Ciencia y Educación
(L-ISSN: 2790-8402 E-ISSN: 2707-3378)
Vol. 7 No. 9.1
Edición Especial IX 2026

INGENIERÍA SOCIAL MEDIANTE DEEPFAKES: ESTRATEGIA DE DETECCIÓN Y PREVENCIÓN PARA USUARIOS WEB
SOCIAL ENGINEERING VIA DEEPFAKES: DETECTION AND PREVENTION STRATEGY FOR WEB USERS

Autores: [1]Pedro Antonio Cevallos Ortiz, y [2]Marlene Guadalupe Castillo Pinargote.
[1]ORCID ID: https://orcid.org/0009-0003-8712-2755
[2]ORCID ID: https://orcid.org/0000-0002-7001-4382
[1]E-mail de contacto: pcevallos6727@utm.edu.ec
[2]E-mail de contacto: marlene.castillo@utm.edu.ec
Afiliación: [1*2*]Universidad Técnica de Manabí, (Ecuador).
Artículo recibido: 3 de Septiembre del 2026.
Artículo revisado: 6 de Septiembre del 2026.
Artículo aprobado: 6 de Septiembre del 2026.
[1]Estudiante de pregrado de la carrera de Tecnologías de la Información y Comunicación de la Universidad Técnica de Manabí, (Ecuador).
[2]Ingeniera en Sistemas Informáticos, graduada de la Universidad Técnica de Manabí, (Ecuador). Magíster en Tecnologías de la Información con mención en Seguridad de Redes y Comunicaciones, graduada de la Universidad Técnica de Manabí, (Ecuador).

Resumen

La inteligencia artificial generativa ha favorecido la creación de deepfakes realistas, incrementando los riesgos de ingeniería social, suplantación de identidad, fraude y desinformación entre usuarios web. El estudio tuvo como objetivo analizar estos riesgos y diseñar una estrategia de detección y prevención. Se empleó un enfoque mixto, con investigación documental y de campo, bajo un diseño no experimental y transversal. Se aplicaron métodos analítico-sintético, inductivo y descriptivo. La población estuvo conformada por 350 habitantes de la comunidad María Luisa Zona 2, Pedernales, Manabí, de la cual se seleccionaron 180 usuarios adultos de internet. Se utilizó una encuesta con preguntas sociodemográficas y de conocimiento, identificación, prevención y verificación de deepfakes. El 35,6 % de los participantes nunca había escuchado el término deepfake y solo el 17,8 % manifestó conocer su significado. Asimismo, el 63,9 % presentó poca o ninguna confianza para identificarlos, mientras que el 62,2 % evidenció prácticas insuficientes de verificación. Ante situaciones de clonación de voz, el 43,3 % adoptaría respuestas inseguras. Sin embargo, el 88,9 % consideró útil o muy útil recibir capacitación. Los resultados evidencian que la exposición a internet, principalmente mediante dispositivos móviles y redes sociales, no garantiza competencias para reconocer contenidos manipulados. Por ello, se propone una estrategia de ocho etapas: pausar, observar, comprobar el origen, verificar, analizar, contrastar, confirmar la identidad y decidir. Su aplicación integra alfabetización digital, pensamiento crítico y herramientas de verificación para fortalecer la prevención y reducir la vulnerabilidad frente a ataques de ingeniería social mediante deepfakes.

Palabras clave: Deepfakes, Ingeniería social, Inteligencia artificial generativa, Detección, Prevención, Alfabetización digital.

Abstract

Generative artificial intelligence has facilitated the creation of realistic deepfakes, increasing the risks of social engineering, identity theft, fraud, and misinformation among web users. This study aimed to analyze these risks and design a detection and prevention strategy. A mixed-methods approach was employed, combining documentary and field research within a non-experimental, cross-sectional design. Analytical-synthetic, inductive, and descriptive methods were applied. The study population consisted of 350 residents of the María Luisa community (Zone 2, Pedernales, Manabí), from which 180 adult internet users were selected. A survey was used to gather data on sociodemographics as well as knowledge, identification, prevention, and verification regarding deepfakes. Results showed that 35.6% of participants had never heard the term "deepfake," and only 17.8% claimed to know its meaning. Furthermore, 63.9% expressed little

to no confidence in their ability to identify deepfakes, while 62.2% demonstrated insufficient verification practices. Faced with potential voice cloning scenarios, 43.3% indicated they would adopt unsafe responses. However, 88.9% considered receiving training to be useful or very useful. The findings demonstrate that internet exposure primarily via mobile devices and social media does not guarantee the competence to recognize manipulated content. Consequently, an eight-stage strategy is proposed: pause, observe, check the source, verify, analyze, cross-reference, confirm identity, and decide. Its implementation integrates digital literacy, critical thinking, and verification tools to strengthen prevention and reduce vulnerability to social engineering attacks involving deepfakes.

Keywords: Deepfakes, Social engineering, Generative artificial intelligence, Detection, Prevention, Digital literacy.

Sumário

A inteligência artificial generativa favoreceu a criação de deepfakes realistas, aumentando os riscos de engenharia social, falsificação de identidade, fraude e desinformação entre usuários da web. O estudo teve como objetivo analisar esses riscos e elaborar uma estratégia de detecção e prevenção. Utilizou-se uma abordagem mista, com pesquisa documental e de campo, sob um delineamento não experimental e transversal. Foram aplicados métodos analítico-sintético, indutivo e descritivo. A população foi composta por 350 habitantes da comunidade María Luisa (Zona 2), em Pedernales, Manabí, da qual foram selecionados 180 usuários adultos de internet. Aplicou-se um questionário com perguntas sociodemográficas e sobre conhecimento, identificação, prevenção e verificação de *deepfakes*. Cerca de 35,6% dos participantes nunca tinham ouvido o termo deepfake, e apenas 17,8% afirmaram conhecer seu significado. Além disso, 63,9% demonstraram pouca ou nenhuma confiança para identificá-los, enquanto 62,2% apresentaram práticas insuficientes de verificação. Diante de situações de clonagem de voz, 43,3% adotariam respostas inseguras. No entanto, 88,9% consideraram útil ou muito útil receber capacitação. Os resultados evidenciam que a exposição à internet, principalmente por meio de dispositivos móveis e redes sociais, não garante competências para reconhecer conteúdos manipulados. Por isso, propõe-se uma estratégia de oito etapas: pausar, observar, verificar a origem, conferir, analisar, confrontar, confirmar a identidade e decidir. Sua aplicação integra letramento digital, pensamento crítico e ferramentas de verificação para fortalecer a prevenção e reduzir a vulnerabilidade a ataques de engenharia social via deepfakes.

Palavras-chave: Deepfakes, Engenharia social, Inteligência artificial generativa, Detecção, Prevenção, Letramento digital.

Introducción

La expansión acelerada de la inteligencia artificial (IA), especialmente de las técnicas de aprendizaje profundo y de la inteligencia artificial generativa (IAG), ha favorecido la producción de contenidos sintéticos cada vez más sofisticados. Entre estas tecnologías destacan los *deepfakes*, entendidos como imágenes, videos o audios generados o manipulados artificialmente para reproducir con elevado realismo la apariencia, voz, gestos o comportamiento de una persona, dificultando su diferenciación respecto de contenidos auténticos (Ramos, 2024; Heidari et al., 2024).

Aunque estas tecnologías poseen aplicaciones legítimas en ámbitos como el entretenimiento, la educación y la producción audiovisual, su utilización maliciosa plantea importantes desafíos relacionados con la seguridad digital, la privacidad, la desinformación y la confianza en los medios electrónicos. La evolución tecnológica ha ampliado considerablemente las posibilidades de generación de contenidos falsificados. Inicialmente, los *deepfakes* estuvieron asociados principalmente con el

intercambio facial y la manipulación de videos; sin embargo, la incorporación de Redes Generativas Adversariales (GAN), autoencoders, modelos de difusión y otras arquitecturas generativas ha permitido producir imágenes, videos, audios y contenidos multimodales con niveles crecientes de realismo (Fernández et al., 2024; Wang et al., 2025; Waseem y Elneel, 2023; Jheelan y Pudaruth, 2025). Abbas y Taeihagh (2024) advierten que esta evolución ha transformado los *deepfakes* de una tecnología experimental en un fenómeno con implicaciones sociales y de seguridad, mientras que Schmitt y Flechais (2024) señalan que la disponibilidad de herramientas generativas está modificando las capacidades empleadas en ataques de ingeniería social y *phishing*.

Esta problemática adquiere especial relevancia en el ámbito de la ingeniería social, entendida como el conjunto de técnicas destinadas a manipular el comportamiento humano para obtener información, recursos, accesos o determinadas acciones. A diferencia de los ataques centrados exclusivamente en vulnerabilidades técnicas, la ingeniería social explota factores como la confianza, la autoridad, la urgencia, el miedo y la familiaridad (Navas y Pérez, 2023; Di Gionantonio y Ligorria, 2022). En este escenario, los *deepfakes* pueden incrementar la credibilidad de una solicitud fraudulenta al incorporar imágenes, videos o voces aparentemente legítimas, fortaleciendo las estrategias tradicionales de engaño (Babaei et al., 2025).

La combinación entre IAG e ingeniería social facilita modalidades de fraude cada vez más personalizadas. Entre sus posibles aplicaciones se encuentran la suplantación de ejecutivos, funcionarios, familiares, profesionales y representantes de organizaciones, así como campañas de *phishing*, solicitudes fraudulentas y clonación de voz (Jabir et al., 2025; Blancaflor et al., 2024). La reproducción artificial de una voz o imagen conocida puede utilizarse para generar situaciones de urgencia, solicitar transferencias económicas, inducir la entrega de información sensible o provocar decisiones basadas en una falsa percepción de autenticidad. El riesgo aumenta cuando los usuarios confían exclusivamente en señales visuales o auditivas sin realizar mecanismos adicionales de comprobación de identidad.

Por ello, los *deepfakes* deben ser comprendidos como una amenaza sociotécnica, en la que la tecnología proporciona los mecanismos para generar la falsificación y los factores psicológicos determinan su capacidad persuasiva. Villamar et al. (2026) señalan que la vulnerabilidad de los usuarios no depende únicamente de sus conocimientos técnicos, sino también de la confianza en la fuente, la familiaridad con la identidad representada, la urgencia del mensaje y la predisposición a aceptar determinada información. Esta situación demuestra que la seguridad frente a contenidos sintéticos requiere considerar conjuntamente factores tecnológicos, cognitivos y conductuales.

La dificultad de los usuarios para identificar contenidos manipulados incrementa esta vulnerabilidad. Diel et al. (2024), mediante una revisión sistemática y metaanálisis de 56 estudios con más de 86.000 participantes, encontraron una precisión humana aproximada del 55,54 % en la identificación de *deepfakes*, un desempeño cercano a una elección aleatoria. De manera similar, Groh et al. (2024) sostienen que la creciente calidad de los contenidos sintéticos reduce la efectividad de la percepción humana como mecanismo independiente de

verificación. Estos resultados evidencian que la inspección visual o auditiva por sí sola resulta insuficiente y justifican la necesidad de combinar alfabetización digital, procedimientos de comprobación y herramientas automatizadas de detección.

Las consecuencias de esta problemática trascienden el fraude individual. Los *deepfakes* pueden facilitar suplantaciones de identidad, extorsión, *phishing*, manipulación informativa y ataques contra organizaciones, afectando dimensiones económicas, reputacionales y operativas (Park et al., 2024; Alzaga, 2025; Zambrano, 2024; Cruz, 2026). Además, la expansión de contenidos sintéticos puede deteriorar progresivamente la confianza en los entornos digitales. García et al. (2025) describen este fenómeno como una crisis de veracidad, mientras que Astudillo Muñoz (2024) considera que la desinformación constituye un desafío que requiere respuestas integrales y colaborativas.

En este mismo sentido, Ma et al. (2025) advierten que la IAG facilita la producción de desinformación a gran escala, especialmente cuando los contenidos circulan mediante plataformas digitales y redes sociales. El problema no se limita a que una falsificación sea aceptada como verdadera, sino que también puede generar un escenario en el que los usuarios desconfíen incluso de contenidos auténticos. Pennycook y Rand (2021) relacionan este fenómeno con un efecto de incertidumbre que debilita la confianza informativa y refuerza la necesidad de desarrollar mecanismos de autenticación, trazabilidad y alfabetización mediática. Ante estas amenazas, la literatura científica ha desarrollado diferentes métodos automatizados para la detección de *deepfakes*. En contenidos visuales destacan las redes neuronales convolucionales, capaces de identificar inconsistencias faciales, patrones de iluminación, texturas y movimientos anómalos (Alrashoud, 2025). Para audio se han empleado arquitecturas como CNN, ResNet, GNN, Transformer, TDNN y DARTS, orientadas a analizar características espectrales, frecuencia, prosodia y patrones propios de voces sintéticas (Li et al., 2025). Asimismo, la detección multimodal permite integrar simultáneamente imagen, video, audio y otros tipos de evidencia, lo que resulta especialmente pertinente frente a falsificaciones que combinan varias modalidades (Javed et al., 2025; Rzayeva et al., 2026).

El análisis de los modelos incluidos en la literatura evidencia una importante diversidad de enfoques técnicos. Modelos como OneClass Variational Autoencoder se orientan a identificar anomalías respecto del contenido auténtico; DFT-MF analiza características biométricas vinculadas con dientes y movimientos bucales; la regresión SVM clasifica contenido a partir de características previamente extraídas; y FG-TEFusionNet combina información geométrica y textural del rostro. A estos se suman arquitecturas como ResNeXt, LSTM, ResNet50V2, ResNet101V2 y ResNet152V2, utilizadas para reconocer patrones espaciales y temporales, así como Vision Transformers, que analizan relaciones globales entre regiones de una imagen, y MEXFIC, que integra técnicas de ensamble y explicabilidad.

Estos modelos muestran que la detección ha evolucionado desde el análisis de características aisladas hacia arquitecturas más profundas, multimodales y explicables. Esta evolución también se refleja en herramientas destinadas a trasladar los modelos técnicos hacia interfaces más accesibles para diferentes usuarios. Entre

las soluciones analizadas se encuentran Sensity AI, orientada al análisis forense de imágenes y videos; Deepware, especializada en videos; DeepFake-O-Meter y Centinel, que admiten diferentes modalidades; Microsoft Video AI Authenticator, centrado en evidencias visuales de manipulación; Facia, que analiza movimientos faciales y sincronización; Hive Moderation, capaz de identificar contenido sintético en imagen, video y audio; y Resemble AI, que utiliza detección multimodal.

El análisis conjunto evidencia una tendencia hacia herramientas que proporcionan puntuaciones o indicadores de probabilidad de manipulación, facilitando la interpretación de los resultados por parte del usuario. Sin embargo, estos sistemas deben considerarse mecanismos de apoyo y no verificadores infalibles, debido a que la evolución continua de los generadores puede reducir el desempeño de modelos entrenados con falsificaciones conocidas.

En el plano técnico, la investigación también evidencia la importancia de los conjuntos de datos empleados para desarrollar y evaluar estos sistemas. Bharati et al. (2026) señalan que la detección continúa siendo un desafío crítico, mientras que conjuntos como FaceForensics++, Synthetic Faces High Quality, FakeFaces, OpenFake, HiDF, DFDC, CelebHQ y HOHA son utilizados para entrenar y comprobar modelos de detección (Vamsi et al., 2022; Momin et al., 2025). La variedad de estos *datasets* refleja la necesidad de entrenar modelos con diferentes tipos de manipulaciones y modalidades, aunque también evidencia que la efectividad de un detector puede depender considerablemente del conjunto de datos con el que fue desarrollado. No obstante, una estrategia de prevención dirigida a usuarios web no puede depender exclusivamente de herramientas técnicas. Buchan et al. (2024) destacan la alfabetización digital como una competencia fundamental para evaluar contenidos, mientras que Tiernan et al. (2023) resaltan habilidades como identificar fuentes, contrastar información, reconocer señales de manipulación y verificar datos mediante fuentes independientes. Colunga y González (2025) amplían esta perspectiva al considerar que la alfabetización digital debe integrar conocimientos básicos sobre IA, pensamiento crítico y comportamiento responsable. De esta manera, la formación preventiva debe preparar al usuario para cuestionar contenidos aparentemente auténticos antes de actuar o compartirlos.

La formación y concienciación adquieren mayor relevancia al considerar que el conocimiento de los entornos digitales no garantiza por sí mismo la capacidad de detectar una falsificación. Ramos (2024) plantea que la formación debe incorporar situaciones reales, análisis de fuentes, comparación de información y utilización guiada de herramientas de verificación. Además, debe abordar los componentes psicológicos de la ingeniería social, puesto que factores como urgencia, miedo, confianza o autoridad pueden provocar decisiones inseguras incluso cuando el usuario reconoce la posibilidad de que el contenido sea falso. Por ello, la prevención debe concebirse como un proceso continuo que articule capacidades cognitivas, técnicas y conductuales.

Desde esta perspectiva, una respuesta preventiva integral puede estructurarse en cuatro capacidades: identificar posibles señales de manipulación, analizar críticamente el contexto de la comunicación, verificar la información mediante fuentes o herramientas independientes y adoptar una conducta segura

antes de realizar acciones sensibles o compartir contenidos. El propósito no consiste en convertir al usuario en especialista en análisis forense, sino en proporcionarle un procedimiento sencillo, reproducible y comprensible que reduzca su vulnerabilidad frente a intentos de manipulación.

A pesar del avance considerable en modelos automáticos de detección, herramientas de verificación y estudios sobre alfabetización digital, persiste una brecha entre el desarrollo tecnológico y la capacidad cotidiana de los usuarios para responder ante ataques de ingeniería social sustentados en *deepfakes*. La literatura aborda ampliamente la detección algorítmica y, de manera paralela, los factores humanos de vulnerabilidad; sin embargo, resulta necesario integrar ambas perspectivas en estrategias prácticas que combinen comprobación tecnológica, análisis contextual, alfabetización mediática y conducta preventiva. Esta necesidad cobra especial importancia ante contenidos cuya calidad puede superar los mecanismos tradicionales de verificación basados únicamente en la percepción humana.

La relevancia del estudio se sustenta, por tanto, en la necesidad de comprender cómo los *deepfakes* están modificando las dinámicas tradicionales de la ingeniería social y ampliando las posibilidades de fraude, suplantación y manipulación en entornos digitales. Desde una perspectiva práctica, sus resultados pueden contribuir al fortalecimiento de las competencias de los usuarios web para evaluar la autenticidad de los contenidos y utilizar herramientas de verificación de manera responsable. Desde el ámbito de la seguridad digital, la investigación aporta una aproximación que reconoce que la prevención requiere combinar soluciones tecnológicas con alfabetización mediática, pensamiento crítico y protocolos de comprobación antes de responder a solicitudes potencialmente fraudulentas. En correspondencia con esta problemática, el objetivo de la investigación es analizar críticamente el fenómeno de los *deepfakes* y su relación con la ingeniería social, identificar sus principales implicaciones para la seguridad digital y diseñar una estrategia de detección y prevención orientada a fortalecer las capacidades de los usuarios web frente a su utilización maliciosa. Para ello, el estudio se sustenta en la revisión de literatura reciente, el análisis de casos documentados y la integración de mecanismos de alfabetización y verificación digital.

Materiales y Métodos

La investigación se desarrolló bajo un enfoque mixto, integrando procedimientos cualitativos y cuantitativos. El componente cualitativo se fundamentó en la revisión y análisis de literatura científica sobre deepfakes, ingeniería social, riesgos, métodos de detección y estrategias de prevención. El componente cuantitativo permitió diagnosticar el nivel de conocimiento y la percepción de los usuarios web respecto a la identificación de imágenes, videos y audios potencialmente manipulados mediante IA.

El estudio fue de tipo descriptivo, documental y de campo, con un diseño no experimental y transversal. La investigación documental permitió establecer los fundamentos teóricos del estudio, mientras que la investigación de campo facilitó la obtención directa de información de los participantes en un momento determinado, sin manipulación de variables. En cuanto a los métodos de investigación, se emplearon los métodos analítico-sintético, inductivo y descriptivo. El método analítico-sintético permitió examinar los principales elementos relacionados con los deepfakes y su

utilización en la ingeniería social, para posteriormente integrarlos en una visión general que fundamentó la construcción de la estrategia propuesta. El método inductivo facilitó la identificación de las principales necesidades y dificultades de los usuarios a partir de los resultados obtenidos en el estudio. Por último, el método descriptivo permitió caracterizar el nivel de conocimiento de los participantes, así como sus prácticas y capacidades relacionadas con la identificación y verificación de deepfakes.

Como técnicas de recolección de información se utilizaron la revisión bibliográfica y la encuesta. La revisión bibliográfica permitió analizar investigaciones y fuentes científicas relacionadas con el objeto de estudio y la encuesta se aplicó mediante un formulario virtual a los participantes seleccionados. El instrumento utilizado fue un cuestionario estructurado, conformado por dos secciones. La primera incluyó seis preguntas de carácter sociodemográfico, orientadas a caracterizar a los participantes y sus principales hábitos de uso de internet. La segunda estuvo integrada por diez preguntas cerradas y de opción múltiple, diseñadas para evaluar el conocimiento, la capacidad de identificación y las prácticas de prevención y verificación de los usuarios frente a posibles deepfakes.

La selección y formulación de los ítems se fundamentó en dimensiones abordadas en investigaciones previas sobre alfabetización en deepfakes y percepción de la confiabilidad de contenidos sintéticos (Plohl et al., 2025; Calabrese et al., 2026). La población de estudio estuvo conformada por los usuarios de internet de la comunidad María Luisa Zona 2, ubicada en el cantón Pedernales, provincia de Manabí, Ecuador. De una población total de 350 habitantes, se consideró como población objetivo a 180 personas mayores de edad que utilizan habitualmente internet, redes sociales, aplicaciones de mensajería o diversas plataformas digitales. La selección de los participantes se realizó considerando criterios de accesibilidad y participación voluntaria. El procedimiento metodológico de la investigación se desarrolla en tres etapas secuenciales e interrelacionadas, las cuales se presentan de manera gráfica en la Figura 1:

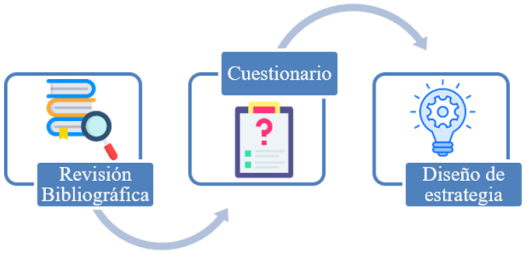

Figura 1. *Procedimiento metodológico de la investigación.*

Fuente: Elaboración propia.

La primera etapa comprende la revisión bibliográfica sobre los deepfakes, la ingeniería social, sus riesgos, los métodos de detección y las estrategias de prevención; la segunda corresponde a la aplicación de un cuestionario virtual para diagnosticar el nivel de conocimiento, la percepción de riesgo y las prácticas de identificación y verificación de los participantes de la comunidad María Luisa Zona 2; y, finalmente, la tercera etapa integra los resultados obtenidos para diseñar una estrategia de detección y prevención dirigida a usuarios web, fundamentada en la observación crítica de los contenidos y en el uso de herramientas digitales gratuitas, accesibles y de fácil utilización. A continuación, se presenta el cuestionario virtual compuesto por seis preguntas de carácter sociodemográfico y diez relacionadas con el conocimiento e identificación de deepfakes.

Tabla 1. *Cuestionario sobre el conocimiento y la identificación de deepfakes.*

Ítem	Dimensión	Pregunta	Opciones
1	Datos sociodemográficos	¿Cuál es su rango de edad?	18–24 años. 25–34 años. 35–44 años. 45–54 años. 55 años o más.
2		¿Con cuál de las siguientes opciones se identifica?	Masculino. Femenino. Otro. Prefiero no responder.
3		¿Cuál es su nivel más alto de educación alcanzado?	Educación básica. Bachillerato. Educación superior. Posgrado. Doctorado. Otro, especifique:
4		¿Con qué frecuencia utiliza internet?	Varias veces al día. Una vez al día. Varias veces por semana. Ocasionalmente. Nunca.
5		¿Qué dispositivo utiliza con mayor frecuencia para acceder a internet?	Teléfono móvil. Computadora. Tableta. Otro, especifique:
6		¿Para qué utiliza principalmente internet?	Redes sociales. Comunicación y mensajería. Trabajo o estudios. Entretenimiento. Compras o trámites. Otro, especifique: ____
7	Conocimiento previo	¿Antes de realizar este cuestionario había escuchado la palabra deepfake?	A) Sí, y sé qué significa. B) Sí, pero sé poco sobre el tema. C) He escuchado la palabra, pero no sé qué significa. D) No, nunca la había escuchado.
8	Conocimiento conceptual	¿Cuál de las siguientes opciones explica mejor qué es un deepfake?	A) Un contenido digital creado o modificado mediante inteligencia artificial para hacer que parezca real. B) Un programa utilizado para proteger dispositivos frente a amenazas digitales. C) Un contenido eliminado o bloqueado de internet. D) Una aplicación utilizada para editar fotografías y videos.
9	Conocimiento sobre tipos de contenido	¿En qué tipo de contenidos pueden aparecer deepfakes?	A) Solo en videos. B) En imágenes y videos. C) En imágenes, videos y audios. D) Solo en audios.
10	Autoeficacia para la detección	Si encontrara un posible deepfake en internet, ¿qué tan seguro/a se sentiría de poder reconocerlo?	A) Muy seguro/a. B) Algo seguro/a. C) Poco seguro/a. D) Nada seguro/a.
11	Identificación de señales de manipulación	Al observar una imagen o video en internet, ¿cuál de estas situaciones podría hacerle sospechar que ha sido manipulado?	A) Los movimientos, expresiones o rasgos de la persona parecen poco naturales. B) El contenido tiene muchas reproducciones o comentarios. C) La publicación aparece en una red social conocida. D) La persona que aparece en el contenido es reconocida por el usuario.
12	Prevención de ingeniería social	Recibe un audio aparentemente enviado por un familiar que le pide dinero con urgencia. ¿Cuál sería la opción más segura?	A) Enviar el dinero porque reconoce la voz del familiar. B) Responder al mismo número para confirmar la solicitud. C) Contactar al familiar por otro medio para confirmar que realmente realizó la solicitud. D) Compartir el audio con otras personas para que decidan si es verdadero.
13	Evaluación de autenticidad	Encuentra en redes sociales un video en el que una persona conocida realiza declaraciones que le parecen inusuales. ¿Qué sería más recomendable hacer?	A) Darlo por verdadero porque la persona aparece claramente en el video. B) Comprobar el origen del video y contrastar la información antes de creerlo o compartirlo. C) Considerarlo verdadero si muchas personas lo han compartido. D) Considerarlo auténtico si la imagen y el sonido tienen buena calidad.
14	Hábitos de verificación	Cuando recibe una noticia, imagen o video que podría ser importante, ¿qué hace habitualmente antes de compartirlo?	A) Lo comparto si parece verdadero. B) Compruebo quién lo publicó y contrasto la información con otras fuentes. C) Pregunto a familiares o amigos antes de compartirlo. D) Generalmente lo comparto sin realizar ninguna comprobación.
15	Uso de herramientas de verificación	Si tuviera dudas sobre la autenticidad de una imagen, video o audio, ¿qué haría para comprobarlo?	A) Buscaría información sobre su origen y utilizaría una herramienta de verificación digital. B) Revisaría únicamente cuántas personas lo han compartido. C) Consultaría solamente los comentarios de otros usuarios. D) Lo consideraría verdadero si tiene buena calidad de imagen o sonido.
16	Necesidad de capacitación	¿Qué tan útil sería para usted recibir una guía sencilla para aprender a identificar posibles deepfakes y utilizar herramientas gratuitas para verificarlos?	A) Muy útil. B) Útil. C) Poco útil. D) Nada útil.

Fuente: Elaboración propia.

Resultados y Discusión

Los resultados obtenidos mediante la aplicación del cuestionario permitieron caracterizar, en primer lugar, a los participantes según sus principales características sociodemográficas y patrones de uso de internet. Posteriormente, se analizaron los resultados relacionados con el conocimiento, la identificación, la prevención y la verificación de contenidos potencialmente manipulados mediante IA.

Tabla 2. *Distribución de los participantes según el rango de edad.*

Respuesta	Frecuencia	Porcentaje
18–24 años.	38	21,1 %
25–34 años.	46	25,6 %
35–44 años.	39	21,7 %
45–54 años.	32	17,8 %
55 años o más.	25	13,9 %
Total	180	100 %

Fuente: Elaboración propia.

La mayor proporción de participantes se ubicó entre los 25 y 34 años (25,6 %), seguida de los grupos de 35 a 44 años (21,7 %) y 18 a 24 años (21,1 %). En conjunto, el 68,4 % de los participantes tiene entre 18 y 44 años. La distribución obtenida permite conocer el nivel de conocimiento y las prácticas frente a los deepfakes en una población con representación de diferentes grupos etarios, lo cual aporta una visión más amplia para el diagnóstico.

Tabla 3. *Distribución de los participantes según su identificación.*

Respuesta	Frecuencia	Porcentaje
Masculino.	86	47,8 %
Femenino.	89	49,4 %
Otro.	2	1,1 %
Prefiero no responder.	3	1,7 %
Total	180	100 %

Fuente: Elaboración propia.

Los resultados muestran una distribución similar entre quienes se identificaron como femenino (49,4 %) y masculino (47,8 %). Las categorías otro y prefiero no responder representaron en conjunto el 2,8 %. Esta distribución contribuye a caracterizar la población estudiada y permite interpretar los resultados obtenidos a partir de una participación equilibrada en las categorías principales.

Tabla 4. *Nivel más alto de educación alcanzado por los participantes.*

Respuesta	Frecuencia	Porcentaje
Educación básica.	47	26,2 %
Bachillerato.	71	39,4 %
Educación superior.	52	28,9 %
Posgrado.	8	4,4 %
Doctorado.	2	1,1 %
Otro.	0	0 %
Total	180	100 %

Fuente: Elaboración propia.

El grupo predominante corresponde a los participantes con bachillerato (39,4 %), seguido por quienes poseen educación superior (28,9 %) y educación básica (26,2 %). En conjunto, estos tres niveles representan el 94,5 % de la población. Este resultado aporta información relevante para el diseño de la estrategia, debido a que evidencia la necesidad de utilizar contenidos comprensibles y adaptados a distintos niveles de formación.

Tabla 5. *Frecuencia de utilización de internet.*

Respuesta	Frecuencia	Porcentaje
Varias veces al día.	128	71,1 %
Una vez al día.	28	15,6 %
Varias veces por semana.	16	8,9 %
Ocasionalmente.	8	4,4 %
Nunca.	0	0,0 %
Total	180	100 %

Fuente: Elaboración propia.

El 71,1 % de los participantes utiliza internet varias veces al día y el 15,6 % lo hace una vez al día. Por tanto, el 86,7 % mantiene una interacción diaria con internet. Este resultado evidencia una alta exposición al entorno digital y, en consecuencia, una mayor posibilidad de estar en contacto con contenidos sintéticos o potencialmente manipulados, lo que justifica la pertinencia del estudio.

Tabla 6. *Dispositivo utilizado con mayor frecuencia para acceder a internet.*

Respuesta	Frecuencia	Porcentaje
Teléfono móvil.	139	77,2 %
Computadora.	29	16,1 %
Tableta.	7	3,9 %
Otro.	5	2,8 %
Total	180	100 %

Fuente: Elaboración propia.

El 77,2 % de los participantes accede principalmente a internet mediante teléfono móvil, mientras que el 16,1 % utiliza una computadora. El predominio del teléfono móvil permite identificar una característica importante de la población: la estrategia de prevención y verificación debe considerar recursos accesibles y funcionales desde este tipo de dispositivo.

Tabla 7. *Principal uso de internet por parte de los participantes.*

Respuesta	Frecuencia	Porcentaje
Redes sociales.	62	34,4 %
Comunicación y mensajería.	39	21,7 %
Trabajo o estudios.	35	19,4 %
Entretenimiento.	27	15,0 %
Compras o trámites.	11	6,1 %
Otro.	6	3,3 %
Total	180	100 %

Fuente: Elaboración propia.

Las redes sociales (34,4 %) constituyen el principal uso de internet, seguidas de la comunicación y mensajería (21,7 %) y del trabajo o estudios (19,4 %). En conjunto, estos resultados muestran que una parte importante de la actividad digital de los participantes se desarrolla en espacios de producción, recepción y circulación de información. Este hallazgo aporta elementos para enfocar la investigación en prácticas de identificación y verificación aplicables a estos contextos.

Tabla 8. *Conocimiento previo sobre el término deepfake.*

Respuesta	Frecuencia	Porcentaje
Sí, y sé qué significa.	32	17,8 %
Sí, pero sé poco sobre el tema.	46	25,6 %
He escuchado la palabra, pero no sé qué significa.	38	21,1 %
No, nunca la había escuchado.	64	35,6 %
Total	180	100 %

Fuente: Elaboración propia.

El 35,6 % de los participantes indicó que nunca había escuchado el término deepfake, mientras que el 25,6 % manifestó conocer poco sobre el tema. Solo el 17,8 % señaló conocer su significado. Estos resultados evidencian un nivel limitado de familiaridad con el fenómeno y aportan un diagnóstico inicial que justifica la inclusión de contenidos básicos de alfabetización sobre deepfakes en la estrategia propuesta.

Tabla 9. *Conocimiento conceptual sobre los deepfakes.*

Respuesta	Frecuencia	Porcentaje
Un contenido digital creado o modificado mediante inteligencia artificial para hacer que parezca real.	88	48,9 %
Un programa utilizado para proteger dispositivos frente a amenazas digitales.	22	12,2 %
Un contenido eliminado o bloqueado de internet.	31	17,2 %
Una aplicación utilizada para editar fotografías y videos.	39	21,7 %
Total	180	100 %

Fuente: Elaboración propia.

La definición correcta fue seleccionada por el 48,9 % de los participantes, mientras que el 51,1 % eligió una respuesta incorrecta. Aunque existe un nivel importante de reconocimiento conceptual, los resultados muestran que más de la mitad de la población no identifica correctamente el significado del término. Este hallazgo permite establecer la necesidad de reforzar la comprensión conceptual como punto de partida para la detección de contenidos sintéticos.

Tabla 10. *Conocimiento sobre los tipos de contenido en los que pueden aparecer deepfakes.*

Respuesta	Frecuencia	Porcentaje
Solo en videos.	29	16,1 %
En imágenes y videos.	41	22,8 %
En imágenes, videos y audios.	86	47,8 %
Solo en audios.	24	13,3 %
Total	180	100 %

Fuente: Elaboración propia.

El 47,8 % reconoció correctamente que los deepfakes pueden presentarse en imágenes, videos y audios. Sin embargo, el 52,2 % seleccionó respuestas que limitan el fenómeno a uno o dos tipos de contenido. Este resultado demuestra que el conocimiento sobre las diferentes manifestaciones de los deepfakes es incompleto, especialmente en relación con los audios sintéticos, aspecto relevante para la prevención de la suplantación de identidad.

Tabla 11. *Seguridad percibida para reconocer un posible deepfake.*

Respuesta	Frecuencia	Porcentaje
Muy seguro/a	14	7,8 %
Algo seguro/a	51	28,3 %
Poco seguro/a	72	40,0 %
Nada seguro/a	43	23,9 %
Total	180	100 %

Fuente: Elaboración propia.

El 40,0 % manifestó sentirse poco seguro/a y el 23,9 %, nada seguro/a. En conjunto, el 63,9 % presenta una baja confianza en su capacidad para reconocer un posible deepfake. Este resultado constituye uno de los principales aportes del diagnóstico, pues evidencia la necesidad de desarrollar competencias prácticas que permitan a los usuarios identificar señales de posible manipulación.

Tabla 12. *Identificación de posibles señales de manipulación.*

Respuesta	Frecuencia	Porcentaje
Movimientos, expresiones o rasgos poco naturales.	91	50,6 %
Muchas reproducciones o comentarios.	28	15,6 %
Publicación en una red social conocida.	33	18,3 %
La persona es reconocida por el usuario.	28	15,6 %
Total	180	100 %

Fuente: Elaboración propia.

La opción relacionada con movimientos, expresiones o rasgos poco naturales fue seleccionada por el 50,6 % de los participantes. Sin embargo, el 49,4 % eligió otros criterios que no constituyen evidencias confiables de autenticidad. Los resultados permiten identificar una comprensión parcial de las señales de manipulación y aportan elementos para orientar la estrategia hacia el análisis crítico de diferentes indicadores, evitando basarse únicamente en la apariencia o popularidad de un contenido.

Tabla 13. Acción ante una posible solicitud fraudulenta mediante audio.

Respuesta	Frecuencia	Porcentaje
Enviar el dinero porque reconoce la voz.	21	11,7 %
Responder al mismo número para confirmar.	37	20,6 %
Contactar al familiar por otro medio para confirmar.	102	56,7 %
Compartir el audio con otras personas.	20	11,1 %
Total	180	100 %

Fuente: Elaboración propia.

El 56,7 % seleccionó correctamente la opción de contactar al familiar por otro medio. Sin embargo, el 43,3 % adoptaría acciones potencialmente inseguras, como confiar en la

voz o responder al mismo número. Este resultado aporta evidencia sobre la existencia de vulnerabilidades frente a posibles casos de clonación de voz y permite identificar la necesidad de incorporar procedimientos de verificación alternativos.

Tabla 14. *Acciones para evaluar la autenticidad de un video.*

Respuesta	Frecuencia	Porcentaje
Darlo por verdadero porque la persona aparece claramente.	19	10,6 %
Comprobar el origen y contrastar la información.	111	61,7 %
Considerarlo verdadero si muchas personas lo compartieron.	27	15,0 %
Considerarlo auténtico por su calidad técnica.	23	12,8 %
Total	180	100 %

Fuente: Elaboración propia.

El 61,7 % señaló que comprobaría el origen del video y contrastaría la información, mientras que el 38,3 % se basaría en criterios como la apariencia, popularidad o calidad técnica. Aunque predomina una respuesta adecuada, los resultados evidencian que una proporción significativa aún utiliza criterios insuficientes. Esto aporta información para orientar la estrategia hacia la verificación del origen y el contexto de los contenidos.

Tabla 15. *Hábitos de verificación antes de compartir contenidos.*

Respuesta	Frecuencia	Porcentaje
Lo comparto si parece verdadero.	38	21,1 %
Compruebo quién lo publicó y contrasto la información.	79	43,9 %
Pregunto a familiares o amigos.	42	23,3 %
Generalmente lo comparto sin realizar comprobación.	21	11,7 %
Total	180	100 %

Fuente: Elaboración propia.

El 43,9 % indicó que comprueba quién publicó el contenido y contrasta la información con otras fuentes. No obstante, el 56,1 % restante señaló prácticas diferentes, entre ellas compartir contenidos si parecen verdaderos o consultar únicamente a personas cercanas. Este resultado revela que la verificación sistemática no constituye una práctica generalizada, lo que representa un aspecto relevante para la prevención de la difusión de contenido sintético.

Tabla 16. *Acciones para verificar la autenticidad de un contenido.*

Respuesta	Frecuencia	Porcentaje
Buscar el origen y utilizar una herramienta de verificación digital.	68	37,8 %
Revisar únicamente cuántas personas lo han compartido.	34	18,9 %
Consultar solamente los comentarios.	47	26,1 %
Considerarlo verdadero por la calidad de imagen o sonido.	31	17,2 %
Total	180	100 %

Fuente: Elaboración propia.

Solo el 37,8 % seleccionó la alternativa de buscar el origen del contenido y utilizar una herramienta de verificación digital. Por el contrario, el 62,2 % se basaría en elementos insuficientes para determinar su autenticidad. Este hallazgo constituye uno de los principales resultados del diagnóstico, debido a que identifica una necesidad concreta de capacitación sobre métodos y herramientas de verificación accesibles para los usuarios.

Tabla 17. *Percepción sobre la utilidad de recibir capacitación.*

Respuesta	Frecuencia	Porcentaje
Muy útil.	109	60,6 %
Útil.	51	28,3 %
Poco útil.	15	8,3 %
Nada útil.	5	2,8 %
Total	180	100 %

Fuente: Elaboración propia.

El 60,6 % de los participantes considera muy útil recibir capacitación y el 28,3 % la considera útil. En conjunto, el 88,9 % presenta una valoración favorable. Este resultado aporta un respaldo directo a la propuesta de investigación, pues evidencia una disposición positiva de la población hacia una estrategia orientada al fortalecimiento de sus conocimientos y capacidades para identificar y verificar posibles deepfakes. En este sentido, los resultados

obtenidos constituyen la base para la elaboración de la estrategia para la detección y prevención de deepfakes dirigida a usuarios web. Las necesidades identificadas en relación con el conocimiento del fenómeno, la capacidad para reconocer posibles señales de manipulación y el uso de procedimientos de verificación permiten orientar la estrategia hacia contenidos y acciones acordes con las características de la población estudiada.

Estrategia para la detección y prevención de deepfakes dirigida a usuarios web

Tabla 18. *Estrategia práctica para la identificación y prevención de deepfakes mediante dispositivos móviles.*

Etapa	Acción del usuario	Señales o aspectos que debe comprobar	Contenido	Herramienta móvil recomendada	Procedimiento sencillo	Resultado esperado
Pausar	Detenerse antes de creer, responder o compartir el contenido.	Mensajes urgentes, alarmistas, solicitudes de dinero, códigos, datos personales o acciones inmediatas.	Imagen, video y audio	No requiere herramienta.	Esperar y analizar el contenido antes de realizar cualquier acción.	Evitar decisiones impulsivas y posibles fraudes.
Observar	Examinar cuidadosamente el contenido recibido.	Rostros o expresiones poco naturales, movimientos extraños, alteraciones en ojos, dientes o manos, iluminación y sombras inconsistentes, así como desincronización entre labios y voz.	Imagen y video	Galería del dispositivo.	Ampliar la imagen y reproducir el video varias veces para identificar posibles irregularidades.	Reconocer señales iniciales que justifiquen una verificación adicional.
Comprobar el origen	Identificar la fuente y el contexto de publicación.	Cuenta desconocida, fuente no identificada, información sin referencias, fecha dudosa o contenido presentado fuera de contexto.	Imagen, video y audio	Google, Brave, Safari u otro navegador.	Buscar el contenido, autor, acontecimiento o información principal en Internet.	Localizar la fuente original y conocer el contexto del contenido.
Verificar la imagen	Comprobar la procedencia y el contexto de la imagen.	Coincidencias con imágenes anteriores, utilización en otros acontecimientos o diferencias entre versiones.	Imagen	Google Lens.	Cargar la imagen o una captura de pantalla y revisar imágenes similares y sitios web relacionados.	Determinar si la imagen corresponde al contexto en que fue compartida.
Analizar el contenido multimedia	Someter el contenido a una herramienta especializada de análisis.	Posibles alteraciones visuales, faciales o de audio que no puedan determinarse mediante observación directa.	Imagen, video y audio	DeepFake-O-Meter, Resemble AI o TruthScan.	Cargar el archivo disponible y consultar el resultado proporcionado por la herramienta.	Obtener evidencia adicional sobre la posible manipulación del contenido.
Contrastar	Comparar la información con fuentes independientes y confiables.	Diferencias en fechas, nombres, lugares, declaraciones o acontecimientos; ausencia de información en fuentes confiables.	Imagen, video y audio	Google, Brave, Safari u otro navegador.	Realizar búsquedas y comparar la información con dos o más fuentes independientes.	Confirmar, contextualizar o cuestionar la información recibida.
Confirmar la identidad	Verificar directamente la identidad de la persona involucrada.	Solicitudes de dinero, transferencias, códigos de seguridad, contraseñas o información privada.	Audio y video	Llamada telefónica o WhatsApp.	Contactar a la persona mediante un canal diferente al utilizado para recibir el contenido.	Confirmar la autenticidad de la persona y de la solicitud.
Decidir	Tomar una decisión considerando todas las evidencias obtenidas.	Resultado sospechoso de una herramienta, falta de fuente confiable, contradicciones o persistencia de dudas.	Imagen, video y audio	Herramientas utilizadas anteriormente.	Si no existe evidencia suficiente para confirmar su autenticidad, evitar compartir el contenido y, cuando corresponda, reportarlo.	Reducir la difusión de contenido falso o manipulado.

Fuente: Elaboración propia.

En relación con el conocimiento sobre los deepfakes, los resultados muestran importantes brechas. El 35,6 % de los participantes nunca había escuchado el término y solo el 17,8 % manifestó conocer su significado. Asimismo, el 51,1 % no identificó correctamente su definición y el 52,2 % desconoció que estos contenidos pueden presentarse simultáneamente en imágenes, videos y audios. Estos resultados coinciden con Dhahir et al. (2024), quienes encontraron niveles bajos de capacidad para identificar deepfakes y

señalaron la necesidad de fortalecer la alfabetización digital mediante conocimientos técnicos y pensamiento crítico. En este sentido, el desconocimiento conceptual constituye una limitación inicial para desarrollar procesos adecuados de identificación y verificación.

La capacidad percibida para reconocer contenidos manipulados también representa una dificultad. El 63,9 % de los participantes manifestó sentirse poco o nada seguro para identificar un posible deepfake, mientras que solo el 50,6 % reconoció como señal de posible manipulación la presencia de movimientos, expresiones o rasgos poco naturales. Este resultado es consistente con la revisión y metaanálisis de Diel et al. (2024) quienes determinaron que el rendimiento humano en la detección de deepfakes es limitado y, en términos generales, no supera ampliamente el nivel esperado por azar.

Por tanto, los resultados obtenidos sugieren que confiar exclusivamente en la percepción visual o en la intuición del usuario no constituye un mecanismo suficiente para determinar la autenticidad de un contenido. Asimismo, los resultados evidencian que las dificultades no se limitan a los contenidos visuales, sino que incluyen los audios sintéticos, debido a que un 43,3 % adoptaría acciones potencialmente inseguras, como confiar en la voz reconocida o responder al mismo número. Esto es tratado por Blancaflor et al. (2024) donde se analizó precisamente el uso de software y servicios de clonación de voz para engañar a víctimas mediante conversaciones telefónicas, evidenciando que la imitación de una voz conocida puede utilizarse como mecanismo de suplantación. En cuanto a las prácticas de verificación, el 62,2 % recurrió a criterios insuficientes, como la cantidad de compartidos, los comentarios o la calidad de imagen y sonido.

Estos resultados determinan que la popularidad, la apariencia y la calidad técnica continúan siendo utilizadas como indicadores de autenticidad, pese a que no garantizan que un contenido sea genuino. Situación que coincide con Momin et al. (2025) quienes destacan la necesidad de desarrollar mecanismos de detección comprensibles y herramientas que permitan a los usuarios interpretar los resultados de manera adecuada.

El 88,9 % de los participantes considera útil o muy útil recibir capacitación sobre la identificación y prevención de deepfakes. Esto respalda la pertinencia de la estrategia propuesta, especialmente porque las principales necesidades identificadas corresponden precisamente a conocimiento conceptual, reconocimiento de señales de manipulación, prevención de la clonación de voz y aplicación de procedimientos de verificación. En función de estos hallazgos, la estrategia planteada mediante las etapas de pausar, observar, comprobar el origen, verificar, analizar, contrastar, confirmar la identidad y decidir, busca transformar la identificación de deepfakes de una actividad basada principalmente en la intuición hacia un proceso sistemático de análisis y verificación.

Conclusiones

Los resultados obtenidos muestran que los participantes mantienen una alta interacción con internet, principalmente a través de teléfonos móviles y redes sociales, lo que los mantiene en constante contacto con contenidos digitales que pueden ser manipulados mediante inteligencia artificial. A pesar de esta frecuente exposición, se identificó un conocimiento limitado sobre los deepfakes, tanto en su definición como en las diferentes formas en que pueden presentarse. Asimismo, una proporción importante de los participantes manifestó poca seguridad para

reconocer este tipo de contenidos y aún considera aspectos como la apariencia, la cantidad de comentarios o la calidad de una imagen o video como elementos para determinar su autenticidad. También se evidenciaron dificultades para responder adecuadamente ante posibles casos de clonación de voz y para aplicar procedimientos de verificación antes de compartir contenidos. Estos resultados muestran que conocer o sospechar que un contenido puede ser falso no siempre se traduce en una acción adecuada para comprobar su autenticidad.

Por otra parte, el 88,9 % de los participantes consideró útil o muy útil recibir capacitación sobre los deepfakes, lo que demuestra una disposición favorable para fortalecer sus conocimientos y habilidades. Para atender las necesidades identificadas, se diseñó una estrategia para la detección y prevención de deepfakes dirigida a usuarios web, estructurada en ocho etapas: pausar, observar, comprobar el origen, verificar, analizar, contrastar, confirmar la identidad y decidir.

Esta estrategia busca orientar a los usuarios mediante acciones sencillas y prácticas que puedan realizarse principalmente desde el teléfono móvil, promoviendo un proceso de verificación que no dependa únicamente de la percepción. De esta manera, la propuesta contribuye al fortalecimiento de las competencias necesarias para analizar los contenidos digitales de forma crítica y tomar decisiones más responsables antes de recibir, verificar y compartir información en internet. La estrategia para la detección y prevención de deepfakes puede utilizarse como una guía práctica para fortalecer las capacidades de los usuarios web. Su aplicación debe considerar las ocho etapas propuestas: pausar, observar, comprobar el origen, verificar, analizar, contrastar, confirmar la identidad y decidir. Las actividades de capacitación deben abordar de manera sencilla las diferentes formas de deepfakes, mediante ejemplos de imágenes, videos y audios, y promover prácticas de verificación basadas en fuentes confiables y el contraste de información.

En situaciones de posible clonación de voz, es importante confirmar la identidad mediante un canal diferente antes de enviar dinero, códigos o información personal. Las herramientas digitales pueden utilizarse como apoyo, complementándolas con el análisis del contexto y la información disponible. Los materiales de la estrategia deben ser accesibles desde teléfonos móviles y emplear un lenguaje claro, considerando las características de la población estudiada. Finalmente, la estrategia debe evaluarse después de su aplicación y actualizarse periódicamente debido a la evolución de las tecnologías de IA.

Agradecimiento

En primer lugar, agradezco a Dios por brindarme la fortaleza, sabiduría y perseverancia necesarias para culminar esta etapa tan importante de mi formación académica. A mi esposa, por creer en mí incluso cuando yo mismo dudaba de mis capacidades, por impulsarme a continuar con mis estudios cuando las circunstancias y las dificultades me hacían pensar en rendirme. Gracias por estar a mi lado, por sus palabras de ánimo, por su paciencia y por recordarme que sí podía lograrlo cuando nadie más me impulsaba a seguir.

Este logro también es de ella, porque detrás de cada paso que di estuvo su apoyo incondicional. A mi hijo, por ser esa razón que me motivó a luchar cada día por un mejor futuro. Todo esfuerzo, cada sacrificio y cada momento de cansancio tuvieron sentido al pensar en él y en

el ejemplo que quiero dejarle. Si hoy estoy culminando esta etapa, es porque tuve una familia que me dio motivos para no rendirme. A mis padres, por sus consejos y enseñanzas, que me han permitido seguir adelante y no rendirme ante las dificultades. De manera especial, expreso mi sincero agradecimiento a mi tutora, por su orientación, dedicación, conocimientos y apoyo durante la elaboración de este artículo científico. Sus recomendaciones y acompañamiento fueron fundamentales para poder llevar a cabo y culminar este trabajo.

Referencias Bibliográficas

Alrashoud, M. (2025). Métodos, enfoques y desafíos para la detección de vídeo deepfake. *Alexandria Engineering Journal*, 125, 265–277. https://doi.org/10.1016/j.aej.2025.04.007

Alzaga, A. (2025). La metamorfosis de la verdad: Deepfakes y el desafío de la autenticidad en la sociedad digital. *Derecom. Revista Internacional de Derecho de la Comunicación y de las Nuevas Tecnologías*, 38(1), 45-57. https://doi.org/10.5209/dere.102344

Babaei, R., Cheng, S., Duan, R. & Zhao, S. (2025). Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis. *Journal of Sensor and Actuator Networks*, 14(1), 17. https://doi.org/10.3390/jsan14010017

Blancaflor, E., Abaleta, R., Achacoso, L., Amper, A. & Ampiloquio, P. (2024). Emerging threat: The use of AI voice cloning software and services to deceive victims through phone conversations and its potential effects on the Filipino population. *Proceedings of the 2024 5th Asia Service Sciences and Software Engineering Conference*, 137–146. https://doi.org/10.1145/3702138.3702145

Buchan, M., Bhawra, J. & Katapally, T. R. (2024). Navigating the digital world: Development of an evidence-based digital literacy program and assessment tool for youth. *Smart Learning Environments*, 11, 8. https://doi.org/10.1186/s40561-024-00293-x

Calabrese, C., Rasul, M., Cho, H., Jeon, M., Kim, T. & Oh, Y. (2026). Development and validation of a deepfake literacy scale: Predictors and correlates of deepfake literacy. *SSRN*. https://doi.org/10.2139/ssrn.6517436

Cruz, C. (2026). La manipulación de la evidencia de políticas públicas con la inteligencia artificial generativa: Los riesgos de los deepfakes. *Universitas*, 43. https://orcid.org/0000-0002-6105-9167

Colunga, R. & González, L. (2025). Perspectiva Teórica de la Alfabetización Digital: Aportes de Heidegger, Freire, Habermas y Vygotsky. *Revista Latinoamericana De Calidad Educativa*, 2(3), 51-54. https://doi.org/10.70625/rlce/278

Dhahir, D., Kenda, N., Dirgahayu, D., Syarifuddin, S., Djaffar, R., Widihastuti, R., Rustam, M. & Pala, R. (2024). The relationship of digital literacy, exposure to AI-generated deepfake videos, and the ability to identify deepfakes in Generation X. *Jurnal Pekommas*, 9(2), 357–368. https://doi.org/10.56873/jpkm.v9i2.5873

Di Gionantonio, A. & Ligorria, L. (2022). Ingeniería Social: Ataques y Estrategias de Defensa. *Revista Tecnología Y Ciencia*, (22), 51–58. https://rtyc.utn.edu.ar/index.php/rtyc/article/view/927

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538. https://doi.org/10.1016/j.chbr.2024.100538

Fernández, Á., Yazidi, A., Vasilakos, A., Haugerud, H. & Youce, D. (2024). Deepfakes: tendencias actuales y futuras. *Artif Intell Rev 57*, 64. https://doi.org/10.1007/s10462-023-10679-x

García, D., Suárez, R. & García, A. (2025). "Todo parece veraz". Credibilidad de la desinformación producida usando IA desde la perspectiva de los estudiantes de comunicación en España. *Revista De*

Comunicación, 24(2), 183-227. https://doi.org/10.26441/RC24.2-2025-3872

Ghiurău, D. & Popescu, D. (2025). Distinguir la realidad de la IA: enfoques para detectar contenido sintético. *Ordenadores*, 14(1), 1. https://doi.org/10.3390/computers14010001

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

Javed, M., Zhang, Z., Dahri, F. & Kumar, T. (2025). Enhancing multimodal deepfake detection with local–global feature integration and diffusion models. *Signal, Image and Video Processing*, 19, 400. https://doi.org/10.1007/s11760-025-03970-7

Jheelan, J. & Pudaruth, S. (2025). Uso de aprendizaje profundo para identificar deepfakes creados mediante redes generativas adversariales. *Ordenadores*, 14(2), 60. https://doi.org/10.3390/computers14020060

Li, M., Ahmadiadli, Y. & Zhang, X. (2025). A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7), 165, 1–38. https://doi.org/10.1145/3714458

Ma, R., Wang, X. & Yang, G. (2025). Fighting fake news in the age of generative AI: Strategic insights from multi-stakeholder interactions. *Technological Forecasting and Social Change*, 216, 124125. https://doi.org/10.1016/j.techfore.2025.124125

Momin, M., Sufian, A., Barman, D., Leo, M., Distante, C. & Damer, N. (2025). Explainable deepfake detection across different modalities: An overview of methods and challenges. *Image and Vision Computing*, 163, 105738. https://doi.org/10.1016/j.imavis.2025.105738

Navas, C. & Pérez, S. (2023). Mapping the Intellectual Structure of the Social Engineering: a Co-citation and Bibliographic Coupling Study. *INVESTIGATIO*, 1(20). https://doi.org/10.31095/investigatio.2023.20.10

Park, P., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 100988. https://doi.org/10.1016/j.patter.2024.100988

Pennycook, G. & Rand, D. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Plohl, N., Mlakar, I., Aquilino, L., Bisconti, P. & Smrke, U. (2025). Development and validation of the perceived deepfake trustworthiness questionnaire (PDTQ) in three languages. *International Journal of Human–Computer Interaction*, 41(11), 6786–6803. https://doi.org/10.1080/10447318.2024.2384821

Quiliche, D., Mauricio, K. & Mendoza de los Santos, A. (2026). Deepfakes y autenticación biométrica: Desafíos y soluciones contra la suplantación de identidad. *Revista Científica Unanchay*, 5(1). https://doi.org/10.64424/rcu51202698

Ramos, F. (2024). Deepfake: Análisis de sus implicancias tecnológicas y jurídicas en la era de la Inteligencia Artificial. Derecho global. *Estudios sobre derecho y justicia*, 9(27), 359-387. https://doi.org/10.32870/dgedj.v9i27.754

Rzayeva, L., Zhetpisbayeva, A., Batkuldin, A., Nyssanov, N., Ryzhova, A. & Saeed, F. (2026). Un método inteligente de historial forense de navegador para el análisis automatizado de registros de actividad web, credenciales y perfiles de comportamiento de usuarios. *Algoritmos*, 19(1), 75. https://doi.org/10.3390/a19010075

Tiernan, P., Costello, E., Donlon, E., Parysz, M. & Scriney, M. (2023). Alfabetización en la información y los medios en la era de la IA: opciones para el futuro. *Ciencias de la Educación*, 13(9), 906. https://doi.org/10.3390/educsci13090906

Vamsi, V., Shet, S., Reddy, S., Rose, S., Shetty, S., Sathvika, S. & Shankar, S. (2022). Deepfake detection in digital media forensics. *Global Transitions Proceedings*,

3(1), 74–79. https://doi.org/10.1016/j.gltp.2022.04.017

Villamar, M., Vera Pico, L., Castillo Pinargote, M., Reyes Romero, D. & López Gorozabel, O. (2026). La incidencia de la inteligencia artificial generativa en los deepfakes y sus repercusiones en la confianza social. *Ciencia y Educación*, 7(1.1), 677 - 685. https://doi.org/10.5281/zenodo.18463188

Wang, T., Liao, X., Cena, K. P., Lin, X. & Wang, Y. (2025). Deepfake detection: A comprehensive survey from the perspective of reliability. *ACM Computing Surveys*, 57(3), 1–35. https://doi.org/10.1145/3699710

Waseem, S. & Elneel, M. (2023). Deepfake face swapping and expression: A review. *IEEE Access*, 11, 117865–117906. https://doi.org/10.1109/ACCESS.2023.3324403

Zambrano, A. (2024). Impacto de la inteligencia artificial en los ciberataques. *Revista Científica Sinapsis*, 24(1). https://doi.org/10.37117/s.v24i1.895

Declaraciones éticas y editoriales del artículo
Contribución de los autores (Taxonomía CRediT) Pedro Antonio Cevallos Ortiz: conceptualización de la investigación, diseño metodológico, desarrollo del proceso investigativo, análisis formal de los datos, redacción del borrador original del manuscrito, revisión crítica del contenido científico y supervisión general del estudio. Marlene Guadalupe Castillo Pinargote: curación y organización de los datos, participación en la recolección de información, validación de los resultados obtenidos y elaboración de representaciones gráficas y visualización de los datos.
Declaración de conflicto de intereses Los autores declaran que no existe conflicto de intereses en relación con la investigación presentada, la autoría del manuscrito ni la publicación del presente artículo.
Declaración de financiamiento La presente investigación no recibió financiamiento específico de agencias públicas, comerciales o de organizaciones sin fines de lucro. En caso de existir financiamiento institucional o externo, este deberá ser declarado explícitamente por los autores en esta sección.
Declaración del editor El editor responsable certifica que el proceso editorial del presente artículo se desarrolló conforme a los principios de integridad científica, transparencia y buenas prácticas editoriales. El manuscrito fue sometido a un proceso de evaluación mediante revisión por pares doble ciego, garantizando la confidencialidad de la identidad de los autores y revisores durante todo el proceso de dictamen académico. Asimismo, el editor declara que el artículo cumple con los criterios científicos, metodológicos y éticos establecidos por la revista.
Declaración de los revisores Los revisores externos que participaron en la evaluación del presente manuscrito declaran haber realizado el proceso de revisión de manera objetiva, independiente y confidencial. Asimismo, manifiestan que no mantienen conflictos de interés con los autores ni con la investigación evaluada, y que sus observaciones y recomendaciones se fundamentan exclusivamente en criterios científicos, metodológicos y académicos.
Declaración ética de la investigación Los autores declaran que la investigación se desarrolló respetando los principios éticos de la investigación científica, garantizando la confidencialidad de los datos y el respeto a los participantes del estudio. En los casos en que la investigación involucre seres humanos, los procedimientos deben ajustarse a los principios éticos establecidos en la Declaración de Helsinki y a las normativas institucionales correspondientes.
Declaración sobre el uso de inteligencia artificial Los autores declaran que el uso de herramientas de inteligencia artificial, en caso de haberse utilizado durante el proceso de investigación o redacción del manuscrito, se realizó únicamente como apoyo técnico para mejorar la claridad del lenguaje o el análisis de información, manteniendo siempre la responsabilidad intelectual sobre el contenido del artículo. Las herramientas de inteligencia artificial no fueron utilizadas como autoras del manuscrito ni sustituyen la responsabilidad académica de los investigadores.
Disponibilidad de datos Los datos que respaldan los resultados de esta investigación estarán disponibles previa solicitud razonable al autor de correspondencia, respetando las normas éticas y de confidencialidad establecidas por la investigación.