

Detección y prevención de material sexual de abuso infantil mediante inteligencia artificial: desafíos y avances

Isabel Leguías ^a , Giovana Garrido ^b 

^aUniversidad Tecnológica de Panamá, República de Panamá.

{isabel.leguias, giovana.garrido}@utp.ac.pa

DOI: 10.33412/pri.v17.1.4215



Resumen

El Material de Abuso Sexual Infantil (CSAM, por sus siglas en inglés) constituye una de las expresiones más graves de explotación en línea y representa un desafío crítico para la ciberseguridad y la protección infantil. El incremento global de su producción y distribución, potenciado por el uso de tecnologías digitales y la inteligencia artificial (IA), ha superado la eficacia de los métodos tradicionales de detección basados en huellas digitales o hashing perceptual. Este artículo presenta una revisión sistemática de literatura científica, informes técnicos y estudios de caso orientados a evaluar el papel de la IA en la detección y prevención de CSAM. Se analizan enfoques algorítmicos como redes neuronales convolucionales (EfficientNet, ResNet, InceptionV3), transformadores (ViT, CLIP), procesamiento de lenguaje natural (BERT, GPT), modelos de detección de anomalías, autoencoders y técnicas de hashing perceptual mejorado (PhotoDNA, NeuraHash). Asimismo, se examinan herramientas aplicadas en la práctica, entre ellas Google Content Safety API, CSAI Match, Safer Predict y la CSAM Scanning Tool de Cloudflare. Los hallazgos evidencian que la IA permite aumentar la precisión en la identificación de material ilícito y reducir significativamente los tiempos de respuesta, aunque persisten limitaciones relacionadas con la generación de CSAM sintético, los falsos positivos y la necesidad de marcos legales y éticos más sólidos. En conclusión, la IA constituye una herramienta prometedora para fortalecer la protección infantil en entornos digitales, siempre que su implementación se realice bajo estrictos estándares de responsabilidad, ética y cooperación internacional.

Palabras clave: CSAM, Inteligencia Artificial, Ciberseguridad, Detección Automática, Protección Infantil.

Title: Detection and Prevention of Child Abuse Sexual Material Using Artificial Intelligence: Challenges and Advances

Abstract: Child Sexual Abuse Material (CSAM) is one of the most severe forms of online exploitation and represents a critical

challenge for cybersecurity and child protection. The global increase in its production and distribution, fueled by digital technologies and artificial intelligence (AI), has surpassed the effectiveness of traditional detection methods based on digital fingerprints or perceptual hashing. This article presents a systematic review of scientific literature, technical reports, and case studies aimed at assessing the role of AI in the detection and prevention of CSAM. It analyzes algorithmic approaches such as convolutional neural networks (EfficientNet, ResNet, InceptionV3), transformers (ViT, CLIP), natural language processing (BERT, GPT), anomaly detection models, autoencoders, and enhanced perceptual hashing techniques (PhotoDNA, NeuraHash). Additionally, it examines practical tools such as Google Content Safety API, CSAI Match, Safer Predict, and Cloudflare's CSAM Scanning Tool. The findings show that AI increases accuracy in the identification of illicit material and significantly reduces response times, although challenges remain related to synthetic CSAM generation, false positives, and the need for stronger legal and ethical frameworks. In conclusion, AI presents a promising tool for strengthening child protection in digital environments, provided its deployment is carried out under strict standards of responsibility, ethics, and international cooperation.

Key words: CSAM, Artificial Intelligence, Cybersecurity, Automatic Detection, Child Protection.

Tipo de artículo: investigación.

Fecha de recepción: 27 de octubre de 2025.

Fecha de aceptación: 20 de febrero de 2026.

1. Introducción

El Material de Abuso Sexual Infantil (CSAM, por sus siglas en inglés) representa una de las formas más graves de explotación infantil. Se refiere a cualquier imagen, fotografía o video que muestre a menores de 18 años en contexto de abuso sexual. Este delito tiene consecuencias devastadoras para las víctimas, generando impactos psicológicos y emocionales profundos a largo plazo, incluso después de haber cesado el contacto con sus agresores. Con el auge de las tecnologías digitales y el acceso masivo a internet, la producción, distribución y consumo de CSAM han experimentado un incremento alarmante a nivel global. Según datos de la Internet Watch Foundation (IWF), en 2021 se registraron más de 29 millones de informes de presunto CSAM, evidenciando su amplia presencia en línea [1]. Recientemente, UNICEF publicó estimaciones alarmantes sobre la prevalencia del abuso sexual infantil a nivel mundial. Según el informe de octubre de 2024, más de 370 millones de niñas y mujeres, es decir, una de cada ocho, han sufrido violaciones o abusos sexuales antes de los 18 años. Esta cifra aumenta a 650 millones (una de cada cinco) cuando se incluyen formas de violencia sexual sin contacto físico, como el abuso verbal o en línea [2]. Aunque las niñas y mujeres son las más afectadas, los niños y hombres también son víctimas de abusos. Se estima que entre 240 y 310 millones de niños y hombres han sufrido violaciones o abusos sexuales durante la infancia, y esta cifra se eleva a entre 410 y 530 millones

si se incluyen formas de abuso sin contacto físico. Estas estadísticas subrayan la urgente necesidad de implementar estrategias integrales de prevención y apoyo para abordar eficazmente todas las formas de violencia y abuso sexual contra la infancia [2].

La naturaleza clandestina y la rápida propagación de este material en plataformas digitales plantean desafíos significativos para las autoridades y organizaciones dedicadas a la protección infantil. Tradicionalmente, la identificación y eliminación de CSAM han dependido de métodos manuales y sistemas de detección basados en huellas digitales conocidas, como el hashing perceptual. Sin embargo, estos enfoques resultan insuficientes ante la creciente sofisticación de los delincuentes y la proliferación de contenido generado mediante técnicas avanzadas, incluyendo el uso indebido de la Inteligencia Artificial (IA) para crear imágenes y videos de abuso sexual infantil [3].

En este contexto, la IA emerge no solo como una herramienta potencialmente mal utilizada, sino también como una solución prometedora para combatir el CSAM. Organizaciones como Thorn han desarrollado sistemas basados en IA, como Safer Predict, capaces de detectar imágenes, videos y conversaciones que podrían involucrar explotación sexual infantil, ampliando así las capacidades de identificación más allá de los métodos tradicionales [4].

A partir de este contexto, surge la necesidad de responder a la siguiente pregunta de investigación: ¿Cómo puede la inteligencia artificial contribuir eficazmente a la detección y prevención del Material de Abuso Sexual Infantil (CSAM) en entornos digitales, considerando los desafíos técnicos, éticos y legales asociados a su implementación? Este artículo analiza este problema mediante una revisión sistemática de literatura especializada y casos de aplicación reales, con el objetivo de identificar las capacidades actuales de la IA, sus limitaciones y los elementos necesarios para un uso responsable y efectivo en la protección infantil en línea.

2. Metodología

Este estudio sigue un enfoque descriptivo analítico, mediante una revisión sistemática de literatura científica, informes técnicos y estudios de caso sobre la aplicación de inteligencia artificial en la detección y prevención del Material de Abuso Sexual Infantil (CSAM). Su objetivo es evaluar las capacidades, limitaciones y desafíos de diferentes modelos y herramientas de IA aplicadas en la lucha contra estos delitos.

Identificación y selección de herramientas basadas en IA: Se recopilaron informes y estudios sobre herramientas ampliamente documentadas en la detección de CSAM, incluyendo tecnologías desarrolladas por Microsoft (PhotoDNA), Google (CSAI Match) y Thorn (Safer Predict). Se seleccionaron aquellas con respaldo en literatura científica o implementaciones en plataformas de seguridad infantil.

Por otro lado, se busca proporcionar una visión integral sobre el uso de la IA en la detección de CSAM sin requerir experimentación directa con contenido sensible, garantizando un

análisis basado en fuentes verificadas y en conformidad con principios éticos de investigación.

2.1 Revisión de Estudios de Casos

Se analizaron casos donde la IA ha demostrado ser efectiva en la detección de CSAM. Entre ellos, destacan herramientas como Safer Predict de Thorn [8], utilizada para identificar patrones sospechosos en imágenes y videos, y PhotoDNA de la empresa Microsoft [9], un sistema de hashing que permite detectar contenido ilícito de manera automatizada. Además, el informe de UNICEF titulado “Combating CSAM with IA” [10] explora el uso de IA en la protección infantil a nivel global, destacando la cooperación internacional para mejorar la eficiencia de la tecnología.

Asimismo, investigaciones como “Caught in the Net” [11] ofrecen una perspectiva valiosa sobre el rastreo de consumidores de CSAM por medio del análisis de registros digitales robados, lo que demuestra el potencial que tiene la IA no solo en la detección de material ilícito, sino también en la detección de redes delictivas.

2.2 Modelos de IA utilizados en CSAM

La detección de CSAM mediante IA se basa en varios enfoques algorítmicos:

- **Visión por computadoras:** Se emplean redes neuronales convolucionales (CNN) como EfficientNet [12], ResNet [13] y InceptionV3 [14], para el análisis de imágenes y videos, capaces de identificar patrones visuales característicos de CSAM incluso si el contenido ha sido alterado.
- **Procesamiento de Lenguaje Natural (NLP):** Se utiliza para analizar metadatos, descripciones y conversaciones en plataformas en línea que puedan estar relacionadas con la distribución o promoción de CSAM. Modelos como BERT [15] y GPT [16] han demostrado ser eficientes en la identificación de lenguaje sospechoso.
- **Aprendizaje Profundo:** Utiliza modelos como EfficientNet o ResNet los cuales se aplican para mejorar la precisión en la clasificación de imágenes sospechosas.
- **Técnicas de Hashing Perceptual Mejorado:** Aunque el hashing tradicional identifica contenido previamente conocido, las técnicas mejoradas permiten detectar variaciones de material ya identificado. PhotoDNA [17] y NeuraHash [18] son ejemplos de este tipo de tecnología aplicada en la detección de CSAM.
- **Redes Generativas Antagónicas (GAN):** Aunque en algunos casos estas redes han sido utilizadas para generar contenido sintético, también han servido para detectar deepfakes [18] y evaluar la autenticidad de imágenes en la lucha contra el CSAM.
- **Modelos de Detección de Anomalías y Autoencoders:** Estos modelos de aprendizaje no supervisado permiten identificar contenido atípico o sospechoso sin necesidad de etiquetas previas, facilitando la detección de nuevo material de abuso infantil no registrado en la base de datos [19].
- **Transformer para Análisis de Imágenes y Texto:** Utiliza modelos avanzados como Vision Transformer (ViT) [20] y CLIP [21] de OpenAI que pueden analizar imágenes sospechosas y

relacionarlas con información textual para mejorar la precisión en la clasificación de contenido ilícito.

- Modelos de Reidentificación de Víctimas y Perpetradores: Se basan en redes neuronales entrenadas para el reconocimiento de personas (ReID), permitiendo rastrear víctimas y agresores en imágenes y videos [22].

Investigaciones como "How AI is Being Abused to Create Child Sexual Abuse Imagery" [9] y "Artificial Intelligence-Produced Child Sexual" [24] analizan cómo la IA puede ser mal utilizada para generar contenido sintético de abuso infantil. Estos estudios destacan los desafíos emergentes en la detección de CSAM, especialmente en entornos como la web oscura donde el material generado por IA complica la identificación de víctimas reales [23].

Estos modelos trabajan en conjunto en soluciones avanzadas de detección de CSAM, combinando diferentes enfoques para maximizar la precisión y minimizar falsos positivos.

2.3 Avances en el uso de Modelos IA para la detección de CSAM

Investigaciones realizadas han profundizado en el uso de IA para mejorar la detección de CSAM. Por ejemplo:

- Según [25], exploran el uso del aprendizaje autosupervisado para el reconocimiento de escenas en imágenes de abuso sexual infantil, utilizando modelos preentrenados para clasificar imágenes de interiores relacionadas con CSAM en colaboración con la Policía Federal de Brasil.
- El estudio [26] presenta una herramienta automatizada para el análisis de informes de abuso sexual infantil en línea. Esta herramienta emplea técnicas de procesamiento de lenguaje natural y aprendizaje automático para categorizar informes en dimensiones como el sujeto involucrado, el grado de criminalidad y el daño causado.
- Según [27], se propone un marco para la detección de CSAM basado en metadatos, utilizando modelos de aprendizaje automático supervisado para identificar patrones sospechosos sin necesidad de procesar el contenido explícito.
- Este estudio [28] compara clasificadores de extremo a extremo con redes basadas en regiones para detectar contenido sexualmente explícito, mejorando la precisión en la detección automática de CSAM.
- Según el [29], se presenta una plantilla de análisis que utiliza modelos preentrenados para extraer señales automáticas de imágenes de CSAM, priorizando la seguridad y la ética en la manipulación de datos sensibles.

Estos estudios destacan cómo la combinación de diferentes modelos de IA puede mejorar significativamente la detección, categorización y análisis del CSAM, optimizando la respuesta de las autoridades y reduciendo la exposición de los analistas a material sensible.

2.4 Herramientas y Dataset utilizados para entrenar los modelos IA

La detección de Material de Abuso Sexual Infantil (CSAM) es un desafío crítico en la ciberseguridad y la protección infantil. El

desarrollo de modelos de inteligencia artificial en esta área depende en gran medida de las herramientas y el conjunto de datos empleados para su entrenamiento. Dada la naturaleza sensible y legalmente restringida de CSAM, el acceso a datos reales es limitado, lo que ha llevado a la creación de métodos y herramientas especializadas para abordar este problema de manera ética [30].

2.4.1 Herramientas para la Detección de CSAM

A continuación, presentamos herramientas para la detección de CSAM.

- Google Content Safety AP: Esta herramienta utiliza algoritmos avanzados de IA para ayudar a las organizaciones a identificar y priorizar la revisión de contenido potencialmente abusivo. Proporciona puntuaciones de probabilidad que indican la posibilidad de que una imagen o un video contenga CSAM, facilitando una revisión más eficiente y segura [31].
- CSAI Match: Desarrollada por Youtube, esta herramienta emplea técnicas de hash para identificar contenido previamente conocido de abuso sexual infantil. Permite una detección rápida y precisa de material ilegal ya registrado, agilizando su eliminación de las plataformas en línea [31].
- ROOST (Robust Open Online Safety Tools): Fundada por empresas como Roblox, Discord, OpenAI y Google, ROOST es una organización sin fines de lucro que desarrolla herramientas de código abierto basadas en IA para la identificación y reporte de CSAM. Estas herramientas están diseñadas para integrarse fácilmente en diversas plataformas en línea, mejorando la capacidad de estas para detectar y gestionar contenido inapropiado [32].
- CSAM Scanning Tool de Cloudflare: Esta herramienta gratuita permite a los propietarios de sitios web escanear y detectar contenido de abuso sexual infantil en sus plataformas, contribuyendo a la creación de un entorno en línea más seguro [33].
- Herramientas de Coincidencia de Hashes de Anthropic: Utiliza una base de datos de valores hash de CSAM conocidos proporcionada por el Centro Nacional para Niños Desaparecidos y Explotados (NCMEC) para detectar y reportar contenido ilícito en las entradas de los usuarios [34].

2.4.2. Datasets para el Entrenamiento de Modelos de IA

El acceso a conjuntos de datos de CSAM está estrictamente controlado debido a consideraciones legales y éticas. Sin embargo, se han desarrollado enfoques para permitir el entrenamiento de modelos de IA sin exponer directamente a los investigadores al contenido sensible:

- Análisis Seguro de Datasets de CSAM: Investigadores han propuesto plantillas de análisis que extraen señales automáticas de imágenes de CSAM sin necesidad de visualizar el contenido. Estas señales incluyen categorías de objetos y detección de pornografía, proporcionando estadísticas agregadas que preservan la privacidad y el anonimato de las víctimas. Este enfoque permite a los modelos de IA aprender características relevantes sin comprometer la ética o la legalidad [35].

- Colaboración con Agencias de Aplicación de la Ley: Debido a las restricciones en el acceso a datos reales de CSAM, es común que los desarrolladores de modelos de IA colaboren directamente con agencias gubernamentales y organizaciones autorizadas. Estas entidades poseen bases de datos seguras y aprobadas legalmente, lo que permite el entrenamiento y evaluación de modelos en entornos controlados y éticamente responsables [35].

Es importante reiterar que el manejo de datos relacionados con CSAM requiere un enfoque extremadamente cuidadoso. La privacidad de las víctimas, el cumplimiento de las leyes y regulaciones y la implementación de salvaguardas para prevenir el uso indebido de las herramientas desarrolladas son aspectos fundamentales que deben ser considerados en todo momento durante el desarrollo y despliegue de modelos IA en este ámbito [35].

3. Resultados y Discusión

Los hallazgos de este estudio evidencian que la IA ha sido una herramienta clave en la detección y mitigación del CSAM. A continuación, se presentan los principales resultados del análisis.

3.1 Eficacia de los modelos IA en la detección de CSAM

Aunque no se encontraron estudios que evalúen directamente el rendimiento de EfficientNet, ResNet e InceptionV3 en la detección de CSAM, investigaciones recientes han explorado su aplicación en la clasificación de imágenes y otras áreas, por ejemplo, un estudio evaluó la clasificación de imágenes de fondo de ojos utilizando Inception V3 y ResNet-50, demostrando su potencial en tareas de clasificación de imágenes [36]. Además, se ha propuesto el uso de clasificadores de extremo a extremo y redes basadas en regiones para detectar contenido sexualmente explícito en el contexto de CSAM, obteniendo una precisión del 90.17% [37].

3.2 Reducción del tiempo de respuesta en la identificación de CSAM

La implementación de sistemas IA en la detección de CSAM ha demostrado mejorar la eficiencia y reducir el tiempo de respuesta en la identificación de contenido ilícito. Un estudio propuso un marco para entrenar y evaluar modelos de aprendizaje automático para la clasificación de CSAM basado en metadatos, logrando una precisión del 97% y demostrando la eficacia de la IA en la detección rápida de dicho material [38].

3.3 Desafíos en la detección de contenido generado por la IA

La generación de CSAM sintético mediante herramientas de IA ha aumentado, presentando desafíos significativos para su detección. La IA ha permitido a los delincuentes crear un volumen ilimitado de CSAM, y estas imágenes generadas son visualmente indistinguibles del material real, incluso para analistas entrenados [39]. Además, los delincuentes revictimizan a las víctimas de CSAM mediante el uso de IA, editando contenido previamente

creado y compartido para generar nuevo material, lo que complica aún más su identificación con herramientas tradicionales [40].

3.4 Discusión

Los resultados indican que, si bien la IA ofrece herramientas poderosas para combatir el CSAM, también presenta desafíos que requieren atención multidisciplinaria.

- Eficiencia en la revisión de contenido: La implementación de herramientas como la Google Content Safety API ha permitido a las organizaciones priorizar la revisión de contenido potencialmente dañino, mejorando la rapidez y precisión en la identificación de material ilícito [41].
- Automatización en la detección de patrones: Modelos como EfficientNet y ResNet han demostrado ser eficaces en la identificación de contenido explícito mediante técnicas de visión por computadoras, lo que reduce la necesidad de intervención humana directa en el análisis de imágenes [42].

3.4.1 Limitaciones y Desafíos Tecnológicos

Presentamos algunas limitaciones y desafíos tecnológicos.

- Falsos positivos y negativos: Aunque los modelos de IA han mejorado en precisión, persisten desafíos relacionados con la identificación errónea de contenido, ya sea clasificando material legítimo como ilícito o viceversa [43]. En la práctica, estos errores suelen estar asociados a umbrales de decisión, ambigüedad contextual y a la ausencia de coincidencias en bases de datos de hashes cuando el contenido es novedoso o ha sido modificado. Por ello, se recomienda combinar señales (hashing + aprendizaje automático + metadatos), incorporar revisión humana en casos dudosos y aplicar mecanismos de auditoría y apelación para equilibrar protección infantil y garantías de debido proceso [30], [43].
- Adaptación a nuevas amenazas: La capacidad de los delincuentes para generar CSAM mediante IA exige una constante actualización y adaptación de las herramientas de detección para abordar estas nuevas formas de explotación [44].

3.4.2 Consideraciones éticas y legales

A continuación, algunas consideraciones éticas y legales.

- Privacidad y protección de datos: El uso de IA en la detección de CSAM debe equilibrarse con la protección de los derechos de privacidad de los usuarios, asegurando que las medidas implementadas no sean invasivas ni vulneren derechos fundamentales [41].
- Marco Legal adecuado: La rápida evolución de las tecnologías de IA ha superado en ocasiones la capacidad de los sistemas legales para adaptarse, lo que dificulta la persecución efectiva de delitos relacionados con CSAM generados artificialmente [44].

4 Desafío ético y consideraciones legales

La integración de la Inteligencia Artificial en la detección y prevención del Material de Abuso Sexual Infantil (CSAM) presenta desafíos legales significativos que deben ser abordados para garantizar una aplicación efectiva y ética de estas tecnologías.

4.1. Desafíos Éticos

- Privacidad y protección de datos: El uso de IA en la identificación de CSAM requiere el análisis de grandes volúmenes de datos, incluyendo imágenes y comunicaciones privadas. Es fundamental asegurar que la recolección y el procesamiento de estos datos respeten la privacidad de los individuos y cumplan con las regulaciones de protección de datos vigentes [41].
- Falsos positivos y negativos: Los sistemas de IA pueden cometer errores, clasificando contenido legítimo como CSAM (falsos positivos) o no detectando material ilícito (falsos negativos). Estos errores pueden tener consecuencias graves, como la acusación injusta de individuos inocentes o la perpetuación del abuso al no identificar contenido ilegal [30].
- Exposición de analistas a material sensible: Conocemos que la IA puede automatizar gran parte del proceso de detección, en ocasiones es necesaria la intervención humana para verificar el contenido. Esto expone a los analistas a material perturbador, lo que plantea preocupaciones sobre su bienestar psicológico y la necesidad de apoyo adecuado [41].
- Uso malintencionado de la IA: Así como esta tecnología es utilizada para combatir el CSAM, también puede ser empleada para generarlo. Por ejemplo, la IA puede crear imágenes hiperrealistas de abuso infantil que no involucran a víctimas reales, complicando su detección y planteando dilemas éticos sobre la naturaleza del contenido y su penalización [41].

4.2. Desafíos Legales

- Marco Legal Insuficiente: La rápida evolución de la IA ha superado el desarrollo de marcos legales adecuados. Muchas jurisdicciones carecen de leyes específicas que aborden la creación y distribución de CSAM generados por IA, lo que dificulta la persecución de estos delitos [44]. Por otro lado, los vacíos suelen aparecer en tres frentes: (a) definiciones y tipificaciones que no contemplan explícitamente contenidos generados o transformados por IA; (b) procedimientos para la obtención, preservación y valoración de la evidencia digital producida por sistemas automatizados; y (c) obligaciones y responsabilidades de las plataformas para la detección, el reporte y la cooperación. Estas asimetrías se agravan por la naturaleza transfronteriza del delito, lo que hace necesaria la armonización normativa y acuerdos de cooperación internacional [44]-[46].
- Jurisdicciones Internacionales: El CSAM se distribuye a menudo a través de plataformas en línea que operan en múltiples países, cada uno con sus propias leyes y regulaciones. Esta disparidad legal complica la cooperación internacional y la aplicación efectiva de las leyes [45].

- Derechos de los acusados: Es vital equilibrar la necesidad de proteger a los niños con los derechos legales de los acusados, incluyendo el derecho a un juicio justo y la presunción de inocencia. La evidencia obtenida mediante sistemas de IA debe ser manejada con cuidado para asegurar su admisibilidad y validez en procedimientos legales [46].
- Actualizaciones Legislativas: Es imperativo que las leyes se actualicen para abordar las nuevas formas de CSAM facilitadas por la IA. Esto incluye la penalización de la creación, posesión y distribución de material generado artificialmente que representa abuso sexual infantil, incluso si no involucra a víctimas reales [44].

Conclusión

La investigación realizada evidencia que la inteligencia artificial se ha convertido en un recurso clave para enfrentar el crecimiento exponencial del Material de Abuso Sexual Infantil (CSAM) en entornos digitales. Herramientas basadas en visión por computadora, procesamiento de lenguaje natural, técnicas de hashing perceptual avanzado y modelos de aprendizaje profundo han demostrado mejorar de manera significativa la precisión y rapidez en la identificación de este tipo de contenidos ilícitos. No obstante, la revisión también identifica limitaciones importantes respaldadas por la evidencia analizada: (i) la generación de CSAM sintético mediante IA y su capacidad de evadir enfoques basados únicamente en huellas digitales cuando no existe coincidencia exacta [39], [40]; (ii) la persistencia de falsos positivos y falsos negativos en sistemas automatizados, especialmente ante cambios en los patrones delictivos y restricciones de datos [43]; y (iii) la existencia de vacíos y heterogeneidad normativa para tipificar y perseguir de forma efectiva estos escenarios, así como para establecer responsabilidades y estándares de evidencia [44]-[46].

Esta afirmación se deriva de tres hallazgos recurrentes. En primer lugar, informes técnicos y boletines especializados describen que los modelos generativos facilitan la producción de imágenes sintéticas y la creación de variaciones a partir de material previamente conocido, incrementando el volumen potencial a analizar y dificultando la detección por hashing perceptual cuando no hay coincidencias previas [39], [40]. En segundo lugar, aun cuando se reportan altas métricas de desempeño en tareas de clasificación (por ejemplo, enfoques basados en metadatos), se reconoce que los errores no desaparecen y que la operación en contextos reales exige umbrales conservadores, flujos de revisión humana y auditoría para mitigar decisiones equivocadas (FP/FN) [43]. En tercer lugar, el análisis jurídico evidencia que la regulación avanza de forma desigual entre jurisdicciones y que persisten vacíos respecto a tipificación, cooperación transfronteriza y responsabilidad de plataformas, por lo que se recomienda fortalecer marcos normativos y estándares de cumplimiento [44]-[46].

Por otro lado, los resultados destacan que la IA, lejos de ser una solución aislada, debe concebirse como un complemento dentro de un ecosistema integral de protección infantil que combine aspectos tecnológicos, legales, éticos y sociales. La

cooperación internacional, la actualización constante de los marcos legales y la integración de prácticas de auditoría ética son indispensables para garantizar un uso responsable de estas herramientas. Asimismo, es prioritario proteger la privacidad de las víctimas, minimizar la exposición de los analistas a material sensible y establecer mecanismos de supervisión que aseguren la transparencia en los procesos de detección.

En conclusión, la IA representa un instrumento prometedor para la lucha contra el CSAM, siempre que se emplee bajo principios de responsabilidad, proporcionalidad y respeto a los derechos fundamentales. Su potencial radica no solo en la detección automatizada, sino en la posibilidad de articular respuestas rápidas, coordinadas y éticas que fortalezcan la ciberseguridad y garanticen la protección de los niños, niñas y adolescentes en el ámbito digital.

Referencias

- [1] A Courageous Voice, *The Global Fight Against Child Sexual Abuse Material (CSAM)*. [En línea]. Disponible en: <https://www.acourageousvoice.org/the-global-fight-against-child-sexual-abuse-material-csam>. [Accedido: 3-feb-2025].
- [2] UNICEF, *Más de 370 millones de niñas y mujeres en todo el mundo se ven sometidas a violaciones o abusos sexuales durante la infancia*. [En línea]. Disponible en: <https://www.unicef.org/es/comunicados-prensa/mas-370-millones-ninas-mujeres-mundo-sometidas-violaciones-abusos-sexuales-infancia>. [Accedido: 3-feb-2025].
- [3] Internet Watch Foundation (IWF), *IWF AI CSAM Report*. [En línea]. Disponible en: https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf. [Accedido: 3-feb-2025].
- [4] Thorn, *Introducing Safer Predict: Using the Power of AI to Detect Child Sexual Abuse and Exploitation Online*. [En línea]. Disponible en: <https://www.thorn.org/blog/introducing-safer-predict-using-the-power-of-ai-to-detect-child-sexual-abuse-and-exploitation-online/>. [Accedido: 3-feb-2025].
- [5] M. Ballin, K. Humphries, A. Lewis, G. Rolan, y C. Ronken, *Survivor Perspectives on Institutional Use of Child Sexual Abuse Material*, *International Journal for Crime, Justice and Social Democracy*, 2023. [En línea]. Disponible en: <https://doi.org/10.5204/ijcsd.3581>. [Accedido: 3-feb-2025].
- [6] UNICEF, *Framing the Future: How the Model National Response framework is supporting national efforts to end child sexual exploitation and abuse online*, 2022. [En línea]. Disponible en: <https://www.unicef.org/reports/framing-future>. [Accedido: 3-feb-2025].
- [7] Nicole Lin Chang, «A million child sex abuse images get 'digital fingerprints' to prevent online sharing», *Euronews.next*, 7 de junio de 2022, <https://www.euronews.com/next/2022/06/07/a-million-child-sex-abuse-images-get-digital-fingerprints-to-prevent-online-sharing> [Accedido: 5-may-2025].
- [8] Thorn. (2024). *Introducing Safer Predict: Using the Power of AI to Detect Child Sexual Abuse and Exploitation Online*. Recuperado de <https://www.thorn.org/blog/introducing-safer-predict-using-the-power-of-ai-to-detect-child-sexual-abuse-and-exploitation-online/>
- [9] Internet Watch Foundation. (2024). *Understanding the Rapid Evolution of AI Generated Child Sexual Abuse Imagery*. Recuperado de <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>
- [10] UNICEF. (2024). *Combating AI Generated CSAM* <https://www.wilsoncenter.org/article/combating-ai-generated-csam> [Accedido: 5-may-2025].
- [11] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv preprint arXiv:1905.11946*, 2019. [En línea]. Disponible en: <https://arxiv.org/pdf/1905.11946>
- [12] K. He, X. Zhang, S. Ren y J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. [En línea]. Disponible en: <https://arxiv.org/pdf/1512.03385>
- [13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens y Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826, 2016. [En línea]. Disponible en: <https://arxiv.org/pdf/1512.00567v3>
- [14] J. Devlin, M. W. Chang, K. Lee y K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2019. [En línea]. Disponible en: <https://arxiv.org/pdf/1810.04805>
- [15] T. B. Brown et al., "Language Models are Few-Shot Learners," *arXiv preprint arXiv:2005.14165*, 2020. [En línea]. Disponible en: <https://arxiv.org/pdf/2005.14165>
- [16] THORN. "Safety by Design for Generative AI: Preventing Child Sexual Abuse." 2024. [En línea]. Disponible en: <https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf>
- [17] Microsoft, "Microsoft PhotoDNA." [En línea]. Disponible en: <https://news.microsoft.com/on-the-issues/2018/09/12/how-photon-dna-for-video-is-being-used-to-fight-online-child-exploitation/>
- [18] Apple, "Expanded Protections for Children," 2021. [En línea]. Disponible en: https://www.apple.com/child-safety/pdf/Expanded_Protections_for_Children_Technology_Summary.pdf
- [18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales y J. Ortega-García, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131-148, 2020. [En línea]. Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S1566253520303110>
- [19] J. An y S. Cho, "Variational Autoencoder based Anomaly Detection using Reconstruction Probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1-18, 2015. [En línea]. Disponible en: <http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf>
- [20] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020. [En línea]. Disponible en: <https://arxiv.org/pdf/2010.11929>
- [21] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021. [En línea]. Disponible en: <https://arxiv.org/pdf/2103.00020>
- [22] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao y S. C. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. [En línea]. Disponible en: <https://arxiv.org/pdf/2001.04193>
- [23] Departamento de Seguridad Nacional de EE. UU., "La Inteligencia Artificial y la Lucha Contra la Explotación y el Abuso Sexual de Menores en Línea", 2025. [En línea]. Disponible en: https://www.dhs.gov/sites/default/files/2025-01/25_0122_k2p_genai-bulletin-spanish.pdf. [Accedido: 12-feb-2025].
- [24] National Center for Missing & Exploited Children. (2024). *Artificial Intelligence-Produced Child Sexual Abuse Material*. Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=10285671>
- [25] Valois, P. H. V., Macedo, J., Ribeiro, L. S. F., dos Santos, J. A., & Avila, S. (2024). *Leveraging Self-Supervised Learning for Scene Recognition in Child Sexual Abuse Imagery*. Recuperado de <https://arxiv.org/abs/2403.01183>
- [26] Puentes, J., Castillo, A., Osejo, W., Calderón, Y., Quintero, V., Saldarriaga, L., Agudelo, D., & Arbeláez, P. (2023). *Guarding the Guardians: Automated Analysis of Online Child Sexual Abuse*. Recuperado de <https://arxiv.org/abs/2308.03880>
- [27] Pereira, M., Dodhia, R., Anderson, H., & Brown, R. (2020). *Metadata-Based Detection of Child Sexual Abuse Material*. Recuperado de <https://arxiv.org/abs/2010.02387>
- [28] Gutfeter, W., Gajewska, J., & Pacut, A. (2024). *Detecting Sexually Explicit Content in the Context of the Child Sexual Abuse Materials (CSAM): End-to-End Classifiers and Region-Based Networks*. Recuperado de <https://arxiv.org/abs/2406.14131>
- [29] Laranjeira, C., Macedo, J., Avila, S., & dos Santos, J. A. (2022). *Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets*. Recuperado de <https://arxiv.org/abs/2204.14110>.
- [30] Google, "Descubre nuestro kit de herramientas de seguridad infantil," *Protecting Children Google*, 2024. [En línea]. Disponible en: <https://protectingchildren.google/intl/es-419/tools-for-partners/>.

- [31] Google, "How we detect, remove, and report child sexual abuse material," *Google Blog*, 2024. [En línea]. Disponible en: <https://blog.google/technology/safety-security/how-we-detect-remove-and-report-child-sexual-abuse-material/>.
- [32] J. Vincent, "Roblox, Discord, OpenAI, and Google found new child safety group," *The Verge*, 2025. [En línea]. Disponible en: <https://www.theverge.com/news/609367/roblox-discord-openai-google-roost-online-safety-tools>.
- [33] Cloudflare, "Anuncio de la herramienta CSAM Scanning, gratuita para todos los clientes de Cloudflare," *Cloudflare Blog*, 2020. [En línea]. Disponible en: <https://blog.cloudflare.com/es-es/the-csam-scanning-tool/>.
- [34] Anthropic, "Detección y Reporte de Material de Abuso Sexual Infantil (CSAM)," *Anthropic Support*, 2025. [En línea]. Disponible en: <https://support.anthropic.com/es/articles/9020328-deteccion-y-reporte-de-material-de-abuso-sexual-infantil-csam>.
- [35] C. Laranjeira, J. Macedo, S. Avila y J. A. dos Santos, "Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets," *arXiv preprint arXiv:2204.14110*, 2022. [En línea]. Disponible en: <https://arxiv.org/abs/2204.14110>.
- [36] S. S. Roy, S. K. Saha, y S. K. Saha, "Fundus image classification using Inception V3 and ResNet-50 for early detection of diabetic retinopathy," *Healthcare Technology Letters*, vol. 8, no. 2, pp. 37-41, 2021. [En línea]. Disponible en: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9975334/>.
- [37] W. Gutfeter, J. Gajewska, y A. Pacut, "Detecting sexually explicit content in the context of the child sexual abuse materials (CSAM): end-to-end classifiers and region-based networks," *arXiv preprint arXiv:2406.14131*, 2024. [En línea]. Disponible en: <https://arxiv.org/abs/2406.14131>.
- [38] M. Pereira, R. Dodhia, H. Anderson, y R. Brown, "Metadata-Based Detection of Child Sexual Abuse Material," *arXiv preprint arXiv:2010.02387*, 2020. [En línea]. Disponible en: <https://arxiv.org/abs/2010.02387>.
- [39] Humanium, "Combatir el aumento de material de abuso sexual infantil generado por la IA," 2024. [En línea]. Disponible en: <https://www.humanium.org/es/combatar-el-aumento-de-material-de-abuso-sexual-infantil-generado-por-la-ia/>.
- [40] Departamento de Seguridad Nacional de EE. UU., "La Inteligencia Artificial y la Lucha Contra la Explotación y el Abuso Sexual Infantil," 2025. [En línea]. Disponible en: https://www.dhs.gov/sites/default/files/2025-01/25_0122_k2p_genai-bulletin_spanish.pdf.
- [41] Google, "How we detect, remove, and report child sexual abuse material," *Google Blog*, 2024. [En línea]. Disponible en: <https://blog.google/technology/safety-security/how-we-detect-remove-and-report-child-sexual-abuse-material/>.
- [42] S. S. Roy, S. K. Saha, y S. K. Saha, "Fundus image classification using Inception V3 and ResNet-50 for early detection of diabetic retinopathy," *Healthcare Technology Letters*, vol. 8, no. 2, pp. 37-41, 2021. [En línea]. Disponible en: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9975334/pdf/fphys-14-1126780.pdf>.
- [43] M. Pereira, R. Dodhia, H. Anderson, y R. Brown, "Metadata-Based Detection of Child Sexual Abuse Material," *arXiv preprint arXiv:2010.02387*, 2020. [En línea]. Disponible en: <https://arxiv.org/pdf/2010.02387>.
- [44] L. Moyano, "El uso de la IA en la lucha contra el CSAM: Desafíos legales y éticos," *Centro de Investigaciones Judiciales de la Provincia de Buenos Aires*, 2023. [En línea]. Disponible en: https://cijur.mpba.gov.ar/files/bulletins/El_Uso_De_Da_IA_-_Dr._Lucas_Moyano_29-8.pdf. [Accedido: 12-ago-2025].
- [45] UNICEF & INTERPOL. (2023). "INTERPOL and UNICEF sign cooperation agreement to address child sexual exploitation and online abuse," [En línea]. Disponible en: <https://www.unicef.org/press-releases/interpol-and-unicef-sign-cooperation-agreement-address-child-sexual-exploitation-and>? [Accedido: 22-may-2025].
- [46] *Indret*, "La evidencia digital y la inteligencia artificial: Implicaciones legales en casos de CSAM," 2023. [En línea]. Disponible en: <https://indret.com/wp-content/uploads/2023/04/1767.pdf>. [Accedido: 22-feb-2025].