UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Tecnura

https://revistas.udistrital.edu.co/index.php/Tecnura/issue/view/1327

DOI: https://doi.org/10.14483/22487638.23438

REFLEXIÓN

# Oportunidades del Aprendizaje Automático Adversarial (AML) para fortalecer la ciberseguridad de la IA en el contexto colombiano

## Opportunities of Adversarial Machine Learning for Strengthening Cybersecurity of AI in Colombian context

Felipe Santiago Valderrama Ballesteros [1], Juan Manuel Cortés Jiménez [2] y Jorge Eliecer Camargo Mendoza [3]

**Fecha de Recepción**: 26 de marzo de 2025 **Fecha de Aceptación**: 18 de abril de 2026

**Cómo citar:** F.S. Valderrama-Ballesteros, J.M. Cortés Jiménez, y J.E. Camargo Mendoza, «Oportunidades del Aprendizaje Automático Adversarial (AML) para fortalecer la ciberseguridad de la IA en el contexto colombiano», *Tecnura*, vol. 30, n.° 88, jun. 2026. 69–84. https://doi.org/10.14483/22487638.23438

## Resumen

**Objetivo:** revisar los fundamentos del Aprendizaje Automático Adversarial (AML) y evaluar su potencial para el refuerzo de la ciberseguridad en sistemas de Inteligencia Artificial (IA) en Colombia.

**Metodología:** se realizó una revisión documental analítica sobre ataques adversariales tradicionales y vulnerabilidades emergentes, con énfasis en la IA Generativa (inyección de *prompt*). Posteriormente, se analizó el marco regulatorio local (CONPES 4144) y se evaluaron cuatro casos de estudio representativos en los sectores de salud, agricultura, planeación pública y asistentes virtuales corporativos (*chatbots*).

**Resultados:** los sistemas de IA en Colombia enfrentan riesgos críticos que abarcan desde el fraude predictivo hasta la exfiltración de datos en Modelos de Lenguaje Grande (LLMs). Para mitigar estas amenazas es imperativo transitar hacia arquitecturas de seguridad por diseño y aplicar estrategias de AML adaptadas al entorno.

**Conclusiones:** la integración segura de la IA en el país requiere superar barreras estructurales significativas como la limitación presupuestal de las MiPymes, la escasez de talento técnico especializado y la actual fragmentación regulatoria. Superar estos retos dependerá de una colaboración estrecha entre el gobierno, el sector privado y la academia para consolidar un entorno digital resiliente.

***Palabras clave:*** Aprendizaje Automático Adversarial, ciberseguridad, ataque cibernético, defensa cibernética, Inteligencia Artificial, inyección de prompt.

1 Estudiante del Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. Miembro del grupo de investigación UNSECURLAB. Email: fvalderramab@unal.edu.co

2 Estudiante del Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. Miembro del grupo de investigación UNSECURLAB. Email: jcortesj@unal.edu.co

3 Doctor en ingeniería y profesor asociado del Departamento de Ingeniería de Sistemas e Industrial de la Universidad Nacional de Colombia. Miembro del grupo de investigación UNSECURLAB. Email: jecamargom@unal.edu.co

## Abstract

**Objective:** To review the fundamentals of Adversarial Machine Learning (AML) and evaluate its potential to strengthen the cybersecurity of Artificial Intelligence (AI) systems in Colombia.
**Methodology:** An analytical documentary review was conducted on traditional adversarial attacks and emerging vulnerabilities, focusing on Generative AI (prompt injection). Subsequently, the local regulatory framework (CONPES 4144) was analyzed, and four representative case studies were evaluated in the healthcare, agriculture, public planning, and corporate virtual assistants (chatbots) sectors.
**Results:** AI systems in Colombia face critical risks ranging from predictive fraud to data exfiltration in Large Language Models (LLMs). To mitigate these threats, it is imperative to transition towards security-by-design architectures and apply AML strategies adapted to the environment.
**Conclusions:** The secure integration of AI in the country requires overcoming significant structural barriers, such as the budget limitations of MSMEs, the shortage of specialized technical talent, and current regulatory fragmentation. Overcoming these challenges will depend on close collaboration among the government, the private sector, and academia to consolidate a resilient digital environment.
***Keywords:*** Adversarial Machine Learning, Cybersecurity, Cyberattack, cyber defenses, Artificial Intelligence, prompt injection.

## Introducción

La inteligencia artificial (IA), en especial el Aprendizaje Automático (ML del inglés *Machine Learning*), continúa siendo un motor de transformación en numerosos sectores: mejora la productividad, impulsa el desarrollo tecnológico, optimiza la atención al cliente y gestiona sistemas complejos, entre muchas otras aplicaciones. En Colombia, la adopción de estas tecnologías ha experimentado un notable crecimiento, lo cual ha impulsado el desarrollo económico y la modernización de diversos servicios [1].

Sin embargo, esta acelerada integración tecnológica, enmarcada en una adopción global sin precedentes [2], trae consigo nuevos desafíos en materia de ciberseguridad. La evolución hacia modelos más complejos, sumada a la reciente irrupción de la inteligencia artificial generativa (GenAI), ha ampliado significativamente la superficie de ataque. Informes recientes advierten que los ciberdelincuentes están explotando estas mismas tecnologías para identificar debilidades algorítmicas y escalar sus operaciones maliciosas [3], [4].

En respuesta a estas vulnerabilidades emergentes, la política pública CONPES 4144 para regular la inteligencia artificial en Colombia subraya el compromiso del Estado con el desarrollo ético y seguro de la inteligencia artificial [5]. Este documento no solo impulsa la adopción de la IA en el país, sino que también señala la necesidad de establecer medidas de seguridad robustas para proteger tanto a las infraestructuras críticas como a los usuarios. En este contexto, el estudio del Aprendizaje Automático Adversarial (AML del inglés *Adversarial Machine Learning*) se presenta como una oportunidad vital para anticipar y mitigar riesgos potenciales pues ofrece estrategias que permiten salvaguardar la integridad y confiabilidad de los sistemas inteligentes.

Este artículo explora los fundamentos del AML, abarcando tanto el aprendizaje automático tradicional como las vulnerabilidades emergentes de los modelos de lenguaje grande (LLM), e identifica las oportunidades que ofrece para reforzar la ciberseguridad en el contexto colombiano, a la vez que enfatiza en la importancia de un enfoque preventivo de la mano con la constante evolución tecnológica.

Para lograr este objetivo, se presentan inicialmente los fundamentos teóricos del AML. Posteriormente, se analizan los desafíos específicos que plantea el contexto colombiano a través de casos de estudio representativos, complementados con la formulación de estrategias de defensa adaptadas. Finalmente, se ofrecen conclusiones sobre los hallazgos encontrados.

## Metodología

Este artículo de reflexión se deriva de una revisión documental analítica. Para su desarrollo, se diseñó un esquema metodológico estructurado en dos fases principales, orientadas a comprender las amenazas teóricas de la inteligencia artificial y evaluar su impacto potencial en el entorno nacional.

### Revisión de fundamentos del AML

La primera fase consistió en una exploración de la literatura centrada en los conceptos teóricos del Adversarial Machine Learning, que abarcó tanto modelos clásicos como las vulnerabilidades emergentes de los Modelos de Lenguaje Grande (LLMs). Las acciones metodológicas iniciaron con la consulta de bases de datos académicas y repositorios reconocidos como *IEEE Xplore*, *Scopus* y *arXiv*, con el empleo de cadenas de búsqueda con términos clave como 'Adversarial Machine Learning', 'AI Cybersecurity', 'Data Poisoning', 'Evasion Attacks', 'Prompt Injection' y 'LLM Security'. Para abordar los avances más recientes en IA Generativa, la búsqueda sistemática se complementó con una técnica de minería de referencias, a través del análisis de las referencias bibliográficas de reportes técnicos de instituciones referentes (ej. NIST, OWASP, Microsoft) para identificar literatura académica de vanguardia sobre estrategias de mitigación [6], [7], [8]. El análisis se enfocó en categorizar los ejes temáticos según el nivel de conocimiento del atacante y el vector de ataque, lo cual consolidó un marco teórico robusto.

### Análisis del contexto colombiano

La segunda fase tuvo como propósito contextualizar estos conceptos globales a la realidad tecnológica del país. Para llevar a cabo este análisis, el estudio se dividió en dos frentes: el entorno regulatorio y la aplicación práctica en la industria. En el frente regulatorio, se seleccionó y analizó la política pública CONPES 4144 [5], al ser el principal y más reciente instrumento gubernamental vigente que dicta los lineamientos éticos y de seguridad para la adopción de IA en Colombia. Por otro lado, para evaluar la materialización empírica de los riesgos, se realizó una selección de casos de estudio representativos en el territorio nacional. Dado que la documentación académica indexada sobre implementaciones locales

específicas de IA aún es emergente, la metodología incluyó la revisión de literatura gris y fuentes primarias corporativas y gubernamentales. Bajo este criterio, se analizaron cuatro casos de uso representativos del entorno nacional: ***Arkangel AI*** en el sector salud [9], ***Demetria*** en el sector agropecuario [10], las iniciativas predictivas del Departamento Nacional de Planeación (DNP) en el sector público [11], y el reciente despliegue de asistentes virtuales corporativos basados en IA Generativa en el Grupo Aval en alianza con Microsoft [12]. El análisis sobre estos casos consistió en identificar la criticidad de los datos que procesan y proyectar cómo las vulnerabilidades adversariales (desde la evasión clásica hasta la inyección de *prompts*) podrían afectar sus operaciones, justificando así la pertinencia de las estrategias de defensa sugeridas.

## Resultados

### Fundamentos del Adversarial Machine Learning

Según Vassilev, A et al. [6], el AML puede definirse como el estudio de técnicas que buscan comprometer el funcionamiento de sistemas de aprendizaje automático mediante la manipulación de datos de entrada o el modelo mismo, con el fin de inducir errores en las predicciones. Esta manipulación intencionada no sólo desafía la integridad de los modelos, sino que representa una preocupación real que puede generar serias afectaciones. Por ello, se describen las bases de estas técnicas y los principales ataques existentes.

Para empezar, los ejemplos adversariales se definen como "las entradas dadas a un modelo Aprendizaje Automático, intencionalmente diseñadas para causar errores en sus predicciones" [13]. Es decir, es la manipulación de los datos de entrada con el fin de generar predicciones incorrectas, con la particularidad de intentar ser lo menos detectable para los seres humanos. Por ejemplo, está la ya conocida imagen de Goodfellow, I et al. [14] que prueba un ejemplo adversarial en un modelo de clasificación de imágenes llamado GoogLeNet.

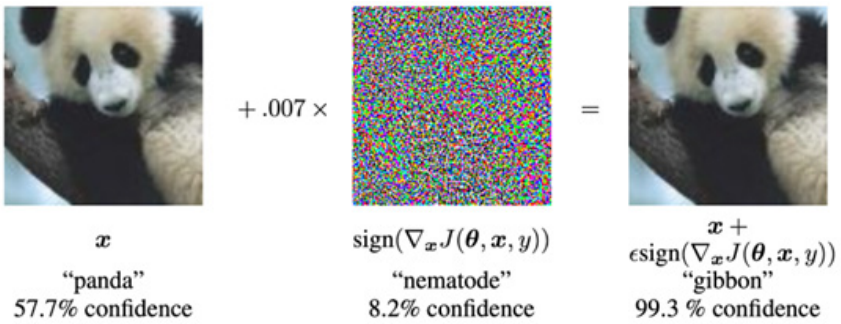


**Figura 1.** Ejemplo adversarial aplicado en el cual una pequeña perturbación imperceptible cambia la clasificación del modelo.

**Fuente:** Goodfellow, J. Shlens y C. Szegedy [14]

Esta perturbación de los datos es problemática para los modelos en diferentes niveles (de acuerdo con las características propias de su algoritmo). No obstante, en general mantienen un objetivo claro: alterar las predicciones del modelo; ahora bien, se pueden mencionar los tipos de ataques, que en términos generales buscan generar ejemplos adversariales y con esto afectar la integridad, disponibilidad o privacidad del modelo [6].

Existen diversas clasificaciones para cada tipo de ataque teniendo en cuenta diferentes factores. Vassilev, A et al. [6] los clasifica de acuerdo con el conocimiento del atacante:

- **Ataques de caja blanca:** el atacante posee un acceso total al modelo, en su arquitectura, hiper parámetros y conjuntos de datos de entrenamiento [6]. Aunque este tipo de casos no es tan probable, pueden servir como base para probar las vulnerabilidades del modelo, con el fin de observar los resultados respecto a los ejemplos adversariales y planificar qué acciones se pueden tomar para mitigar las posibles afectaciones al modelo. Wiyatno, R et al. [13] provee una extensa serie de técnicas usadas de acuerdo con el algoritmo de aprendizaje automático, destacándose las basadas en la Optimización del Gradiente o la Optimización Restringida.

- **Ataques de caja negra:** a diferencia de los anteriores, estos reflejan la situación contraria en la que los atacantes desconocen por completo los detalles de funcionamiento del modelo y solamente pueden acceder a la salida que este genera [13]. Debido a esto, este tipo es uno de los más comunes y sencillos de llevar a cabo.

- **Ataques de caja gris:** estos son un híbrido entre los de caja negra y blanca. En este escenario, el atacante tiene información parcial sobre el modelo, como por ejemplo el acceso a una distribución de datos similar a la usada para entrenar el modelo [6]. Por consiguiente, puede generar ejemplos adversariales con mayor efectividad.

A partir del conocimiento que posee un atacante, este puede utilizar diferentes ataques adversariales como los que se reseñan en la Tabla 1.

Con la reciente irrupción de la Inteligencia Artificial Generativa (GenAI), las técnicas adversariales han evolucionado más allá de la alteración de píxeles o datos numéricos, hasta abarcar ahora la manipulación del lenguaje natural. Los Modelos de Lenguaje Grande (LLMs) introducen nuevas superficies de ataque, siendo la inyección de *prompt* y el *jailbreaking* las amenazas más críticas, clasificadas por la *Open Web Application Security Project* (OWASP) como la vulnerabilidad principal en estas aplicaciones [7]. A diferencia de los ataques clásicos, estas técnicas explotan la inmensa flexibilidad lingüística del modelo para comprometer sistemas integrados, como arquitecturas de generación aumentada por recuperación (RAG) o agentes autónomos, con lo cual logran exfiltrar datos corporativos mediante instrucciones maliciosas de forma casi indetectable [16], [17].

**Tabla 1.** *Técnicas adversariales y sus descripciones*

| Ataque | Descripción |
|---|---|
| Envenenamiento de datos | Introducción de datos maliciosos durante el entrenamiento de los modelos para comprometer su comportamiento [15]. Por ejemplo, introducir imágenes manipuladas en un modelo de clasificación para generar precisiones incorrectas que afecten su veracidad. |
| Evasión | Ingreso de datos alterados con el fin de "confundir" al modelo y que este genere predicciones incorrectas. La diferencia con el anterior es que estos ataques no se realizan durante el entrenamiento sino cuando el modelo está en una etapa productiva. |
| Extracción de modelos | Obtención de información sobre el modelo (sus parámetros o estructura) a través de consultas y técnicas de ingeniería inversa [15], con el único fin de encontrar vulnerabilidades y explotarlas. |
| Inyección de *prompt* (*Prompt Injection*) | Manipulación de las entradas de un Modelo de Lenguaje Grande (LLM) para evadir sus barreras de seguridad o extraer información sensible. Puede ser directa (por el usuario) o indirecta (oculta en documentos procesados por el modelo). Un subtipo crítico es el '*jailbreaking*', que fuerza al modelo a ignorar sus protocolos de seguridad [7], [16]. |

**Fuente:** Elaboración propia con base en Birch [15].

La comprensión detallada de estas técnicas adversariales, desde la evasión en algoritmos tradicionales hasta la manipulación semántica de los LLMs, trasciende la mera curiosidad teórica; constituye, pues, una herramienta analítica indispensable para evaluar la robustez de cualquier implementación. Al analizar estas vulnerabilidades con un enfoque crítico, resulta evidente que la simple adopción de modelos predictivos o generativos de terceros introduce riesgos sistémicos si no se auditan la calidad de los datos de entrenamiento o las superficies de interacción. Si bien estas amenazas representan retos significativos a nivel global, en el contexto colombiano adquieren matices particulares debido a factores como el nivel de madurez tecnológica de las infraestructuras, las capacidades específicas de los actores de ciberseguridad nacional y el marco regulatorio existente que sigue en desarrollo. Es precisamente esta intersección entre las vulnerabilidades algorítmicas y nuestra realidad institucional la que hace imperativo analizar cómo se están abordando estos desafíos desde la política pública.

## Avances en política pública respecto a IA y ciberseguridad en Colombia

En Colombia, el uso de la inteligencia artificial ha aumentado en contextos gubernamentales, empresariales y cotidianos debido a su efecto disruptivo y las numerosas ventajas que trae implementarla. Esta adopción es clave para enfrentar "desafíos sociales, económicos y ambientales, tales como mejorar la seguridad alimentaria, reducir la pobreza, y avanzar hacia una economía basada en el conocimiento." [5]. Por esto, el gobierno ha centrado sus esfuerzos en la creación de una "Política Nacional sobre la Inteligencia Artificial, CONPES 4144" [5], la cual sirve como el insumo más adecuado para analizar los desafíos del AML en la ciberseguridad en Colombia.

Si bien el CONPES 4144 marcó un hito directivo fundamental, el ecosistema normativo colombiano se encuentra en una fase de rápida evolución legislativa para responder a las nuevas amenazas. Recientemente, este panorama ha comenzado a materializarse con iniciativas como la Ley 2502 de 2025, enfocada en penalizar severamente la suplantación de identidad agravada por el uso de IA, y el Proyecto de Ley 043 de 2025, que propone un marco de supervisión basado en categorías de riesgo y la creación de entornos de prueba controlados (*sandboxes*) para validar sistemas de manera segura [18], [19]. A pesar de estos notables avances, la academia e investigadores del sector advierten que el entorno regulatorio nacional aún es fragmentado [20]. Diversos análisis del ecosistema [21] coinciden en que el país enfrenta el enorme desafío de estructurar políticas que mitiguen los riesgos de ciberseguridad sin asfixiar la innovación tecnológica local.

Es precisamente debido a esta etapa de transición legislativa que los diagnósticos estructurales del documento CONPES 4144 mantienen plena vigencia. En él se exponen una serie de dificultades iniciales que están estrechamente relacionadas con las vulnerabilidades frente a ataques de AML en la ciberseguridad colombiana:

*Falta de políticas públicas y regulación sobre IA y ciberseguridad*

Quizás uno de los mayores desafíos es el no reconocer el riesgo latente, pues "En Colombia no se han generado políticas públicas específicas con respecto a los riesgos y efectos no deseados relacionados con la IA" [5]. Esto es un serio problema, ya que a pesar de que en la nueva política se mencionan avances en el tema, como la Política Nacional de Confianza y Seguridad Digital (CONPES 3995) [5], aún no hay una regulación específica que contemple los ataques adversariales a modelos de IA y su impacto en la seguridad nacional.

*Infraestructura tecnológica vulnerable*

Colombia tiene un obstáculo considerable: el país no posee la infraestructura adecuada para "desarrollar y operar de forma eficiente y sostenible los sistemas de IA" [5]. Es decir, el país tiene dificultades para abordar este tipo de avances tecnológicos y así mismo se encuentra en una posición vulnerable debido a ser uno de los países "con más ataques de ciberseguridad en Latinoamérica" [22]. Esto recalca la importancia de una mayor conciencia e inversión sobre la ciberseguridad en la infraestructura tecnológica del país.

*Desconocimiento sobre la seguridad en el campo de la IA*

A pesar del crecimiento en la adopción de la inteligencia artificial en sectores empresariales y gubernamentales, las estrategias de seguridad en estos sistemas no han avanzado al mismo ritmo. El documento CONPES 4144 advierte sobre el desconocimiento generalizado en torno a la seguridad digital

en los sistemas de IA, lo que representa una preocupación significativa [5]. Asimismo, diversos índices internacionales califican a Colombia como un país con un compromiso moderado en la implementación de políticas de ciberseguridad [5]. Esta brecha de conocimiento sugiere que muchos de los principales usuarios de IA, tanto a nivel institucional como individual, podrían no ser plenamente conscientes de las vulnerabilidades inherentes a estos sistemas ni de su posible explotación mediante ataques de AML.

El CONPES 4144 destaca una extensa serie de oportunidades de mejora para que el país pueda tener una adopción de la IA sostenible, segura y beneficiosa. No obstante, es importante mencionar que el principal desafío es la falta de concienciación y capacitación en seguridad digital, lo cual expone a organizaciones y ciudadanos a riesgos crecientes como la AML en entornos de IA. Por esto se resalta la necesidad de estrategias educativas y regulatorias para mitigar los impactos negativos de estas amenazas y garantizar una implementación responsable de la IA en Colombia.

## Aplicaciones de la IA y ML en Colombia

Colombia ha mostrado un creciente interés en la adopción de tecnologías de inteligencia artificial, como se pudo evidenciar con el desarrollo de políticas públicas que regulen el tema. Sin embargo, aquí se enfatiza en tres aplicaciones puntuales de estas nuevas tecnologías en iniciativas nacionales para distintos sectores, a partir de su relevancia y beneficios para el contexto colombiano.

*Caso 1: Sector de la salud*

El ML está optimizando el diagnóstico y la gestión de recursos médicos. Un caso de estudio es el de la *startup* ***Arkangel AI***, empresa colombiana que emplea algoritmos de ML para analizar imágenes médicas y detectar enfermedades como malaria, leucemia y COVID-19 en segundos. Con esta tecnología se han realizado más de 21.000 detecciones en Colombia y otros países, mejorando el acceso a diagnósticos precisos en zonas rurales y urbanas [9]. Además, ***Arkangel AI*** ofrece estas soluciones como servicio, lo que permite que instituciones de salud creen modelos predictivos sin la necesidad de conocimientos técnicos, en un intento de democratizar el uso de la inteligencia artificial para esta área [23]. Su colaboración con entidades como UNICEF demuestra su impacto en la atención médica y su alineación con la necesidad de soluciones innovadoras en un país con desafíos de cobertura sanitaria.

*Caso 2: Sector agropecuario*

La agricultura colombiana también hace uso de técnicas de aprendizaje de máquina para mejorar sus procesos. Por ejemplo, la *startup* ***Demetria*** utiliza ML para analizar datos ambientales y de calidad, con el fin de ayudar a los productores a optimizar el sabor y la gestión del café en la cadena de suministro [10]. Además, investigaciones locales han aplicado ML para predecir el rendimiento de cosechas y clasificar

zonas aptas para el cultivo de cacao, lo cual fortalece la seguridad alimentaria y la economía rural [24], [25], [26]. Estas aplicaciones son vitales en un país cuyo sector agrícola es un pilar económico.

*Caso 3: Sector Público*

El gobierno colombiano ha integrado el ML para mejorar la eficiencia en la gestión pública. El Departamento Nacional de Planeación (DNP) emplea modelos predictivos para evaluar la viabilidad de proyectos de inversión, a través del análisis de más de 8.000 iniciativas históricas con un 76 % de precisión. Esta herramienta ha identificado proyectos inviables, con un potencial ahorro de $3 000 000 000 de pesos si se hubiera implementado antes [11]. El DNP planea expandir su uso a programas sociales, lo que refleja el potencial del ML para optimizar recursos y apoyar los objetivos del CONPES 4144 [5]. Como propuesta inicial, el gobierno colombiano, a través del MinTIC, ha impulsado la educación en ML como parte de su estrategia de transformación digital [27].

*Caso 4: IA Generativa y asistentes virtuales en el sector corporativo*

El ecosistema corporativo e institucional colombiano ha acelerado vertiginosamente la adopción de IA generativa, especialmente en la interacción con sus usuarios mediante asistentes virtuales y arquitecturas basadas en Modelos de Lenguaje Grande (LLMs). Organizaciones como Empresas Públicas de Medellín (EPM), a través de su asistente "Ema", o diversas entidades del sector financiero mediante la integración de herramientas corporativas como Copilot, evidencian una masificación de estos esquemas conversacionales [12], [21]. Paralelamente, el ecosistema de *startups* locales robustece esta oferta, lo cual amplía el acceso a asistentes impulsados por IA para el servicio al cliente y la optimización de procesos [21], [28].

Estas nuevas iniciativas de empresas salientes, junto con las propuestas gubernamentales, están posicionando a Colombia como un actor emergente en el ecosistema global de IA. Sin embargo, desafíos como la privacidad de datos, los sesgos en los modelos y la necesidad de educación continua deben abordarse para mitigar riesgos en un contexto en el que la adopción de estas tecnologías sigue creciendo y se busca maximizar los beneficios.

## Riesgos y oportunidades de mejora

El aumento en la adopción de modelos de aprendizaje automático (ML) en Colombia, como se describió anteriormente, trae consigo riesgos asociados a vulnerabilidades que pueden ser explotadas mediante técnicas de *Adversarial Machine Learning* (AML). Estos riesgos, detallados en la sección de fundamentos, incluyen ataques como el envenenamiento de datos, la evasión y la extracción de modelos [29]. En este contexto, el AML no solo representa una amenaza, sino también una oportunidad para desarrollar estrategias de defensa que protejan los sistemas de IA, y el subsecuente fortalecimiento de la ciberseguridad en el país.

*Vulnerabilidades en el contexto colombiano*

Los modelos algorítmicos aplicados en sectores clave enfrentan riesgos específicos que varían según su arquitectura:

- En salud, un ataque de evasión mediante alteraciones imperceptibles a nivel de píxel en imágenes médicas puede engañar a los sistemas de clasificación diagnóstica para que emitan resultados erróneos con altísima confianza. Más allá del evidente riesgo vital para los pacientes, en el ámbito administrativo estas vulnerabilidades algorítmicas podrían ser explotadas para cometer fraudes a gran escala burlando los sistemas automatizados de reclamaciones y facturación del sistema de salud [30].

- En agricultura, los modelos predictivos dependen en gran medida de datos climáticos y registros empíricos provistos por los campesinos para optimizar el manejo de cultivos y mitigar riesgos asociados a fenómenos ambientales [24]. Un envenenamiento de estos conjuntos de datos podría alterar la evaluación de los entornos e inducir al algoritmo a recomendar decisiones catastróficas en la planificación agrícola, lo que amenazaría directamente la economía rural y la seguridad alimentaria regional.

- En políticas públicas, la manipulación de datos históricos o demográficos de entrada en modelos de ML del DNP [11] podría alterar las predicciones de éxito de los proyectos de infraestructura propuestos y desviar millonarios recursos públicos hacia obras inviables.

- En el sector corporativo y financiero, la reciente adopción de asistentes virtuales basados en LLMs introduce una superficie de ataque crítica. Como quedó demostrado con el descubrimiento de la vulnerabilidad EchoLeak (CVE-2025-32711) en plataformas como Microsoft 365 Copilot, los atacantes pueden lograr la exfiltración de datos corporativos privados y escalar privilegios sin interacción del usuario (*zero-click*), utilizando simplemente correos maliciosos diseñados para evadir los clasificadores de seguridad [31]. Para instituciones colombianas altamente integradas a estos ecosistemas, la inyección de *prompt* representa un riesgo de compromiso total.

Especialmente en el último caso, la magnitud de estas amenazas emergentes ha llevado a firmas de ciberseguridad a catalogar la inyección de *prompt* como "la nueva inyección SQL" [32]. Esta analogía advierte que, así como la inyección de código fue la vulnerabilidad reina en los inicios de las bases de datos web, la manipulación de instrucciones (*prompts*) se ha convertido en la vulnerabilidad número uno y fundacional en esta nueva era de la IA [7]. Para Colombia, donde el costo promedio de una brecha de datos podría desestabilizar severamente a las organizaciones locales [33], la contención se vuelve una urgencia. Estas vulnerabilidades se agravan por factores intrínsecos como la infraestructura tecnológica limitada y el desconocimiento generalizado sobre seguridad en IA, ambos señalados en el CONPES 4144 [5]. Además, dado que Colombia se posiciona como uno de los países con más ataques cibernéticos en América Latina [22], la necesidad de implementar defensas robustas es inaplazable.

## Estrategias de defensa

Para contrarrestar estos riesgos, se proponen las siguientes estrategias basadas en AML, adaptadas al contexto colombiano:

1. **Entrenamiento adversarial**

   Esta técnica consiste en entrenar modelos con ejemplos adversariales para aumentar su resistencia a manipulaciones [29]. Por ejemplo, en el sector salud, fortalecería la robustez de modelos predictivos frente a alteraciones en imágenes médicas en la fase de entrenamiento de los modelos.

2. **Destilación defensiva**

   Se refiere al uso de un modelo secundario para reducir la sensibilidad a perturbaciones [34]. En el sector público, su aplicación ayudaría a estabilizar las predicciones del DNP, protegiendo la asignación de recursos y con ello la viabilidad de los proyectos de inversión.

3. **Privacidad diferencial**

   Consiste en añadir ruido estadístico para proteger datos sensibles y mitigar ataques [35]. Aplicada al ecosistema sanitario, podría salvaguardar información confidencial de los pacientes frente a intentos de extracción.

4. **Monitoreo continuo**

   Detección de anomalías y desviaciones en el rendimiento del modelo en producción, que indiquen probables ataques [36]. En el agro, identificaría envenenamiento en predicciones de cosechas, asegurando su integridad.

5. **Defensa en profundidad (*Defense in Depth*) para sistemas de IA generativa**

   La naturaleza autorregresiva de los LLMs impide garantizar la seguridad con una sola barrera [37]. Es imperativo adoptar arquitecturas multicapa que incluyan el particionamiento de instrucciones (como la técnica de *Spotlighting* [8], [38]), filtros rigurosos de entrada y salida (*guardrails*), y estrictas políticas de control de acceso basadas en el principio de menor privilegio [31].

6. **Auditoría y estandarización normativa (seguridad por diseño)**

   Adoptar marcos internacionales como el OWASP Top 10 for GenAI [7] o los catálogos de tácticas de MITRE ATLAS [17] funcionaría como una defensa preventiva fundamental. Estos lineamientos

proporcionan reglas arquitectónicas estrictas que permiten a los equipos de desarrollo realizar pruebas de estrés (*Red Teaming*) e identificar vulnerabilidades antes de que el modelo salga a producción. Es decir, evitan que las organizaciones desplieguen sistemas con fallos de diseño estructurales, pasando de una seguridad reactiva a una resiliencia planificada y auditable.

Estas estrategias de defensa señalan oportunidades de mejora tangibles para la robustez de los modelos presentes en el contexto colombiano. Sin embargo hay que tener en cuenta algunas consideraciones para implementar con éxito dichas estrategias: primero, es crucial formar profesionales en de AML y ciberseguridad, aspecto que hay que tener en cuenta en los programas de capacitación como los que se están adelantando por parte del MinTIC [27], lo cual demanda más inversión en el área de la tecnología; segundo, dado que ninguna estrategia es infalible por sí sola, lo recomendable es combinar múltiples enfoques para lograr una protección más robusta, adaptada a los posibles desafíos del contexto colombiano [29].

## Conclusiones

La creciente integración de la inteligencia artificial en Colombia representa una oportunidad transformadora para impulsar la innovación en sectores estratégicos. Sin embargo, la evolución desde el aprendizaje automático tradicional hacia la adopción masiva de modelos generativos (LLMs) ha expandido drásticamente la superficie de ataque, posicionando vulnerabilidades como la inyección de *prompt* como riesgos sistémicos para las organizaciones [7]. Aunque aún no se han documentado incidentes específicos de ataques adversariales en el país, las vulnerabilidades inherentes de los modelos de inteligencia artificial hacen indispensable adoptar estrategias preventivas. Ante esta realidad, la incorporación de técnicas derivadas del *Adversarial Machine Learning* (AML) deja de ser una opción teórica para convertirse en un imperativo operativo.

No obstante, materializar estas defensas preventivas en el contexto colombiano trasciende la voluntad técnica y se enfrenta a barreras estructurales específicas. En primer lugar, existe una profunda barrera económica: los altos costos asociados al despliegue de infraestructuras de ciberseguridad avanzadas, auditorías algorítmicas y monitoreo continuo resultan prohibitivos para las micro, pequeñas y medianas empresas (MiPymes). Estas empresas constituyen la base del tejido empresarial del país y, al adoptar aceleradamente estas nuevas tecnologías sin el presupuesto adecuado para asegurarlas, quedan altamente expuestas a posibles vulneraciones [21]. En segundo lugar, Colombia enfrenta una aguda escasez de talento humano especializado [21]; el ecosistema requiere urgentemente profesionales capacitados no solo en ciberseguridad tradicional, sino en pruebas de estrés adversarial *(Red Teaming*) y mitigación de amenazas nativas de la IA. Finalmente, la actual fragmentación y transición del marco regulatorio nacional genera un escenario de incertidumbre que dificulta la adopción estandarizada de políticas de seguridad corporativa [20].

Superar estos obstáculos estructurales exige que el impulso normativo actual, ejemplificado en el CONPES 4144 y las iniciativas legislativas en curso, se materialice mediante una colaboración articulada entre el gobierno, el sector privado y la academia. Esta sinergia es vital para democratizar el acceso a herramientas de defensa, replicar modelos internacionales de aseguramiento (como *Project Glasswing* [39] o los marcos de MITRE ATLAS [17]) y fomentar programas de capacitación técnica y retención de talento en el país.

En definitiva, el camino a seguir para Colombia no es frenar la adopción de la inteligencia artificial ante el panorama adversarial, sino blindar su integración. Asumir la seguridad por diseño, cerrar la brecha de talento especializado y consolidar un marco regulatorio unificado son pasos ineludibles para garantizar que el avance tecnológico se traduzca en beneficios sostenibles, competitivos y seguros para la sociedad colombiana.

## Referencias

[1] Fedesoft, "Colombia avanza en la adopción de la Inteligencia Artificial Generativa," Fedesoft. Accessed: Apr. 16, 2026. [Online]. Available: https://fedesoft.org/colombia-avanza-en-la-adopcion-de-la-inteligencia-artificial-generativa-el-29-de-las-empresas-estan-en-fase-de-experimentacion-activa-revela-sondeo-de-fedesoft/

[2] Stanford HAI, "The 2026 AI Index Report," Stanford University Human-Centered Artificial Intelligence. Accessed: Apr. 16, 2026. [Online]. Available: https://hai.stanford.edu/ai-index/2026-ai-index-report

[3] IBM, "IBM 2026 X-Force Threat Index: AI-Driven Attacks are Escalating as Basic Security Gaps Leave Enterprises Exposed," IBM Newsroom. Accessed: Apr. 16, 2026. [Online]. Available: https://newsroom.ibm.com/2026-02-25-ibm-2026-x-force-threat-index-ai-driven-attacks-are-escalating-as-basic-security-gaps-leave-enterprises-exposed

[4] P. Girnus, V. Ciancaglini, M. Swimmer, D. Fiser, A. Oliveira, and B. Zigh, "Fault Lines in the AI Ecosystem: TrendAI™ State of AI Security Report | Trend Micro (US)," Trend. Accessed: Apr. 16, 2026. [Online]. Available: https://www.trendmicro.com/vinfo/us/security/news/threat-landscape/fault-lines-in-the-ai-ecosystem-trendai-state-of-ai-security-report

[5] G. F. Petro Urrego *et al.*, *CONPES 4144: POLÍTICA NACIONAL DE INTELIGENCIA ARTIFICIAL*. 2025. [Online]. Available: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/4144.pdf

[6] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," National Institute of Standards and Technology, Gaithersburg, MD, NIST AI 100-2e2025, 2025. doi: https://doi.org/10.6028/NIST.AI.100-2e2025

[7] OWASPGenAIProject Editor, "LLM01:2025 Prompt Injection," OWASP Gen AI Security Project. Accessed: Apr. 15, 2026. [Online]. Available: https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[8] A. Paverd, "How Microsoft defends against indirect prompt injection attacks," Microsoft. Accessed: Apr. 16, 2026. [Online]. Available: https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks

[9] J. Zea, "Arkangel Ai AI use cases for HealthCare." Aug. 15, 2025. [Online]. Available: https://arkangel.ai/en/research/arkangel-ai-predictive-models-reduce-hospital-admissions-45-68-million-chronic-patients

[10] "Sobre Demetria," Demetria. Accessed: Apr. 16, 2026. [Online]. Available: https://www.demetria.ag/colombia/sobre-a-demetria

[11] Emilio, "Machine Learning enhances Public Policy in Colombia," Technology and Operations Management. Accessed: Apr. 16, 2026. [Online]. Available: https://d3.harvard.edu/platform-rctom/submission/machine-learning-enhances-public-policy-in-colombia/

[12] News Center Microsoft Latinoamérica, "Grupo Aval y Microsoft se unen para impulsar la revolución de la inteligencia artificial en todas sus entidades," News Center Latinoamérica. Accessed: Apr. 16, 2026. [Online]. Available: https://news.microsoft.com/es-xl/grupo-aval-y-microsoft-se-unen-para-impulsar-la-revolucion-de-la-inteligencia-artificial-en-todas-sus-entidades/

[13] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker, "Adversarial Examples in Modern Machine Learning: A Review," Nov. 15, 2019, *arXiv*: arXiv:1911.05268. doi: https://doi.org/10.48550/arXiv.1911.05268

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," Mar. 20, 2015, *arXiv*: arXiv:1412.6572. doi: https://doi.org/10.48550/arXiv.1412.6572

[15] L. Birch, "AI Under Attack: Six Key Adversarial Attacks and Their Consequences," Mindgard. Accessed: Apr. 16, 2026. [Online]. Available: https://mindgard.ai/blog/ai-under-attack-six-key-adversarial-attacks-and-their-consequences

[16] S. Gulyamov *et al.*, "Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms," *Information*, vol. 17, no. 1, p. 54, Jan. 2026, doi: https://doi.org/10.3390/info17010054

[17] The MITRE Corporation, "Case Studies | MITRE ATLAS™." Accessed: Apr. 16, 2026. [Online]. Available: https://atlas.mitre.org/studies

[18] *Ley 2502 de 2025 Congreso de la República de Colombia*. Accessed: Apr. 16, 2026. [Online]. Available: https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?dt=S&i=188454

[19] MINCIENCIAS, *Proyecto de Ley "Por medio del cual se regula la inteligencia artificial en Colombia para garantizar su desarrollo ético y responsable y se dictan otras disposiciones"*. 2025. Accessed: Apr. 16, 2026. [Online]. Available: https://minciencias.gov.co/sites/default/files/upload/noticias/pl_ia_finalizado.pdf

[20] A. S. Barliza, I. C. Gómez, J. M. Caballero, and S. V. Muñoz, "Towards a National Artificial Intelligence Policy in Colombia: A Comparative Analysis of International Frameworks," *OnBoard Knowl. J.*, pp. 1–13, Feb. 2026, doi: https://doi.org/10.70554/OBJK2025.v01n01.02

[21] S. Defelipe Díaz, "IA en Colombia: Innovación y casos de éxito reciente," Impacto TIC. Accessed: Apr. 16, 2026. [Online]. Available: https://impactotic.co/inteligencia-artificial/ia-en-colombia-innovacion-y-casos-de-exito-reciente/

[22] Forbes Staff, "Colombia sigue siendo el país con más ataques de ciberseguridad en Latinoamérica, según IBM," Forbes Colombia. Accessed: Apr. 16, 2026. [Online]. Available: https://forbes.co/2024/02/28/tecnologia/colombia-es-el-pais-con-mas-ataques-de-ciberseguridad-en-latinoamerica/

[23] J. Zea, "2024_Hippocrates_Ark_Whitepaper." Aug. 15, 2025. [Online]. Available: https://arkangel.ai/en/research/no-code-hippocrates-automl-builds-pediatric-leukemia-ai-models-tenfold-faster

[24] J. Cock, D. Jiménez, H. Dorado, and T. Oberthür, "Operations research and machine learning to manage risk and optimize production practices in agriculture: good and bad experience," *Curr. Opin. Environ. Sustain.*, vol. 62, p. 101278, Jun. 2023, doi: https://doi.org/10.1016/j.cosust.2023.101278

[25] C. A. Ramírez Gómez, "Aplicación del Machine Learning en agricultura de precisión," *Rev. CINTEX*, vol. 25, no. 2, pp. 14–27, Dec. 2020, doi: https://doi.org/10.33131/24222208.356

[26] L. Talero-Sarmiento, S. Roa-Prada, L. Caicedo-Chacon, and O. Gavanzo-Cardenas, "A Data-Driven Approach to Improve Cocoa Crop Establishment in Colombia: Insights and Agricultural Practice Recommendations from an Ensemble Machine Learning Model," *AgriEngineering*, vol. 7, no. 1, p. 6, Jan. 2025, doi: https://doi.org/10.3390/agriengineering7010006

[27] J. D. Ayazo, "Formación gratuita en ciencia de datos e IA – Convocatorias," Impacto TIC. Accessed: Apr. 16, 2026. [Online]. Available: https://impactotic.co/innovacion/convocatorias-tic/oportunidad-para-formarse-de-manera-gratuita-en-ciencia-de-datos-e-ia-convocatorias/

[28] Redacción Canal Trece Colombia, "Inteligencia Artificial hecha en Colombia: empresas y creadores que están marcando la diferencia | Canal Trece." Accessed: Apr. 16, 2026. [Online]. Available: https://canaltrece.com.co/noticias/inteligencia-artificial-hecha-en-colombia-empresas-y-creadores-que-estan-marcando-la-diferencia/

[29] J. E. Fonseca Núñez and Nestlé Global Cyber SOC, "Adversarial Machine Learning for Cyber Security," MASTER'S DEGREE THESIS, Universitat Politecnica de Catalunya Barcelonatech, 2022. [Online]. Available: https://upcommons.upc.edu/server/api/core/bitstreams/5a8cde84-4c2c-4ee4-9a2b-ac74cbca634c/content

[30] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: https://doi.org/10.1126/science.aaw4399

[31] P. Reddy and A. S. Gujral, "EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System," Sep. 06, 2025, *arXiv*: arXiv:2509.10540. doi: https://doi.org/10.1609/aaaiss.v7i1.36899

[32] G. Tziakouris and Y. Kramarz, "Prompt injection is the new SQL injection, and guardrails aren't enough," Cisco Blogs. Accessed: Apr. 16, 2026. [Online]. Available: https://blogs.cisco.com/ai/prompt-injection-is-the-new-sql-injection-and-guardrails-arent-enough

[33] IBM, "Cost of a data breach 2025 | IBM." Accessed: Apr. 16, 2026. [Online]. Available: https://www.ibm.com/reports/data-breach

[34] P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis, "Defense strategies for Adversarial Machine Learning: A survey," *Comput. Sci. Rev.*, vol. 49, p. 100573, Aug. 2023, doi: https://doi.org/10.1016/j.cosrev.2023.100573

[35] G. W. Muoka *et al.*, "A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense," *Mathematics*, vol. 11, no. 20, p. 4272, Jan. 2023, doi: https://doi.org/10.3390/math11204272

[36] Y. Wang *et al.*, "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," arXiv.org. Accessed: Apr. 16, 2026. [Online]. Available: https://arxiv.org/abs/2303.06302v1

[37] M. Russinovich, A. Salem, S. Zanella-Béguelin, and Y. Zunger, "The Price of Intelligence: Three risks inherent in LLMs," *Queue*, vol. 22, no. 6, pp. 38–61, Dec. 2024, doi: https://doi.org/10.1145/3711679

[38] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kiciman, "Defending Against Indirect Prompt Injection Attacks With Spotlighting," Mar. 20, 2024, *arXiv*: arXiv:2403.14720. doi: https://doi.org/10.48550/arXiv.2403.14720

[39] "Project Glasswing: Securing critical software for the AI era," Anthropic. Accessed: Apr. 16, 2026. [Online]. Available: https://www.anthropic.com/glasswing