*(3), 278-294. https://doi.org/10.55813/gaea/rcym/v3/n3/75

World Economic Forum. (2025). *Global Cybersecurity Outlook 2025*. https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf