Revista Facultad de Ingeniería

# Comparative Review of Deep Learning Models for the Voice Deepfakes Detection in Bank Fraud Scenarios

## Revisión Comparativa de Modelos de Aprendizaje Profundo para la Detección de Deepfakes de Voz en Fraudes Bancarios

**Iarley Felipe Granobles Prieto, https://orcid.org/0009-0002-2896-9254**
**Universidad de Cundinamarca**
**Autopista Chía - Cajicá | Sector "El Cuarenta"**
**Chía, Colombia**

**KEYWORDS**
Deep Learning, Deepfake, Bank Fraud, Financial Security
Aprendizaje Profundo, Deepfake, Fraude Bancario, Seguridad Financiera

**ABSTRACT:**

The article presents a systematic review guided by the PRISMA 2020 guidelines, focused on the analysis of the state of the art of voice detection generated by deep learning models applied to bank fraud. Similarly, the evolution of generative artificial intelligence techniques is analyzed, a tool used for the creation of voice *deepfakes*, which in turn represent a significant risk to financial digital security, since it facilitates the falsification of identities and compromises sensitive processes that involve voice verification.

In the architectures studied, convolutional networks and Transformers stand out for their ability to identify patterns and adapt to various contexts. However, its practical application mostly results in a decrease in its performance, which is associated with factors such as acoustic noise, environmental variability and the possible disconnect between training data and the specific characteristics of the banking context.

The article proposes a conceptual model based on prioritized attention and segmented audio analysis, with the aim of optimizing the computational resources used in the process. In this way, it seeks to improve the accuracy, efficiency and reliability of real-time verification systems to contribute to the strengthening of cybersecurity and in the same way, the prevention of fraud through artificial intelligence in the banking sector.

**RESUMEN:**

El artículo presenta una revisión sistemática guiada por los lineamientos PRISMA 2020, centrada en el análisis del estado del arte de la detección de voces generadas por modelos de aprendizaje profundo

aplicadas al fraude bancario. De igual manera, se analiza la evolución de las técnicas de inteligencia artificial generativa, herramienta que se usa para la creación de *deepfakes* de voz, que a su vez representan un riesgo significativo para la seguridad digital financiera, ya que facilita la falsificación de identidades y compromete procesos sensibles que involucren verificación por voz.

En las arquitecturas estudiadas, se destacan las redes convolucionales y las Transformers por su capacidad para identificar patrones y adaptarse a diversos contextos. Sin embargo, su aplicación práctica se produce en su gran mayoría una disminución en su rendimiento, que se asocia a factores como el ruido acústico, la variabilidad del entorno y la posible desconexión entre los datos de entrenamiento y las características específicas del contexto bancario.

El artículo propone un modelo conceptual basado en atención priorizada y análisis segmentado en audio, con el objetivo de optimizar los recursos computacionales usados en el proceso. De esta manera, se busca mejorar la precisión, eficiencia y confiabilidad de los sistemas de verificación en tiempo real para contribuir al fortalecimiento de la ciberseguridad y de igual manera, la prevención del fraude mediante inteligencia artificial en el sector bancario.

## 1. Introducción

El avance de la inteligencia artificial (IA) ha impulsado el desarrollo de técnicas para la generación de contenido sintético [1]. En consecuencia, esta tecnología ha transformado la manera en que las organizaciones manejan la información y automatizan los métodos. En el caso del sector financiero se ha hecho uso de herramientas basadas en IA generativa, para optimizar procesos y mejorar la experiencia del cliente. Sin embargo, el uso indebido de esta tecnología ha llevado a una nueva generación de amenazas y vulnerabilidades que no sean fáciles de detectar para los usuarios y con un alto impacto económico.

La IA generativa alcanzó un nivel de sofisticación al punto de poder imitar patrones y capacidades que se consideraban exclusivamente humanas, como hablar o escribir. Estas tecnologías benefician la automatización de procesos bancarios, pero de igual manera exponen las entidades a nuevas formas de fraude. “Así mismo, se presenta el riesgo asociado a la autorización indebida de pagos a causa de la suplantación de identidad de la persona titular de los servicios financieros, así como la suplantación de directivos dentro de las entidades.” [2]; por lo tanto, el sector bancario tiene el reto de mantener un nivel de innovación tecnológica sin afectar la seguridad en las operaciones, ya que de esta depende la confianza de los usuarios.

Existen casos en los que se ha demostrado que las amenazas tienen una base real y tangible. “A principios de 2020, en Hong Kong, un gerente bancario recibió lo que creyó que era una llamada de voz del director de una empresa cuya voz reconoció; este director llamó y le pidió un favor de que autorizara transferencias por valor de 35 millones de dólares” [3]. Este hecho confirma que la clonación de voz apoyada por modelos de IA generativa puede engañar a los sistemas de verificación que se tienen actualmente, causando pérdidas millonarias a bancos y clientes, disminuyendo la confianza de los usuarios en las entidades y desencadenando problemas económicos de gravedad.

En este contexto, el sector bancario reconoce que es necesaria una reestructuración a las estrategias de ciberseguridad actuales para dar frente al fraude potenciado por IA. Es fundamental que las entidades bancarias orienten los recursos en lidiar con el fraude estimulado por la IA para conservar o conseguir una posición en el área financiera [4]. Por tanto, la detección de *deepfakes* se ha convertido en una prioridad estratégica en las entidades financieras, ya que abarca dimensiones tecnológicas, económicas y éticas de las que depende la confianza del usuario y los servicios que pueden ofrecer las entidades bancarias.

La literatura actual se limita a describir los métodos y recursos existentes para abordar esta problemática, pero carece de comparaciones entre estas técnicas; además, no se ha hecho una evaluación en entornos reales ni alguna aplicabilidad concreta. Asimismo, no se dispone de estudios que evalúen el desempeño entre los diferentes modelos de aprendizaje, ni de investigaciones que integren datos reales con las soluciones propuestas para generar una evaluación de rendimiento en entornos no controlados.

Los aportes esperados se centran en ofrecer una revisión sistemática de las arquitecturas de aprendizaje profundo usadas para la detección de *deepfakes* de voz, identificando capacidades y limitaciones que presentan al momento de aplicarse en el sector bancario. A nivel local, se busca responder a los crecientes casos de fraude por medio de suplantación de identidad reportados en Colombia, donde los bancos se enfrentan a la necesidad de resguardar la integridad y garantizar la verificación de identidad en transacciones remotas. De igual forma, a nivel global, el trabajo se alinea con los desafíos en cuanto al avance de los ataques de ingeniería social mediante IA generativa de voz. En este contexto, el artículo se alinea con los principios del Modelo Educativo Digital Transmoderno (MEDIT), estableciendo un vínculo entre las problemáticas del sector financiero local y los retos de la seguridad bancaria de la sociedad digital transmoderna.

El artículo tiene como objetivo analizar y diferenciar teóricamente las arquitecturas de aprendizaje profundo, que se hayan diseñado para la detección de voces generadas de manera artificial, enfocándose en la viabilidad como herramienta para la prevención de fraudes en contextos bancarios, especialmente los que involucran comunicaciones que tengan dentro contenido crítico, como solicitudes de transferencias, autorizaciones de pagos o acceso a información sensible de los usuarios. Además, se plantea examinar los retos técnicos y funcionales que surgen cuando estos sistemas se implementan en entornos reales. De esta manera, el artículo busca responder ¿Qué modelos de aprendizaje profundo resultan más eficaces para detectar voces sintéticas en escenarios de fraude bancario, y cuáles son sus limitaciones prácticas principales?

## 2. Metodología

La presente revisión sistemática se diseñó con un enfoque cualitativo, esto con el objetivo de comprender, comparar e interpretar el estado del arte sobre los modelos propuestos recientemente, su aplicabilidad y el desempeño que tienen en escenarios reales de fraude bancario por suplantación de voz. Este tipo de estudio requiere una aproximación interpretativa y comprensiva de los hallazgos de diferentes estudios relacionados al tema. De esta manera, se llevó a cabo un registro riguroso y trazable del proceso de revisión de los proyectos precedentes y sus avances para abordar la problemática propuesta.

De acuerdo con Siddaway et al. [5], las revisiones sistemáticas requieren de un proceso que sea metófico y reproducible. De igual manera, exige de búsquedas profundas en la literatura y un resumen crítico de lo que se haya encontrado respecto a lo que sea relevante para el artículo. Los lineamientos PRISMA 2020 tienen como objetivo identificar, seleccionar, evaluar y sintetizar estudios relevantes para el proyecto en cuestión [6], [7]. De este modo, se asegura que el análisis cuente con una base sólida y cumpla con los estándares de transparencia y consistencia metodológica. Estos factores permiten validar los resultados y respaldar futuras investigaciones con un fundamento científico que sea confiable.

### 2.1. Estrategia de busqueda

La busqueda de literatura se ejecutó en múltiples bases de dato académicas reconocidas como Scopus, IEEE Xplore, ScienceDirect, J-Gate, McGraw-Hill e IC Editorial, considerando únicamente artículos publicados en el periodo de 2019-2025, que estuvieran redactados en inglés o español, limitados a investigaciones originales revisadas por pares y relacionadas directamente con el tema de estudio o que presentaron un aporte significativo para el desarrollo del artículo. Las palabras clave empleadas combinaron términos en español e inglés; la estrategia combinó descriptores técnicos en inglés mediante operadores booleanos. Los términos clave incluyeron frases literales como “deepfake audio”, “synthetic voice”, “voice spoofing detection”, “AI-generated speech”, “audio forgery”, junto con conceptos de aprendizaje profundo y fraude (“deep learning”, “fraud detection”).

Para la gestión bibliográfica y trazabilidad de las decisiones durante el proceso de cribado y selección, se utilizó la herramienta Zotero, que permitió organizar los estudios por etapas, registrar motivos de exclusión y el mantenimiento de la transparencia y rigor metodológico.

### 2.1. Estrategia de búsqueda

Para asegurar una pertinencia en el análisis, se establecieron criterios rigurosos de selección basados en la pregunta de investigación.

- Artículos académicos originales publicados entre 2019. 2025.
- Revisiones por pares en revistas o conferencias de alto impacto
- Relevancia directa con escenarios de fraude o seguridad, especialmente en el ámbito bancario y financiero, así como evaluaciones de robustez frente a *deepfakes* empleados para la suplantación de identidad o manipulación de transacciones.

De este modo, se enfoca la revisión en fuentes pertinentes que aborden arquitecturas de redes neuronales y enfoques de aprendizaje profundo, dando prioridad a los estudios que traten la detección de fraudes en entornos financieros.

### 2.1. Proceso de selección de estudios

El proceso siguió las etapas del diagrama de flujo PRISMA. En un principio, la estrategia de búsqueda arrojó un número total de registros que se gestionó por etapas. La población de esta revisión corresponde

a artículos académicos publicados entre 2019 y 2025, centrados en técnicas de detección de voces sintéticas generadas por IA. Específicamente mediante modelos de aprendizaje profundo.

En una primera fase se realizó la identificación de registros de las bases de datos seleccionadas. Durante este proceso se efectuó una depuración de registros duplicados. Posteriormente, se evaluaron los títulos y los resúmenes de estos para descartar trabajos que no presentaban relación con la temática. Los textos que superaron esta etapa fueron leídos íntegramente aplicando los criterios de inlusión/exlusión detallados anteriormente. Finalmente, se incluyeron los estudios que cumplían con todos los criterios, resultando en 27 artículos que aportaron al presente artículo.

El proceso seguido se representa en la **Figura 1** mediante el diagrama de flujo PRISMA, que apoya en la ilustración del proceso de selección con las diferentes etapas, proporcionando una representación clara del número de registros identificados, excluidos o incluidos en cada una de ellas [8]. Asimismo, el diagrama testifica un seguimiento documentado de cada artículo viable desde la caracterización hasta la posible inclusión en el análisis final. Por consiguiente, se obtiene una visión transparente y sistemática del procedimiento de revisión, permitiendo evaluar el rigor del proceso de selección y la validez de los resultados.

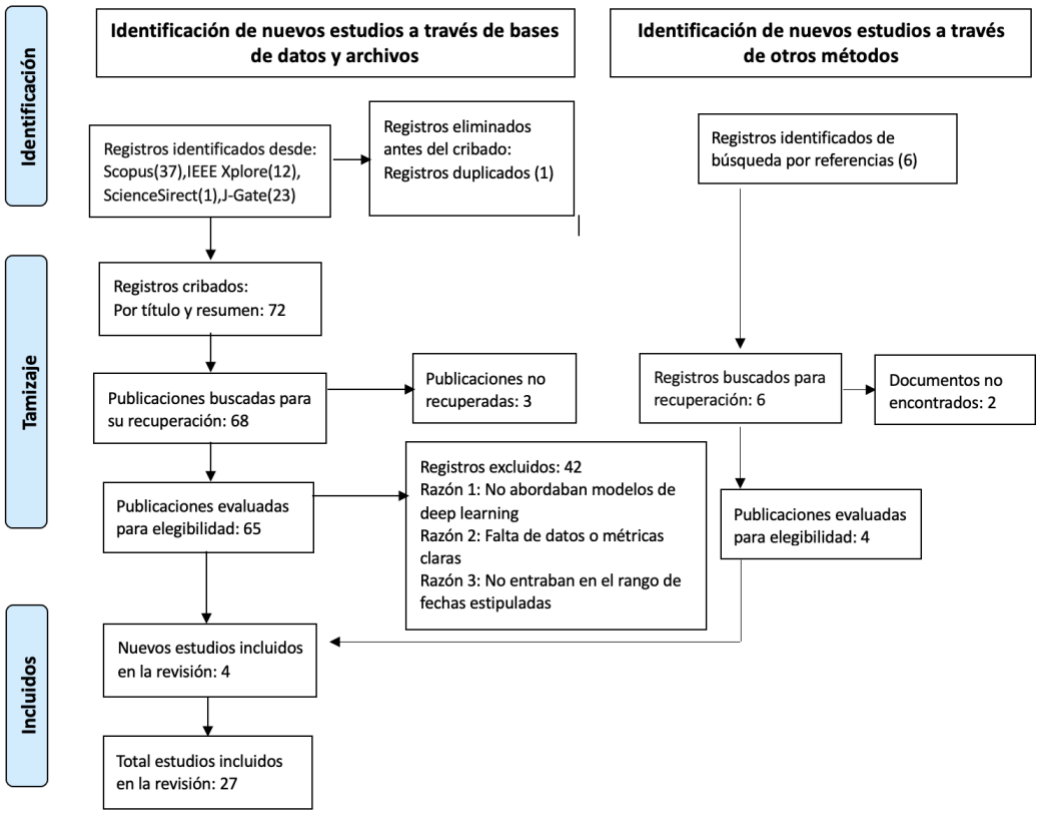


**Figure 1** Diagrama de flujo PRISMA

## 3. Resultados

Al aplicar el proceso de cribado sistemático, se seleccionaron finalmente 27 estudios que cumplen con los criterios de inclusión establecidos en el apartado metodológico. Estos trabajos se centraban en la detección de voces sintéticas generadas por IA con fines maliciosos, sin profundizar en el sector bancario. De esta manera, se identificaron las principales tendencias y limitaciones actuales en los sistemas detección y reconocer las posibles áreas de mejora en la literatura actual.

La detección de *deepfakes* es un campo en desarrollo dentro del aprendizaje automático, en el cual persisten debates sobre la efectividad y generalización de los modelos que se han propuesto [9], destacan la diversidad de enfoques fundamentados en aprendizaje profundo, entre los que se encuentran las redes convolucionales, las redes recurrentes y las arquitecturas de atención, todas orientadas a la identificación de señales manipuladas en su plenitud. Estos modelos procesan el espectrograma del audio para identificar diferencias por más sutiles que sean, entre voces auténticas y sintéticas desde un punto de vista técnico.

De acuerdo con los avances, los detectores de voces sintéticas suelen basarse en principios de aprendizaje profundo, examinando diversas arquitecturas y sus variantes más eficaces de acuerdo con las métricas dispuestas para cada ejercicio[10], [11], [12]. Entre ellas, sobresale la tendencia de combinar modelos en un mismo sistema, para así, aprovechar las ventajas particulares de cada arquitectura. Por ejemplo, las CNN y sus versiones ligeras (LCNN) se emplean con frecuencia como base en proyectos de detección, por su capacidad para identificar patrones en los espectrogramas. En consecuencia, se refuerza la robustez del modelo, dado que se le aplica un pre-filtro con las LCNN que detecta las regiones donde exista una mayor presencia de ondas en el espectrograma, mientras que la CNN se encarga de analizar esos fragmentos seleccionados y realizar una clasificación más precisa.

Asimismo, se han propuesto otros modelos como las arquitecturas Transformer, aplicadas al procesamiento de audio en general. En combinación con las CNN, han demostrado una mejor capacidad de generalización en pruebas cruzadas, aunque sin aplicarse en un entorno real o simulado dentro del contexto bancario [13]. De igual forma, se han explorado las redes de cápsulas (CapsNet) para preservar relaciones espaciales de alto nivel. Cheng et al. [10] incorporó ruido controlado y una función softmax optimizada, logrando una mejora en la extracción de características falsificadas y en el enrutamiento dinámico de los datos de prueba. Este proceso permite determinar de manera más eficiente qué información es relevante para pasar a la siguiente capa. Aunque estos modelos no se han implementado en aplicaciones reales, sus resultados evidenciaron un porcentaje de éxito mayor a otros proyectos que se han establecido previamente dada su capacidad de reacomodar los pesos de las conexiones entre capas. Dentro de los enfoques estudiados se destacan:

- CNN y sus variantes: Destacan por la capacidad para extraer patrones locales en cuanto a la frecuencia y tiempo de un audio, representadas gráficamente en elementos como puede ser el espectrograma [10].
- Transformers: Tienen la capacidad de aprender dependencias de largo alcance en la señal [9].

- Modelos auto supervisados: Se resaltan los sistemas AASIST o AASIST con módulos de squeeze-and-excitation, que combinan pre-entrenamiento masivo con bloques CNN y han dominado recientes competiciones [13].

Igualmente, los estudios analizados emplean diversos corpus públicos, diseñados para detección de *deepfakes* de audio y anti-spoofing. Desafíos como el ASVspoof (2015-2022) aportaron significativamente a las bases de voz sintética encaminadas hacia ataques lógicos, ya sea por conversión de voz o generación de audio a partir de texto [14]. De igual forma, se utilizan conjuntos de datos creados específicamente para la detección de audio falso, como WaveFake, que contiene más de 100.000 clips de audio de siete redes tipo conversoras de voz de última generación [14]. De esta manera, se demuestra que los datos artificiales generados para este tipo de ejercicios existen en cantidad, constituyendo una opción viable para las personas que desarrollan una solución tangible para esta problemática. Sin embargo, de acuerdo con los resultados de la revisión sistemática, estos conjuntos presentan el problema al llegar a aplicarse por la falta de contextualización que presentan en casos específicos como el sector bancario.

En escenarios de fraude específicos, han surgido conjuntos de datos especializados para el entrenamiento de los modelos. Uno de ellos es HAD (Half-truth Audio Detection), que incluye audios parcialmente falsificados donde solo algunas palabras se manipularon para la generación de los datos, y por otro lado, cuentan con una etiqueta de la región editada [15]. Sin embargo, una posible limitación de este tipo de conjuntos de datos es su dependencia del contexto en el que fueron recopilados, dado que los patrones de manipulación o las condiciones acústicas pueden diferir de los entornos reales que enfrentan las entidades financieras. Por lo tanto, la utilidad de los datos puede verse condicionada por el contexto de origen, ya que las técnicas utilizadas en la generación de los conjuntos de datos pueden no coincidir con las que se observan en escenarios financieros reales.

En concordancia, Lu y Ebrahimi [15] postulan que los modelos de detección muestran una disminución significativa en su rendimiento al momento de aplicarse a datos no vistos previamente o con características no esperadas por la red. Dentro de estas particularidades, en el contexto financiero puede encontrarse ruido, acentos no reconocidos, extranjerismos inexistentes en los datos de entreno o variaciones que se pueden dar en una conversación operacional de una llamada bancaria.

Ante estas limitaciones, se identifica la importancia de contar con métricas de evaluación estandarizadas para medir imparcialmente el desempeño de los modelos bajo diferentes condiciones de uso. En este sentido, se propone el EER, que se define como el punto en el que la tasa de falsos positivos es igual a la tasa de falsos negativos en un contexto donde se realice una clasificación [16]. Asimismo, los estudios usan métricas de precisión, y en contextos más específicos, se emplea la métrica t-DCF (función de costo tándem), que cuantifica el impacto del detector de falsificaciones en el rendimiento global del sistema [10]. Estas métricas permiten determinar la utilidad y capacidad de generalización de las redes neuronales en entornos reales y de prueba para generar comparaciones y resaltar su efectividad.

Con base en los resultados, algunos estudios han determinado que las IAs generativas dejan rastro en sus productos creados que se pueden detectar en el dominio de la frecuencia [17]. Ese hallazgo abre la

posibilidad de identificar contenido sintético sin depender del entorno controlado, ampliando el alcance de las estrategias de detección hacia escenarios más reales.

No obstante, la mayoría de los detectores alcanzan resultados favorables únicamente en entornos controlados, disminuyendo en ambientes reales o no preparados para la situación propuesta. En el caso de Pierno et al. [16] se presenta un EER de 0.25% en un dominio conocido que aumenta hasta un 16.4% al ensayar fuera de sus dominios conocidos. Aunque esta cifra pueda parecer reducida, en el ámbito financiero los sistemas bancarios no tienen la posibilidad de tolerar de aceptar márgenes de error.

De manera concordante, estudios recientes confirman esta tendencia, ya que existen algoritmos que presentan una alta precisión en pruebas de laboratorio, pero pierden eficiencia cuando existe una variación en el ruido, el sonido ambiente o el lenguaje utilizado [18]. Esto sugiere que el déficit de desempeño observado proviene, en parte, de la falta de contextualización de los datos de entrenamiento, pues estos no reflejan las condiciones reales de interacción en los entornos financieros.

De acuerdo con lo anterior Ma et al. [19] señala que todo modelo que supere el 99% de acierto en una evaluación vista, tiene alta probabilidad de producir porcentajes de error altos. Esto hace énfasis en la dificultad que existe en trasladar los resultados de un entorno controlado y predecible a aplicarlo en el entorno bancario real, que se considera impredecible al tener condiciones nuevas, como el ruido ambiente, el sonido generado por diversos tipos de micrófonos o la compresión de audio que suelen ser adversas a lo que esta designado en el entrenamiento de la red.

En este sentido, se infiere que, incluso si se lograra mantener la eficacia en situaciones reales, el coste computacional de evaluar conversaciones enteras sería muy alto. Por ello, la segmentación y prelación en el análisis debe concentrarse en momentos críticos de alto riesgo dentro de la conversación, lo que conlleva la necesidad de identificar momentos sin procesar datos innecesariamente.

En aplicaciones de alto riesgo, como las transacciones bancarias, el mínimo error en detección puede resultar en consecuencias graves. Dentro de los escenarios ilustrados en los estudios analizados, se destacan sistemas TTS empleados para la creación de audios falsos de políticos influyentes o de directivos empresariales ordenando transferencias ilegítimas [10], demostrando que los *deepfakes* de voz representan amenazas económicas. En este contexto, enfocar los modelos en momentos críticos incrementa la utilidad práctica de los sistemas de detección y fortalece su aplicación en la ciberseguridad financiera.

El uso de detección priorizada se justifica, además, por la dificultad que tienen las personas para detectar audios falsos en contextos sensibles [20]. Invertir recursos computacionales y diseñar modelos más sofisticados tiene mayor valor en audios de alto riesgo en el sector financiero, dado que los falsos negativos o positivos pueden presentar consecuencias graves [21]. Esta orientación permite optimizar la respuesta de los sistemas de detección hacia escenarios con mayor impacto y se plantea como línea de mejora la implementación de modelos adaptativos capaces de priorizar segmentos críticos de audio según el nivel de riesgo transaccional o el tipo de interacción con el cliente.

Esta priorización influye directamente en la selección del modelo y las métricas de evaluación, dado que favorece a aquellos que tengan una baja tasa de falsos negativos, a pesar de que implique un incremento de falsos positivos. Abbasi et al. [22] realiza un estudio de detección profunda cuyo objetivo se alinea con el del presente artículo: se priorizo no perder ningún deepfake con un 90.5% de recall (baja tasa de falsos negativos). Aplicando este enfoque a la detección de audio en entornos financieros o bancarios, se optaría por arquitecturas que maximicen la detección de suplantaciones a pesar de reducir la precisión del modelo.

De manera complementaria, esta priorización impacta por igual la composición del conjunto de entrenamiento, que permite cubrir audios sensibles de transacciones bancarias, lo que sugiere incluir datos de dominio relevante como grabaciones de llamadas transaccionales, actualizaciones de datos vía llamada o registro de nuevos usuarios. Esto permitiría que el modelo aprenda las características de cada situación, detectando anomalías en contextos específicos. Dentro de la literatura relacionada con el procesamiento de voz, se han desarrollado clasificadores de contenido sensible basados en el texto hablado [23], los cuales podrían integrarse en un sistema híbrido que pueda etiquetar el tipo de contenido y activar los módulos de detección diseñados para cada escenario.

Conformemente, se presenta una propuesta conceptual de detección de suplantación por voz fundamentada en la priorización y el análisis segmentado del audio, planteando una contribución al campo de la ciberseguridad bancaria. Dicha propuesta se sustenta en la revisión comparativa de diversas arquitecturas de aprendizaje profundo y en la evaluación de su eficacia puesta en escenarios de fraude bancario.

Por consiguiente, el aporte central del estudio reside en la transformación de resultados teóricos en una propuesta aplicable, en la cual el modelo conceptual no procese el audio completo, sino que adopte una lógica de análisis adaptativa al nivel de riesgo y al contenido, dando prioridad a la verificación en las partes de mayor sensibilidad. De acuerdo con [24] “Este mecanismo de atención selectiva se perfila como una estrategia innovadora frente a los sistemas tradicionales, al canalizar el procesamiento intensivo solo hacia los fragmentos de audio con mayor probabilidad de suplantación”. De esta manera, la propuesta consolida un enfoque adaptativo que busca mantener un equilibrio entre el rendimiento y la confiabilidad en el entorno financiero.

Esta propuesta ofrece una alternativa posible para entidades financieras interesadas en integrar tecnologías para la detección de *deepfakes* de voz en los canales de autenticación que se aplican en los procesos bancarios. Se propone un modelo conceptual que constituya las bases para una aplicación práctica de la IA, con el fin de que el sistema financiero pueda fortalecer su seguridad frente a las amenazas de suplantación.

## 4. Conclusiones

La revisión sistemática permitió establecer que los modelos de aprendizaje profundo, en especial las CNN, constituyen una base para la detección de voces sintéticas en contextos financieros. Estas arquitecturas sobresalen por su capacidad para identificar patrones en los espectrogramas del audio que se esté analizando, y de esta manera alcanzar niveles bajos de error, únicamente en entornos controlados. No obstante, su rendimiento disminuye cuando se enfrenta a datos reales debido a la variabilidad, el ruido ambiental y las diferencias acústicas, factores que reducen la precisión del modelo.

Asimismo, se añade que los modelos basados en arquitecturas Transformer presentan una mayor adaptabilidad y generalización en los diferentes tipos de deepfakes de audio. Sin embargo, el desempeño de igual manera se ve afectado por el costo computacional que se utiliza y la disminución en la precisión cuando se realizan pruebas en entornos no controlados. Esto evidencia que la eficacia de los modelos se encuentra limitada por la calidad y diversidad de los conjuntos de datos utilizados en su entrenamiento, confirmando que la brecha existente entre los resultados de laboratorio y las condiciones reales del sector financiero donde es posible encontrar ruido y calidad de grabación variable en cada proceso.

En este contexto, el articulo propone un modelo conceptual de detección segmentada y priorizada, donde los recursos computacionales sean utilizados únicamente en momentos críticos que involucren información sensible o procesos de alto riesgo. Este enfoque busca optimizar la eficiencia, reducir el costo de procesamiento y mejorar la capacidad de respuesta en momentos importantes del área financiera, como solicitudes de transferencia o validaciones de identidad.

Los resultados concluyen que, no existe una arquitectura que sea completamente efectiva frente a condiciones reales de fraude bancario. No obstante, un modelo híbrido CNN-Transformer, complementado con estrategias de atención priorizada y un entrenamiento contextualizado con datos financieros reales, podría ser un fortalecimiento para los sistemas de seguridad bancaria. En consecuencia, futuros trabajos podrían centrarse en validar este enfoque mediante pruebas en entornos financieros reales, evaluando su desempeño frente a variaciones de ruido, acentos y calidad de grabación. En este sentido, integrar mecanismos de aprendizaje continuo permitiría que el sistema se adapte de manera progresiva a nuevas formas de deepfake de voz.

## 5. Declaración de intereses concurrentes

(X) Declaro que no tengo intereses significativos en competencia, incluidos intereses financieros o no financieros, profesionales o personales que interfieran con la presentación completa y objetiva del trabajo descrito en este manuscrito.

## 6. Declaración de disponibilidad de datos

Todos los textos y datos referenciados en este artículo están disponibles públicamente mediante sus respectivo DOI, los cuales están indicados en la bibliografía. De esta manera, se garantiza el acceso a las fuentes originales utilizadas en esta investigación.

## 7. Referencias


[1] L. Banh y G. Strobel, "Generative artificial intelligence", *Electron Markets*, vol. 33, núm. 1, p. 63, dic. 2023, doi: 10.1007/s12525-023-00680-1.

[2] J. M. González, *The global risks report 2024: insight report*, 19th ed. Geneva: World Economic Forum, 2024.

[3] D. Delić, "Deepfake technology used in $35 million Hong Kong bank heist", ProPrivacy.com. Consultado: el 22 de octubre de 2025. [En línea]. Disponible en: https://proprivacy.com/privacy-news/deepfake-technology-used-in-hong-kong-bank-heist

[4] S. Lalchand, V. Srinivas, B. Mayor, y J. Henderson, "Generative AI is expected to magnify the risk of deepfakes and other fraud in banking", Deloitte Insights. Consultado: el 22 de octubre de 2025. [En línea]. Disponible en: https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html

[5] A. P. Siddaway, A. M. Wood, y L. V. Hedges, "How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses", *Annual Review of Psychology*, vol. 70, núm. Volume 70, 2019, pp. 747–770, ene. 2019, doi: 10.1146/annurev-psych-010418-102803.

[6] L.-U. ta'Malta, "What is PRISMA and how to use it?", L-Università ta' Malta. Consultado: el 30 de septiembre de 2025. [En línea]. Disponible en: https://www.um.edu.mt/library/help_az/systematicreviews/prisma/

[7] M. J. Page *et al.*, "Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas", *Revista Española de Cardiología (English Edition)*, vol. 74, núm. 9, pp. 790–799, sep. 2021, doi: 10.1016/j.rec.2021.07.010.

[8] S. Moukhliss, "LibGuides: Nursing: PRISMA". Consultado: el 30 de septiembre de 2025. [En línea]. Disponible en: https://libguides.unf.edu/c.php?g=1054238&p=8780672

[9] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, y Y. Zhao, "Audio Deepfake Detection: A Survey", el 29 de agosto de 2023, *arXiv*: arXiv:2308.14970. doi: 10.48550/arXiv.2308.14970.

[10] H. Cheng, T. Wang, X. Chang, L. Nie, Y. Guo, y Q. Li, "Voice-Face Homogeneity Tells Deepfake", *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, nov. 2023, doi: 10.1145/3625231.

[11] K. Jędrasiak, "Audio Stream Analysis for Deep Fake Threat Identification", *Civitas et Lex*, vol. 41, núm. 1, pp. 21–35, 2024.

[12] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, y K. Böttinger, "Does Audio Deepfake Detection Generalize?", el 27 de agosto de 2024, *arXiv*: arXiv:2203.16263. doi: 10.48550/arXiv.2203.16263.

[13] J. Yi *et al.*, "Half-Truth: A Partially Fake Audio Detection Dataset", presentado en Proc. Interspeech 2021, 2021, pp. 1654–1658. doi: 10.21437/Interspeech.2021-930.

[14] K. Samrouth, P. El Housseini, y O. Deforges, "Siamese Network-Based Detection of Deepfake Impersonation Attacks with a Person of Interest Approach.", *ACM Transactions on Multimedia Computing, Communications & Applications*, vol. 21, núm. 3, pp. 1–23, mar. 2025, doi: 10.1145/3708352.

[15] Y. Lu y T. Ebrahimi, "Assessment framework for deepfake detection in real-world situations", *EURASIP Journal on Image and Video Processing*, vol. 2024, núm. 1, p. 6, feb. 2024, doi: 10.1186/s13640-024-00621-8.

[16] A. D. Pierno, L. Guarnera, D. Allegra, y S. Battiato, "End-to-end Audio Deepfake Detection from RAW Waveforms: a RawNet-Based Approach with Cross-Dataset Evaluation", el 29 de abril de 2025, *arXiv*: arXiv:2504.20923. doi: 10.48550/arXiv.2504.20923.

[17] L. Guarnera, O. Giudice, C. Nastasi, y S. Battiato, “Preliminary Forensics Analysis of DeepFake Images”, en *2020 AEIT International Annual Conference (AEIT)*, sep. 2020, pp. 1–6. doi: 10.23919/AEIT50178.2020.9241108.
[18] G. Channing, J. Sock, R. Clark, P. Torr, y C. S. de Witt, “Toward Robust Real-World Audio Deepfake Detection: Closing the Explainability Gap”, el 9 de octubre de 2024, *arXiv*: arXiv:2410.07436. doi: 10.48550/arXiv.2410.07436.
[19] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, y C. Wang, “Continual Learning for Fake Audio Detection”, en *Interspeech 2021*, ago. 2021, pp. 886–890. doi: 10.21437/Interspeech.2021-794.
[20] K. Warren *et al.*, “‘Better Be Computer or I’m Dumb’: A Large-Scale Evaluation of Humans as Audio Deepfake Detectors”, en *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City UT USA: ACM, dic. 2024, pp. 2696–2710. doi: 10.1145/3658644.3670325.
[21] M. G. M. Arcila y F. O. O. Pabon, “Repercusiones éticas sobre el uso indebido del deepfake en el ámbito de las TIC mediante un análisis cualitativo documental”, *Reto*, vol. 9, núm. 1, pp. 36–47, 2021, doi: 10.23850/reto.v9i1.3040.
[22] M. Abbasi, P. Váz, J. Silva, y P. Martins, “Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks”, *Applied Sciences*, vol. 15, núm. 3, p. 1225, ene. 2025, doi: 10.3390/app15031225.
[23] R. Tripathi, B. Dhamodharaswamy, S. Jagannathan, y A. Nandi, “Detecting Sensitive Content in Spoken Language”, en *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, oct. 2019, pp. 374–381. doi: 10.1109/DSAA.2019.00052.
[24] OpenAI, “ChatGPT”, ChatGPT. Consultado: el 30 de septiembre de 2025. [En línea]. Disponible en: https://chatgpt.com/?locale=es-419