FACULTAD DE MEDICINA HUMANA

ESCUELA PROFESIONAL DE MEDICINA HUMANA

TESIS

VALIDACIÓN DE UNA INTELIGENCIA ARTIFICIAL USANDO CHATBOT PARA LA CLASIFICACIÓN BIRADS EN EL TAMIZAJE DE CÁNCER DE MAMA EN LORETO, 2024

PARA OPTAR EL TÍTULO PROFESIONAL DE

MÉDICO CIRUJANO

PRESENTADO POR:

HELMER HUANIO CACHIQUE

ASESOR:

MC. JORGE MIGUEL SIBINA VELA, Dr.

IQUITOS, PERÚ

2025

ACTA DE SUSTENTACIÓN DE TESIS
N°040 / CGT- FMH-UNAP-2025

En la ciudad de Iquitos, distrito de Punchana, departamento de Loreto, a los **21 días del mes de octubre del 2025 a horas 13:00 pm,** se dio inicio a la sustentación pública de la Tesis titulado **"VALIDACIÓN DE UNA INTELIGENCIA ARTIFICIAL USANDO CHATBOT PARA LA CLASIFICACIÓN BIRADS EN EL TAMIZAJE DE CÁNCER DE MAMA EN LORETO, 2024",** aprobada la sustentación con Resolución Decanal **Nº 527-2025-FMH-UNAP** del bachiller **Helmer Huanio Cachique,** para optar el título profesional de Médico Cirujano.

El jurado calificador y dictaminador designado mediante **Resolución Decanal N° 340-2025-FMH-UNAP:**

- MC. Francisco Flores Echevarría, Mgtr. GSS. — Presidente
- MC. Alain Elías Arévalo Mera — Miembro
- MC. Cecilia del Carmen Cueller Aleman — Miembro
- MC. Jorge Miguel Sibina Vela, Dr. — Asesor

Luego de haber escuchado con atención y formulado las preguntas necesarias, las cuales fueron respondidas: SATISFACTORIAMENTE

El jurado después de las deliberaciones correspondientes, llegó a las siguientes conclusiones:

La sustentación pública de la tesis ha sido APROBADO con la calificación de MUY BUENA

Estando el bachiller APTO para obtener título profesional de Médico Cirujano.

Siendo las 2:00 PM se dio por terminado el acto académico.

MC. Francisco Flores Echevarría, Mgtr. GSS.
Presidente

MC. Alain Elías Arévalo Mera
Miembro

MC. Cecilia del Carmen Cueller Aleman
Miembro

MC. Jorge Miguel Sibina Vela, Dr.
Asesor

MIEMBROS DEL JURADO CALIFICADOR Y DICTAMINADOR Y JURADO

MC. Francisco Flores Echevarría, Mgtr. GSS.
Presidente

MC. Alain Elías Arévalo Mera
Miembro

MC. Cecilia del Carmen Cueller Aleman
Miembro

MC. Jorge Miguel Sibina Vela, Dr.
Asesor

RESULTADO DEL INFORME DE SIMILITUD

HELMER HUANIO CACHIQUE

FMH_MH_TESIS_HUANIO CACHIQUE.pdf

20-24 OCT
20-24 OCT
Universidad Nacional De La Amazonia Peruana

Detalles del documento

Identificador de la entrega
trn:oid:::20208:517064238

Fecha de entrega
23 oct 2025, 12:05 p.m. GMT-5

Fecha de descarga
23 oct 2025, 12:13 p.m. GMT-5

Nombre del archivo
FMH_MH_TESIS_HUANIO CACHIQUE.pdf

Tamaño del archivo
879.3 KB

50 páginas

9428 palabras

53.667 caracteres

12% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Coincidencias menores (menos de 10 palabras)

Fuentes principales

8% Fuentes de Internet
2% Publicaciones
9% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

DEDICATORIA

A mis padres, por su amor incondicional, su sacrificio constante y por ser mi mayor inspiración para perseguir mis sueños. Gracias por guiarme y recordarme que la perseverancia es la clave del éxito.
A mis hermanos, por su apoyo y por compartir conmigo cada paso de este camino.

A todos aquellos que, de una u otra forma, encendieron en mí la llama del conocimiento. A quienes me enseñaron a cuestionar, a dudar y a buscar. Este trabajo es la culminación de un largo viaje, pero también el comienzo de muchos más.

AGRADECIMIENTO

La culminación de este trabajo de investigación ha sido posible gracias al apoyo y la colaboración de numerosas personas e instituciones a lo largo de este proceso. Expreso mi sincero agradecimiento a todos aquellos que, de una u otra forma, contribuyeron a su realización.

- **A mi universidad:** Por haberme brindado la oportunidad de crecer académica y profesionalmente, y por ofrecer los recursos necesarios para llevar a cabo esta investigación.
- **A mis profesores:** Por su orientación, su conocimiento y su invaluable apoyo durante mi formación. Su dedicación fue fundamental para mi desarrollo.
- **A mi asesor de tesis:** Por su guía, sus sabias sugerencias y su paciencia. Su experiencia y acompañamiento han sido clave para la estructuración y conclusión de este proyecto.
- **A mis compañeros de estudios:** Por el intercambio de ideas, la solidaridad y el apoyo mutuo durante esta etapa académica.
- **A mis familiares y amigos:** Por el apoyo constante, la comprensión y el aliento brindados a lo largo de este camino.

Concluyo con una sincera gratitud a todos los que formaron parte de esta etapa. Sus aportes han sido esenciales para la culminación de este trabajo.

ÍNDICE

ÍNDICE DE TABLAS

RESUMEN

Introducción: El cáncer de mama constituye una de las principales causas de mortalidad femenina a nivel mundial, con más de 2,3 millones de casos anuales según GLOBOCAN 2022. Las diferencias en la supervivencia entre países desarrollados y en vías de desarrollo reflejan desigualdades en el acceso a diagnóstico precoz y tratamiento oportuno. En la región de Loreto, Perú, la escasez de especialistas y las limitaciones en la interpretación de mamografías representan un desafío para el tamizaje efectivo. El sistema BI-RADS, clave en la clasificación mamográfica, depende de la experiencia del radiólogo, lo que introduce variabilidad diagnóstica. La inteligencia artificial (IA) surge como una alternativa prometedora para mejorar la precisión diagnóstica y reducir brechas de acceso. **Objetivo:** Validar la eficacia de una inteligencia artificial usando un chatbot (IA) para la clasificación BIRADS en el tamizaje de cáncer de mama en el Hospital Regional de Loreto, 2024. **Método:** Se diseñó un estudio de tipo Observacional con un diseño descriptivo, transversal y correlacional. **Población:** La muestra estuvo conformada por 854 imágenes seleccionadas de manera probabilística y clasificables según el sistema BI-RADS. **Resultados:** El modelo GMIC (GREYC's Magic for Image Computing) presentó una exactitud de 46,49%, con sensibilidad de 57,89% y especificidad de 46,23%. El valor predictivo positivo fue bajo (2,39%), con tasas elevadas de error: 53,77% de falsos positivos y 42,11% de falsos negativos. El coeficiente Kappa de Cohen fue 0,003, evidenciando concordancia prácticamente nula respecto al radiólogo, y la prueba de McNemar mostró diferencias estadísticamente significativas en las discordancias. **Conclusiones:** La validación del chatbot de inteligencia artificial para la clasificación BI-RADS permitió evidenciar limitaciones en su desempeño frente al criterio del radiólogo. No obstante, los hallazgos constituyen un aporte local al conocimiento sobre la aplicabilidad de estas herramientas en contextos con recursos limitados, ofreciendo una base para futuros ajustes metodológicos y validaciones en la práctica clínica regional.

Palabras clave: Cáncer de Mama, Inteligencia Artificial, Sensibilidad, Especificidad.

ABSTRACT

Introduction: Breast cancer is one of the leading causes of female mortality worldwide, with more than 2.3 million annual cases according to GLOBOCAN 2022. Differences in survival between developed and developing countries reflect inequalities in access to early diagnosis and timely treatment. In the Loreto region of Peru, the shortage of specialists and the limitations in mammography interpretation represent a challenge for effective screening. The BI-RADS system, essential for mammographic classification, largely depends on the radiologist's expertise, introducing diagnostic variability. Artificial intelligence (AI) has emerged as a promising alternative to improve diagnostic accuracy and reduce access gaps. **Objective:** To validate the effectiveness of an artificial intelligence chatbot for BI-RADS classification in breast cancer screening at the Regional Hospital of Loreto, 2024. **Method:** An observational study with a descriptive, cross-sectional, and correlational design was conducted. **Population:** The sample consisted of 536 probabilistically selected images that were classifiable according to the BI-RADS system. **Results:** The GMIC model showed an accuracy of 46.49%, with sensitivity of 57.89% and specificity of 46.23%. The positive predictive value was low (2.39%), with high error rates: 53.77% false positives and 42.11% false negatives. Cohen's Kappa coefficient was 0.003, indicating virtually no agreement with the radiologist's evaluation, and the McNemar test revealed statistically significant differences in the observed discordances. **Conclusions:** The validation of the artificial intelligence chatbot for BI-RADS classification revealed limitations in its performance compared with the radiologist's assessment. Nevertheless, the findings provide local evidence on the applicability of these tools in resource-limited settings, offering a foundation for future methodological refinements and clinical validations in the regional context.

Keywords: Breast Cancer; Artificial Intelligence; Sensitivity; Specificity.

INTRODUCCIÓN

1.1 Descripción de la situación problemática:

El cáncer de mama es una de las principales causas de mortalidad en mujeres a nivel mundial, afectando de manera significativa la calidad de vida de las pacientes. De acuerdo con los datos de GLOBOCAN 2022, se diagnosticaron alrededor de 2,3 millones de casos nuevos de cáncer de mama en el mundo, lo que representa aproximadamente 23,8% de todos los cánceres en mujeres, y se registraron cerca de 666 mil muertes por esta enfermedad. La detección temprana continúa siendo clave para mejorar las tasas de supervivencia. En países con mayores recursos, la supervivencia a cinco años supera el 90%, mientras que en naciones de ingresos bajos y medios esta cifra puede descender hasta 40-66%, debido en gran medida a las limitaciones en la detección precoz y a la falta de acceso oportuno al tratamiento (1).

En Perú, y particularmente en la región de Loreto, el acceso a especialistas como radiólogos y oncólogos es limitado, lo que genera retrasos en el diagnóstico y tratamiento del cáncer de mama. Un estudio realizado en un hospital peruano en 2019 identificó brechas significativas en el acceso al tratamiento oncológico, evidenciando la necesidad de mejorar la disponibilidad de especialistas en regiones como Loreto (2).

Esta situación es crítica, ya que el sistema BI-RADS (Breast Imaging Reporting and Data System), diseñado para estandarizar la clasificación de mamografías, depende en gran medida de la experiencia de los radiólogos. Un estudio en Colombia encontró deficiencias en la calidad de las imágenes y en su interpretación, lo que puede introducir variabilidad y posibles errores en la

clasificación (3).

La inteligencia artificial (IA) ha demostrado un gran potencial para mejorar el diagnóstico mamográfico. Un estudio evaluó el impacto de la IA en el cribado del cáncer de mama utilizando mamografía digital, encontrando que los radiólogos pudieron asignar con mayor precisión las categorías BI-RADS sin aumentar el tiempo de interpretación, concluyendo que el uso de IA facilita una mejor clasificación de las mamografías y optimiza la detección de cáncer de mama de manera eficiente y sin afectar el tiempo de lectura (4).

En China, un sistema de clasificación BI-RADS mediante inteligencia artificial logró una precisión del 84.5%, demostrando su utilidad para asistir a los radiólogos en decisiones clínicas clave (5). De manera similar, estudios han evaluado el uso de IA en la detección del cáncer de mama mediante mamografía, destacando su potencial para mejorar la precisión diagnóstica y la concordancia con radiólogos expertos (6). Además, un estudio en España utilizó un sistema basado en redes neuronales para detectar y clasificar lesiones mamarias, obteniendo una precisión del 96.5% en la detección y logrando concordancia diagnóstica con radiólogos expertos (7).

En Perú, un modelo desarrollado en el Hospital Almanzor Aguinaga Asenjo alcanzó una sensibilidad del 91.4% y una especificidad del 85.7%, demostrando que las herramientas computacionales pueden optimizar la detección temprana del cáncer de mama en contextos locales (8). De manera similar, un sistema inteligente basado en redes neuronales convolucionales implementado en Chiclayo clasificó mamografías con una precisión general del 88.5%, resaltando la viabilidad de esta tecnología para el diagnóstico temprano (9).

En Hawái, EE. UU., se desarrolló un modelo de inteligencia artificial explicable para la detección de lesiones en ecografías mamarias, utilizando una capa de embudo conceptual basada en BI-RADS, lo que permite a los radiólogos revisar y modificar las predicciones del sistema, mejorando su precisión del 0.876 al 0.885 en el AUC (10) En Colombia, la Universidad ECCI validó un software para detectar regiones anormales en mamografías, basado en un detector de objetos en cascada, obteniendo una precisión del 73% y destacando la densidad mamaria como un factor clave en las imágenes procesadas (11). Finalmente, en Perú, la Universidad Peruana de Ciencias Aplicadas diseñó modelos de clasificación de densidad mamaria con redes neuronales convolucionales, concluyendo que el uso de múltiples validaciones cruzadas puede alcanzar hasta un 99% de precisión en la clasificación, lo que refuerza el potencial de la IA en el diagnóstico mamográfico (12).

A pesar de estos avances, la implementación de una IA enfrenta desafíos en entornos con recursos limitados como Loreto, incluyendo la calidad variable de las mamografías y la necesidad de validaciones locales para garantizar la eficacia de los modelos. Este estudio propone validar el uso de una inteligencia artificial para la clasificación BI-RADS en mamografías de pacientes atendidas en el Hospital Regional de Loreto, comparando su desempeño con el de médicos radiólogos, con el objetivo de optimizar el tamizaje y reducir la inequidad en el acceso a servicios especializados.

1.2 Formulación del problema

¿Cuál es la eficacia de una Inteligencia Artificial usando un chatbot para la clasificación BIRADS como herramienta para el tamizaje de Cáncer de Mama, en Loreto 2024?

1.3 Objetivos:

1.3.1 General

Validar la eficacia de una inteligencia artificial usando un chatbot para la clasificación BIRADS en el tamizaje de cáncer de mama en el Hospital Regional de Loreto, 2024.

1.3.2 Específicos

- Comparar la clasificación BIRADS realizada por el chatbot con la de los médicos radiólogos en el tamizaje de cáncer de mama.
- Evaluar la sensibilidad, especificidad y el grado de concordancia diagnóstica del chatbot frente al criterio del radiólogo.
- Determinar la tasa de falsos positivos y falsos negativos generados por el chatbot en la clasificación BIRADS.

1.4 Justificación:

1.4.1 Importancia:

Este proyecto aborda una necesidad crítica en el sistema de salud de Loreto, una región donde la limitación de recursos humanos especializados afecta la capacidad de diagnóstico temprano del cáncer de mama. La inteligencia artificial se ha convertido en una herramienta clave para la automatización y estandarización del análisis de imágenes médicas, ofreciendo soluciones prometedoras que permiten reducir la carga de trabajo de los especialistas y mejorar la precisión en el diagnóstico. (13)

El uso del sistema BIRADS es fundamental para la clasificación de mamografías y la identificación de lesiones mamarias. Sin embargo, la interpretación de imágenes está sujeta a variaciones entre los médicos, lo que puede llevar a diagnósticos inconsistentes. Estudios

recientes han demostrado que la IA (Inteligencia Artificial) puede igualar e incluso superar la precisión de los radiólogos en la clasificación de imágenes, lo que la convierte en una herramienta valiosa para contextos donde la disponibilidad de personal especializado es limitada (14).

Implementar esta Inteligencia artificial validada podría mejorar el acceso a diagnósticos precisos en áreas con menos recursos, como Loreto, donde la infraestructura hospitalaria no siempre permite contar con médicos radiólogos. Este enfoque no solo mejoraría la calidad del diagnóstico, sino que también ayudaría a reducir la mortalidad por cáncer de mama al facilitar el acceso a diagnósticos tempranos y oportunos.

La implementación de inteligencia artificial en la clasificación BIRADS no solo representa un avance tecnológico en la atención médica, sino que también conlleva beneficios económicos y sociales significativos, especialmente en regiones con recursos limitados como Loreto. Desde el punto de vista económico, esta tecnología puede reducir de manera considerable los costos de diagnóstico al disminuir la necesidad de derivaciones a centros especializados. En una región donde los pacientes deben trasladarse a grandes ciudades para obtener un diagnóstico preciso, esta reducción de costos no solo implica un ahorro directo en traslados, sino también una optimización de los recursos sanitarios disponibles. Además, al facilitar un diagnóstico temprano, la IA puede evitar tratamientos más costosos asociados con el cáncer detectado en etapas avanzadas, lo que no

solo beneficia a los pacientes, sino que también reduce la carga económica sobre el sistema de salud. La estandarización en la interpretación de imágenes ayuda, además, a minimizar tratamientos innecesarios, asegurando un uso más eficiente de los recursos médicos (15).

Desde una perspectiva tecnológica, la inteligencia artificial aplicada al diagnóstico mamográfico representa un avance clave en la medicina personalizada y predictiva. Su capacidad para analizar imágenes con mayor rapidez y precisión que el ojo humano no solo optimiza el proceso diagnóstico, sino que también disminuye el margen de error asociado a la variabilidad en la experiencia de los médicos. Este aspecto es crucial en zonas con escasez de especialistas, donde la IA permite llevar diagnósticos de alta precisión a comunidades rurales, eliminando la dependencia de profesionales presenciales y garantizando un acceso más equitativo a servicios de salud de calidad (16).

A nivel social, la implementación de esta tecnología impactaría directamente en la mejora de la salud pública, beneficiando particularmente a las mujeres de Loreto, quienes actualmente enfrentan barreras significativas para acceder a diagnósticos especializados. Un acceso más oportuno y preciso a la detección temprana del cáncer de mama podría mejorar los resultados del tratamiento y reducir la mortalidad en la región. Además, la incorporación de la IA en centros de primer nivel de atención ayudaría a disminuir las inequidades en salud entre zonas rurales y urbanas,

promoviendo un sistema más inclusivo y eficiente. Así, la combinación de beneficios económicos, tecnológicos y sociales posiciona a la inteligencia artificial como una herramienta clave para transformar la atención médica y reducir las desigualdades en el acceso a diagnósticos de calidad.

1.4.2 Viabilidad:

La viabilidad del proyecto está respaldada por la colaboración del Hospital Regional de Loreto y la disponibilidad de imágenes mamográficas para realizar la validación. El chatbot (Inteligencia Artificial) cuenta con un algoritmo previamente desarrollado y ha sido entrenado en diversos contextos, lo que facilita su adaptación al entorno local sin grandes obstáculos. Solo requiere ajustarse al proyecto con los datos proporcionados por el Hospital Regional de Loreto. Además, su tecnología es escalable, permitiendo su implementación en distintos niveles de atención.

Recursos disponibles: La infraestructura del hospital, junto con la participación de médicos radiólogos locales, garantiza que el proceso de validación se realice de manera efectiva. Se cuenta con el apoyo de los especialistas para comparar los resultados de la IA con los diagnósticos humanos, lo que asegura la calidad del estudio.

Escalabilidad: Una vez validada, la IA podrá implementarse en otros centros de salud de la región, facilitando el acceso a diagnósticos rápidos y precisos en áreas de difícil acceso.

CAPITULO I: MARCO TEÓRICO

1.1 Antecedentes:

INTERNACIONALES

En 2023, se realizó un estudio experimental en Bogotá, Colombia, para validar un software basado en inteligencia artificial diseñado para detectar regiones anormales en mamografías mediante un detector de objetos en cascada. El software fue entrenado utilizando imágenes de repositorios internacionales y luego probado con imágenes locales pre-procesadas para asegurar el anonimato. Los resultados, evaluados mediante una matriz de confusión, mostraron una precisión del 73% en ciertas proyecciones, aunque con una sensibilidad baja (1%) y especificidad variable según las vistas mamográficas analizadas. Este estudio resalta el potencial del software como herramienta de apoyo para radiólogos, aunque se requieren mejoras en su sensibilidad y especificidad para aplicaciones clínicas más robustas (11).

En 2023, se realizó un estudio en el Instituto Politécnico Nacional de México para evaluar un modelo basado en redes neuronales convolucionales ensambladas (CNN) aplicado a la clasificación BI-RADS de tumores de mama en ultrasonido, abordando categorías 2 a 5. Se implementaron estrategias de descomposición binaria que permitieron una clasificación más precisa al convertir problemas multiclase en binarios. El modelo denominado D5_v1, identificado como el mejor, alcanzó un coeficiente de correlación de Matthews de 0.77, superando en especificidad (0.92) al promedio de radiólogos (0.79) al reducir la necesidad de biopsias innecesarias en lesiones probablemente benignas. Sin embargo, la sensibilidad (0.87) fue inferior a la de los especialistas, evidenciando áreas de mejora en la detección de malignidad. El estudio concluye que las estrategias de ensamble y descomposición binaria representan un avance significativo en la

precisión y objetividad de la clasificación BI-RADS mediante inteligencia artificial. (17).

En 2023, se realizó un estudio en Costa Rica que evaluó el impacto de la inteligencia artificial (IA) en la detección temprana del cáncer de mama, utilizando algoritmos de aprendizaje profundo aplicados a mamografías. Este trabajo destacó la capacidad de la IA para identificar patrones patológicos hasta cinco años antes de su manifestación clínica. Los resultados mostraron que la IA alcanzó una precisión diagnóstica del 96%, reduciendo falsos positivos en un 30% en comparación con los métodos tradicionales, y detectando un 20% más de tumores en tamizajes asistidos en comparación con la revisión de dos radiólogos. Además, se observó una disminución del 50% en la carga laboral de los especialistas, optimizando significativamente los flujos de trabajo. El estudio concluyó que la implementación de IA en la práctica radiológica mejora la precisión diagnóstica, aumenta las tasas de detección temprana y tiene un impacto transformador en la salud pública, especialmente en regiones con recursos limitados (18).

En 2022, se realizó un estudio experimental en la Institución Universitaria Pascual Bravo, Medellín, Colombia, donde se desarrolló un algoritmo basado en inteligencia artificial para clasificar lesiones benignas y malignas en imágenes de resonancia magnética de mama, utilizando datos de 1,206 imágenes etiquetadas según el sistema BIRADS por radiólogos expertos. El estudio comparó diferentes modelos, como máquinas de soporte vectorial (SVM), redes neuronales y aprendizaje profundo, mostrando que el modelo SVM alcanzó una precisión del 66% en la clasificación de lesiones malignas. Los resultados sugieren que los modelos de IA pueden ser herramientas útiles para apoyar el diagnóstico del

cáncer de mama, aunque se requiere optimización de los parámetros para mejorar su desempeño y aplicabilidad clínica (19).

En 2022, se realizó un trabajo de investigación en la Universidad Nacional de Educación a Distancia (UNED), España, para desarrollar un sistema basado en inteligencia artificial capaz de detectar, describir y clasificar tumores mamarios en ecografías utilizando el sistema BI-RADS. El modelo combinó el algoritmo YOLO (You Only Look Once) para detección de tumores y un sistema encoder-decoder para describir en lenguaje natural las características BI-RADS y estimar la malignidad del tumor. Los resultados mostraron una precisión de detección del 96.5%, una exhaustividad del 95%, y un área bajo la curva de precisión y exhaustividad de 0.97. En la etapa de descripción, se lograron niveles de concordancia con radiólogos expertos equivalentes a la intercorrelación e intracorrelación observada entre profesionales, demostrando la capacidad del modelo para acercarse al razonamiento humano en la clasificación de lesiones mamarias. Este estudio destacó la utilidad del sistema como herramienta de apoyo en la evaluación de tumores en tiempo real (20).

En 2022, se desarrolló un estudio en la Universidad de Málaga para implementar y comparar algoritmos de inteligencia artificial, incluyendo redes neuronales, K-Nearest Neighbors (KNN) y razonamiento basado en casos (CBR), con el fin de apoyar el diagnóstico médico en pacientes con cáncer de mama. El estudio utilizó datos de mamografías, citologías y exámenes histológicos para evaluar la precisión y el área bajo la curva (AUC) de cada modelo. Los resultados mostraron que el modelo de redes neuronales alcanzó una AUC de hasta 0.97 en mamografías, mientras que el CBR ofreció ventajas adicionales al proporcionar representaciones visuales que facilitaron la interpretación por parte

de médicos. Este trabajo concluyó que el modelo CBR es una opción válida y de bajo costo computacional para el apoyo clínico, con áreas de mejora relacionadas con la transformación de datos y optimización de algoritmos (7).

En 2021, se realizó un estudio en el Centro de Investigación y de Estudios Avanzados (CINVESTAV) de Tamaulipas, México, para desarrollar sistemas de clasificación histopatológica explicables basados en el léxico BI-RADS para mamografías. Este trabajo utilizó 1,897 mamografías de la base de datos pública DDSM, evaluando dos sistemas CAD: uno basado en tendencia patológica (CAD_TP) y otro en descripción de atributos (CAD_DA). Los resultados mostraron que el sistema CAD_TP alcanzó una precisión del 90% en la clasificación general, mientras que el CAD_DA obtuvo un 93%. Para la clasificación de forma, los sistemas lograron precisiones del 90% y 82%, respectivamente; para el margen, 90% y 86%; y para la densidad, 82% y 70%. En comparación con sistemas CAD convencionales, los nuevos enfoques demostraron ser competitivos, ofreciendo mayor explicabilidad al correlacionar criterios internos con el estándar BI-RADS. Este trabajo resalta la importancia de herramientas explicables para mejorar la confianza en el diagnóstico asistido por computadora (16).

NACIONALES

En 2023, se desarrolló un estudio en la Universidad Peruana de Ciencias Aplicadas (UPC) para evaluar modelos de clasificación de densidad mamaria mediante redes neuronales convolucionales. Se recopilaron y analizaron 20 artículos seleccionados de bases de datos como ScienceDirect, Scopus e IEEE, utilizando criterios de inclusión específicos como el uso de técnicas de inteligencia artificial aplicadas a la clasificación mamaria en mamografías. Los

resultados destacaron que la precisión de los modelos evaluados, como redes convolucionales profundas y aprendizaje de transferencia, alcanzó hasta un 99% mediante validaciones cruzadas. Además, se identificó que la implementación de estrategias avanzadas de segmentación y normalización de imágenes mejora significativamente la precisión y sensibilidad de los modelos. Por lo tanto, se concluye que las redes neuronales convolucionales representan una herramienta eficaz para la clasificación de densidad mamaria, con alto potencial de aplicación clínica en el diagnóstico temprano de cáncer de mama (12).

En 2021, se realizó un estudio en el Hospital Las Mercedes de Chiclayo, Perú, para desarrollar un sistema inteligente basado en redes neuronales convolucionales (CNN) que apoye el análisis mamográfico en la detección de tumores de mama en mujeres de 40 a 60 años. El sistema utilizó bases de datos públicas como MINI-MIAS y DDSM para el entrenamiento y validación del modelo. El modelo alcanzó una precisión del 85.7% y una exactitud del 88.5% al clasificar mamografías como sanas o con tumores, con una sensibilidad del 71.4% y una especificidad del 88.5%. Además, se integró el modelo en un sistema web local, optimizando los tiempos de diagnóstico y mejorando la capacidad de atención del hospital. El estudio concluyó que el sistema es una herramienta fiable y eficaz para apoyar el diagnóstico temprano del cáncer de mama, destacando la necesidad de bases de datos más amplias para mejorar la precisión (9).

En 2020, se realizó un estudio en el Hospital Almanzor Aguinaga Asenjo, EsSalud, en Chiclayo, Perú, que desarrolló un sistema de detección automática de micro-calcificaciones en mamografías digitales utilizando técnicas de tratamiento de imágenes e inteligencia artificial. El modelo implementó filtros

espaciales, segmentación mediante la transformada Top Hat, y redes neuronales para clasificar micro-calcificaciones como benignas o sospechosas de malignidad. Los resultados mostraron una sensibilidad del 91.4% y una especificidad del 85.7%, con una precisión general del 88%. El sistema logró identificar micro-calcificaciones con diámetros tan pequeños como 0.4 mm, mejorando significativamente la capacidad de detección temprana del cáncer de mama. Este estudio concluyó que el uso de herramientas computacionales basadas en inteligencia artificial puede optimizar la precisión diagnóstica y reducir la carga de trabajo de los especialistas (8).

1.2 Bases Teóricas

Anatomía

La mama es un órgano presente tanto en hombres como en mujeres, ubicado sobre el músculo pectoral mayor. Se extiende verticalmente desde el borde inferior de la segunda costilla hasta la sexta, y transversalmente desde el borde externo del esternón hasta la línea axilar anterior. Por encima de esta estructura se encuentra el complejo areola-pezón (CAP). La composición de la mama incluye tejido glandular, tejido conjuntivo fibroso y tejido adiposo, cuya proporción varía según la edad, estado hormonal y peso de la persona. La mama está constituida por entre 10 y 15 lóbulos, rodeados por crestas fibroglandulares (crestas de Duret) que se insertan en el ligamento de Cooper, una estructura subcutánea. Tradicionalmente, la mama se divide en cuatro cuadrantes: superior externo, superior interno, inferior, y la región del CAP (22).

Densidad mamaria

La densidad mamaria refleja la composición del tejido mamario, el cual está formado por tejido glandular, epitelial conjuntivo fibroso y tejido adiposo. Se evalúa mediante la mamografía y es importante reconocerla debido a su relación con el riesgo de desarrollar cáncer de mama. Diversos estudios han demostrado que cuanto mayor es la proporción de densidad mamaria, mayor es el riesgo de cáncer de seno. Además, una alta densidad mamaria puede dificultar el diagnóstico temprano, ya que el tejido denso puede enmascarar u ocultar tumores (23).

El sistema BI-RADS (Breast Imaging Data Reporting System) es una herramienta estandarizada, cualitativa y cuantitativa, utilizada para determinar la proporción de tejido fibroglandular frente al tejido adiposo. Este sistema, desarrollado y actualizado por el Colegio Americano de Radiología (ACR), facilita la comunicación entre los especialistas (23).

La clasificación cualitativa BI-RADS establece cuatro categorías:

- A: Los senos son mayormente grasos, con un 25% de tejido mamario.
- B: Hay áreas dispersas de tejido fibroglandular que constituyen entre el 25% y el 50% del tejido mamario.
- C: Los senos son heterogéneamente densos, representando entre el 51% y el 71% del tejido mamario.
- D: Los senos son extremadamente densos, con un ≥ 75% de tejido mamario.

Las categorías C y D se consideran factores de riesgo para el cáncer de mama y dificultan la detección de anomalías debido a la disminución de la sensibilidad mamográfica.

CLASIFICACIÓN BI-RADS NÚMERICA	
BI-RADS 0	No concluyente.
BI-RADS 1	Mama normal.
BI-RADS 2	Hallazgos benignos (probabilidad de cáncer similar a la población general)
BI-RADS 3	Hallazgos probablemente benignos (<2% de riesgo de malignidad
BI-RADS 4	Probablemente maligna (valor predictivo positivo para cáncer entre 29-34%-70%).
BI-RADS 5	Altamente sugestivo de malignidad (valor predictivo positivo para cáncer >70%)
BI-RADS 6	Malignidad confirmada histológicamente, pero antes de iniciar manejo definitivo.

Cáncer de mama

El cáncer de mama es una enfermedad oncológica de origen clonal, caracterizada por una alteración en las células de la glándula mamaria debido a su rápida proliferación. Esto se debe a cambios en los mecanismos de división y muerte celular, lo que conduce al crecimiento de tumores malignos o masas anormales. Estas células cancerosas tienen el potencial de diseminarse a otras partes del cuerpo, proceso que se conoce como metástasis. El desarrollo del cáncer de mama está influenciado por diversos factores, entre ellos los genéticos, hormonales y relacionados con el estilo de vida (24).

Estatificación del cáncer de mama

La estatificación es el proceso mediante el cual se evalúa el grado de avance del cáncer en el cuerpo, lo que permite al médico identificar en qué etapa se encuentra la enfermedad y definir las opciones de tratamiento más adecuadas (Estatificación del cáncer, s.f.). Para el cáncer de mama, la estatificación se basa en la clasificación desarrollada por el American Joint Committee on Cancer (AJCC) y la International Union for Cancer Control (UICC). Esta clasificación se realiza según el sistema **TNM**, que se centra en criterios anatómicos (24). Cada letra del sistema tiene un significado específico:

- **T**: Tamaño del tumor.
- **N**: Afectación de los ganglios linfáticos.
- **M**: Presencia de metástasis a distancia.

Exámenes diagnósticos para el cáncer de mama

Cuando en el examen físico se detecta una masa o protuberancia, que en la mayoría de los casos es indolora y puede estar acompañada de signos clínicos como eritema, equimosis, piel de naranja, retracción del pezón o secreción, se indican varios estudios diagnósticos. Estos incluyen mamografías, exámenes paraclínicos y biopsias, entre otros, para confirmar el diagnóstico (25).

Imágenes diagnósticas del cáncer de mama

Se utilizan tres pruebas principales de imagen para identificar posibles anomalías en el tejido mamario:

- Mamografía: Esta técnica emplea dosis bajas de rayos X para generar imágenes del seno desde dos ángulos: medio lateral oblicuo (MLO) y

cráneo-caudal (CC). Al combinar estas vistas, se obtiene una representación tridimensional que facilita la detección de patologías. Esto es útil porque una característica sospechosa en una proyección puede no ser visible en la otra (Hasan et al., 2021).

- Ecografía: El ecógrafo genera imágenes al emplear ondas de ultrasonido, las cuales se basan en las diferencias de densidad de los tejidos (Villavicencio-Romero et al., 2019).
- Resonancia Magnética: Utiliza campos magnéticos para crear imágenes diagnósticas. Esta técnica permite obtener una serie de imágenes detalladas que muestran con precisión los órganos y estructuras óseas (25).

1.3 Definición de términos básicos

- BIRADS (Breast Imaging Reporting and Data System): Sistema de clasificación utilizado por radiólogos para estandarizar la interpretación de mamografías y evaluar el riesgo de malignidad de las lesiones mamarias. BIRADS clasifica las imágenes en siete categorías que van desde 0 (incompleto) hasta 6 (cáncer confirmado por biopsia) (26).
- Sensibilidad: Capacidad de la inteligencia artificial para identificar correctamente los casos positivos de cáncer de mama (lesiones clasificadas como BIRADS 4 o 5). Se calcula como la proporción de verdaderos positivos entre todos los casos que realmente son positivos (verdaderos positivos + falsos negativos) (27).
- Especificidad: Capacidad de la IA para identificar correctamente los casos negativos (lesiones benignas o BIRADS 1-2). Se calcula como la proporción de verdaderos negativos entre todos los casos que realmente

son negativos (verdaderos negativos + falsos positivos) (27).

- Precisión diagnóstica: Capacidad global de la IA para clasificar correctamente las mamografías en sus correspondientes categorías BIRADS, evaluada mediante la sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo (VPN) (27).
- Inteligencia artificial (IA): Tecnología que utiliza algoritmos de aprendizaje automático (machine learning) para analizar grandes volúmenes de datos y realizar tareas que normalmente requieren inteligencia humana. En este contexto, la IA es utilizada para interpretar imágenes de mamografías y clasificar las lesiones mamarias en el sistema (28).
- Falsos positivos: Casos en los que la IA clasifica una mamografía como sospechosa de malignidad (BIRADS 4 o 5), pero el diagnóstico final por biopsia resulta ser benigno. Un alto número de falsos positivos puede llevar a procedimientos innecesarios (28).
- Falsos negativos: Casos en los que la IA clasifica una mamografía como negativa o benigna (BIRADS 1 o 2), pero el diagnóstico final resulta ser maligno. Un alto número de falsos negativos puede retrasar el tratamiento necesario (28).
- Concordancia: Grado de acuerdo entre las clasificaciones realizadas por la IA y las realizadas por los radiólogos. Se mide a través del coeficiente kappa de Cohen, que evalúa el nivel de coincidencia más allá del azar (29).
- Validación externa: Proceso mediante el cual se evalúa la efectividad de una tecnología, como la IA, en un entorno clínico diferente al que fue entrenada originalmente, asegurando que su desempeño sea robusto y generalizable (30) .

- VP (Verdaderos positivos): Casos enfermos correctamente identificados.
- VN (Verdaderos negativos): Casos sanos correctamente identificados.
- FN (Falsos negativos): Enfermos no detectados.
- FP (Falsos positivos): Sanos erróneamente clasificados como enfermos.

CAPITULO II: HIPÓTESIS Y VARIABLES

2.1 Formulación de la hipótesis

- H1: El chatbot basado en inteligencia artificial es eficaz y preciso en la clasificación BI-RADS para el tamizaje de cáncer de mama en pacientes atendidas en el Hospital Regional de Loreto, durante el año 2024.
- H0: El chatbot basado en inteligencia artificial no es eficaz ni preciso en la clasificación BI-RADS para el tamizaje de cáncer de mama en pacientes atendidas en el Hospital Regional de Loreto, durante el año 2024.

2.2 Variables y definiciones operacionales:

Variable dependiente:

- Precisión del diagnóstico del chatbot

Variable independiente:

- Validación de una inteligencia artificial (sensibilidad y especificidad del Chatbot).
- Tasa de falsos positivos/negativos en el tamizaje de cáncer de mama en Loreto.
- Capacidad para la clasificación por categorías BIRADS

Variable	Definición	Tipo por su naturaleza	Indicador	Escala de medición	Categorías	Valores de las categorías	Medio de verificación
Precisión del diagnóstico	Grado en que el diagnóstico por IA coincide con el diagnóstico por métodos tradicionales	Cuantitativa	Porcentaje de coincidencia	Racional	Alta, Media, Baja	>90%, 70-90%, <70%	Resultados de pruebas
Validación de una IA en la sensibilidad del Chatbot	Capacidad del sistema para identificar correctamente a los pacientes que sí padecen una enfermedad, evitando falsos negativos (casos enfermos no detectados).	Cuantitativa	Proporción de verdaderos positivos entre el total de casos positivos reales [VP / (VP + FN)]	Razón	Alta, Media, Baja	>90%, 70-90%, <70%	Resultados de pruebas
Validación de una IA en la especificidad del Chatbot	Capacidad del sistema para reconocer con precisión a los individuos que no tienen la enfermedad, evitando falsos	Cuantitativa	Proporción de verdaderos negativos entre el total de casos	Razón	Alta, Media, Baja	>90%, 70-90%, <70%	Resultados de pruebas

	positivos (casos sanos diagnosticados como enfermos).		negativos reales [VN / (VN + FP)]				
Tasa de falsos positivos/negativos en el tamizaje de cáncer de mama en Loreto.	Proporción de diagnósticos incorrectos que son positivos o negativos	Cuantitativa	Porcentaje de falsos positivos/negativos	Racional	Alta, Media, Baja	>10%, 5-10%, <5%	Resultados de pruebas
Interacción Usuario-Chatbot	Evaluación de la experiencia del usuario al interactuar con el sistema, enfocándose en la facilidad de uso, la satisfacción general y la claridad en la comprensión de las respuestas o diagnósticos proporcionados	Cualitativa	Nivel de satisfacción medido mediante una escala Likert aplicada en encuestas post-interacción.	Ordinal	Alta, Moderado, Baja	Alta, Moderado, Baja	Encuesta

	por el Chatbot.						
Capacidad para la clasificación por categorías BIRADS	Habilidad del chatbot para identificar las diferentes categorías BIIRADS	Cuantitativa	Porcentaje de clasificaciones correctas por categoría	Intervalo	Alta, Media, Baja	>90%, 70-90%, <70%	Resultados de pruebas

CAPITULO III: METODOLOGÍA

3.1 Diseño metodológico

Tipo de Investigación: Observacional.

Diseño: Descriptivo, transversal y correlacional. Este estudio evaluó el desempeño de una inteligencia artificial (IA) para la clasificación BIRADS, comparando su sensibilidad y especificidad con los resultados obtenidos por médicos radiólogos del Hospital Regional de Loreto.

3.2 Diseño muestral

Población de estudio

La población está compuesta por todas las imágenes mamográficas de pacientes realizadas en el Hospital Regional de Loreto durante el período del 2024. Las mamografías seleccionadas están clasificadas según el sistema BIRADS.

Tamaño de la Muestra

Fórmula general para el cálculo de tamaño de muestra con población finita:

$$n = \frac{Z^2 \; x \; P \; x \; (1 - P)}{d^2}$$

donde:
Donde:

$\mathrm{Z} = 1.96$ (95% de nivel de confianza)

$\mathrm{P} = 0.05$ (prevalencia del 5%)

$\mathrm{D} = 0.02$ (margen de error del 2%)

$$n = \frac{(1.96)^2 \; x \; 0.05 \; x \; (1 - 0.05)}{(0.02)^2}$$

$$n = \frac{3.8416 \; x \; 0.05 \; x \; 0.95}{0.0004}$$

$$n = \frac{3.8416 \; x \; 0.0475}{0.0004}$$

$$n = 456$$

Haciendo un ajuste por pérdidas de un 15% tenemos una muestra de 536 imágenes radiográficas.

Tipo de muestreo y procedimiento de selección de muestra

Se utilizó un muestreo no probabilístico por conveniencia, donde se analizaron todas las mamografías digitalizadas disponibles durante el período de estudio. Esta estrategia es adecuada debido a las limitaciones logísticas y a la necesidad de obtener datos específicos para evaluar el desempeño de la IA en un entorno clínico real.

Criterios de Selección

Inclusión

- Mamografías con clasificación BIRADS realizada previamente por radiólogos.
- Pacientes mayores de 18 años.
- Imágenes en formato digital y con calidad suficiente para el análisis.

Exclusión

- Mamografías con artefactos o de mala calidad que no permitan una evaluación adecuada.
- Mamografías que no cuenten con la clasificación de BIRADS realizada previamente por radiólogos u oncólogos.

3.3 Procedimiento, técnicas e instrumentos de recolección de datos

Instrumento y Fuentes de datos:

Algoritmo de IA (Chatbot)

Se empleó un algoritmo con Inteligencia Artificial previamente desarrollado como parte de otro estudio, adaptado para la clasificación de mamografías.

El objetivo principal del Chatbot fue asignar una categoría BIRADS a cada imagen mamográfica.

Ficha de recolección de datos

Se utilizó una ficha de recolección diseñada por el equipo de investigación.

Esta ficha no cuenta con un proceso formal de validación y confiabilidad, dado que se trata de un instrumento de uso interno para sistematizar los resultados.

En dicha ficha se registraron las características de cada imagen, la categoría BIRADS asignada tanto por la IA como por los especialistas.

Base de datos de mamografías

Se accedió al archivo digital de mamografías del Hospital Regional de Loreto, que cuenta con estudios previos de pacientes diagnosticadas durante el período de investigación.

Procedimiento

Selección de mamografías digitalizadas

Se identificó y seleccionó las mamografías de pacientes atendidas en el Hospital Regional de Loreto durante el período del estudio.

Todas las imágenes seleccionadas cumplieron con criterios de calidad y disponibilidad de diagnóstico clínico/radiológico previo.

Evaluación por especialistas

Un equipo de médicos radiólogos del hospital revisó cada mamografía y la clasifico siguiendo las categorías BIRADS.

Esta clasificación se estableció como el estándar de referencia (“Gold standard”) para la comparación posterior con el diagnóstico provisto por la IA.

Clasificación con inteligencia artificial

El algoritmo de IA procesó las mismas mamografías e igualmente asigno una categoría BIRADS.

El Chatbot utilizó técnicas de visión por computador y/o redes neuronales para el reconocimiento de patrones en las imágenes.

Etiquetado de resultados

Para cada mamografía, se registró el resultado emitido por la IA y el resultado emitido por el especialista, junto con los datos clínicos disponibles.

Se generó un identificador único para cada imagen a fin de asegurar la trazabilidad y evitar duplicaciones.

Creación de la base de datos

Toda la información (imágenes, clasificaciones, datos clínicos y metadatos) se centralizó en una única base de datos.

Los registros se exportaron a un archivo de Microsoft Excel para su organización preliminar.

Posteriormente, estos datos se importaron al software estadístico IBM SPSS v29 para su análisis.

Procesamiento y análisis de la información

Almacenamiento y preparación de datos

Los resultados de la clasificación de la IA y de los radiólogos se almacenaron en paralelo en la base de datos, facilitando la comparación directa.

Se verificó la calidad de los datos y se realizó una limpieza preliminar de los registros que pudieran presentar inconsistencias o datos incompletos.

Análisis descriptivo

Se realizó un análisis descriptivo de los datos recolectados con el objetivo de caracterizar la muestra:

- Se calculó medidas de tendencia central y dispersión para la variable edad (media, mediana, desviación estándar).
- Se presentó la frecuencia absoluta y relativa de las clasificaciones BIRADS otorgadas tanto por el radiólogo como por el chatbot de inteligencia artificial.

- Los resultados se representaron mediante tablas de frecuencia y gráficos de barras.

Evaluación de concordancia diagnóstica

Para determinar el nivel de acuerdo entre el chatbot y el radiólogo en la clasificación BIRADS, se utilizó el índice de concordancia Kappa de Cohen (κ), que cuantifica el grado de coincidencia entre dos evaluadores más allá del azar.

Se construyó una tabla de contingencia entre ambas clasificaciones.

El valor de Kappa será interpretado según los rangos estándar (Landis y Koch):

< 0.20: concordancia muy baja 0.21–

0.40: baja

0.41–0.60: moderada

0.61–0.80: buena

0.80: muy buena o excelente

Este análisis permitirá establecer la confiabilidad del chatbot como herramienta diagnóstica.

Evaluación del rendimiento diagnóstico del chatbot

Con el radiólogo como estándar de referencia, se determinaron las siguientes métricas de rendimiento del chatbot:

- Sensibilidad (capacidad para detectar casos sospechosos de malignidad)
- Especificidad (capacidad para identificar correctamente los casos no sospechosos)

- Exactitud (proporción total de clasificaciones correctas)

Para este análisis, se agruparán las categorías BIRADS de la siguiente manera:

- Positivo o sospechoso: BIRADS 4, 5 y 6
- Negativo o no sospechoso: BIRADS 1, 2 y 3

Se construirá una matriz de confusión para obtener los valores de:

- Verdaderos positivos (TP)
- Verdaderos negativos (TN)
- Falsos positivos (FP)
- Falsos negativos (FN)

Análisis de errores diagnósticos

Se analizaron las tasas de error del chatbot:

- Tasa de falsos positivos (FP): cuando el chatbot clasifica como sospechoso, pero el radiólogo no.
- Tasa de falsos negativos (FN): cuando el chatbot no detecta una lesión sospechosa identificada por el radiólogo.

Este análisis permitió identificar el potencial riesgo clínico del uso del chatbot como herramienta de apoyo.

Representación de resultados

Los resultados obtenidos fueron presentados en:

- Tablas de resumen estadístico.
- Gráficos de barras y gráficos circulares para distribución de categorías.

- Gráficos de matriz de confusión para visualización del rendimiento del chatbot.
- Informes tabulados comparativos entre el rendimiento del chatbot y el radiólogo.

3.4 Aspectos éticos

Este estudio ha sido aprobado por el Comité de Ética del Hospital Regional de Loreto. Las mamografías se obtuvieron de manera retrospectiva y se anonimizaron, eliminando todo dato identificable. Dado el diseño observacional y el uso exclusivo para validación de la IA (Inteligencia artificial), no se solicitó una exención de consentimiento informado individual, fundamentada en el mínimo riesgo y la inviabilidad de contactar a las pacientes. Los datos se almacenaron en servidores seguros con acceso restringido al equipo investigador, y se eliminarán tras 5 años. El estudio cumplió con la Declaración de Helsinki y la Ley de Protección de Datos Personales del Perú (Ley N° 29733).

CAPITULO IV: RESULTADOS

Tabla 1. Características de los Conjuntos de Datos para Evaluación.

Características	**N (%)**
Conjunto de Datos: Modelo GMIC	
Resumen del Procesamiento de Datos	
Imágenes iniciales procesados	1207
Imágenes finales procesados	854
Distribución a Nivel Paciente (n = 427 pacientes)	
Clase de Referencia (Binaria)	
Negativo (BI-RADS 0-3)	408 (95,6%)
Positivo (BI-RADS ≥4)	19 (4,4%)
Distribución a Nivel Mama (n = 854 imágenes/lados)	
Clase de Referencia (Binaria)	
Negativo (BI-RADS 0-3)	835 (97,8%)
Positivo (BI-RADS ≥4)	19 (2,2%)

Fuente: Datos recopilados por el presente estudio (Hospital Regional de Loreto)

Durante el período de evaluación se procesaron inicialmente 1207 imágenes, de los cuales se generaron 854 exámenes válidos. Tras el proceso de emparejamiento con la base de datos de referencia, se evaluaron 427 pacientes únicos correspondientes a 854 mamas analizadas.

Tabla 2. Matriz de confusión. Se muestra un desglose detallado de las predicciones correctas e incorrectas para cada clase

Real \ Predicho	Negativo	Positivo	Total real
Negativo	TN=386	FP=449	835
Positivo	FN=8	TP=11	19
Total predicho	394	460	854

Fuente: resultados reales de la base de datos, con los resultados predictivos de la IA

La matriz de confusión, que comprende 854 casos analizados, ilustra el desempeño de un modelo predictivo. Se registraron 386 verdaderos negativos (TN) y 11 verdaderos positivos (TP), lo que indica clasificaciones correctas. Sin embargo, se detectaron 449 falsos positivos (FP) y 8 falsos negativos (FN), lo cual constituye las clasificaciones erróneas. Del total de casos reales, 835 fueron negativos y 19 positivos, en tanto que las predicciones del modelo arrojaron 394 negativos y 460 positivos.

Tabla 3. Métricas principales con IC95%. Intervalo de confianza del 95%, esta tabla muestra un rango de valores donde es probable que se encuentre el valor real del parámetro en la población general, con un 95% de confianza.

Métrica	Valor	IC95%
Exactitud	46,49%	43,17–49,84%
Sensibilidad	57,89%	36,28–76,86%
Especificidad	46,23%	42,87–49,62%
VPP	2,39%	1,34–4,23%
VPN	97,97%	96,05-98,97%
Prevalencia (BI-RADS ≥4)	2,22%	1,43–3,45%

Fuente: base de datos del Hospital Regional de Loreto, cruzado con los resultados de la inteligencia artificial.

La exactitud fue del 46,49% (IC95%: 43,17–49,84%), lo cual representa la proporción de predicciones correctas. La sensibilidad fue del 57,89% (IC95%: 36,28–76,86%), indicando la capacidad para identificar casos positivos. La especificidad fue del 46,23% (IC95%: 42,87–49,62%), reflejando la capacidad para identificar casos negativos. El Valor Predictivo Positivo (VPP) fue del 2,39% (IC95%: 1,34–4,23%), mostrando la probabilidad de que un resultado positivo sea verdadero. El Valor Predictivo Negativo (VPN) alcanzó un 97,97% (IC95%: 96,05–98,97%), representando la probabilidad de que un resultado negativo sea verdadero. La prevalencia de BI-RADS ≥4 fue del 2,22% (IC95%: 1,43–3,45%), indicando la proporción de casos positivos en la población estudiada.

Tabla 4. Métricas complementarias. Agrupación de métricas, de evaluación de modelos de clasificación utilizada en el campo del aprendizaje automático, especialmente útil cuando se trabaja con conjuntos de datos desbalanceados.

Métrica	Valor
Exactitud balanceada	52,06%
F1-score	4,59%
MCC (Coeficiente de Correlación de Matthews)	0,012

Fuente: Resultados de la inteligencia artificial.

En relación con las métricas complementarias, se presenta una Exactitud Balanceada (Balanced Accuracy) del 52,06%, una medida que ajusta el desempeño del modelo para clases desequilibradas. El puntaje F1-score se sitúa en un 4,59%, reflejando el equilibrio entre la precisión y la exhaustividad. Finalmente, el Coeficiente de Correlación de Matthews (MCC), con un valor de 0,012, ofrece una evaluación integral de la calidad de las clasificaciones binarias, considerando todos los tipos de predicciones correctas e incorrectas.

Tabla 5. Tasas de error. Métrica que mide la frecuencia con la que ocurren fallos o errores en un proceso, sistema o conjunto de datos durante un período de tiempo determinado.

Indicador	Valor
FPR* (1 – Especificidad)	53,77%
FNR (1 – Sensibilidad)**	42,11%
FDR° (1 – PPV)	97,61%
FOR°° (1 – NPV)	2,03%

***** Tasa de Falsos Positivos (FPR). ** Tasa de Falsos Negativos (FNR). ° Tasa de Falsos Descubrimientos (FDR). °°Tasa de Omisión de Falsos (FOR).

Analizando las tasas de error, se observa una Tasa de Falsos Positivos (FPR) del 53,77%, calculada como uno menos la Especificidad. La Tasa de Falsos Negativos (FNR), que es uno menos la Sensibilidad, se sitúa en un 42,11%. Además, la Tasa de Falsos Descubrimientos (FDR), correspondiente a uno menos el Valor Predictivo Positivo (PPV), asciende a un 97,61%. Finalmente, la Tasa de Omisión de Falsos (FOR), o uno menos el Valor Predictivo Negativo (NPV), es del 2,03%.

Tabla 6. Likelihood ratios y DOR. Cuadro que evalúa la utilidad de una prueba, combinando su sensibilidad y especificidad en una única métrica que indica cuánto influye el resultado en la probabilidad de que una enfermedad esté presente.

Indicador	Estimación	IC95%
LR+	1,08	0,73–1,59
LR−	0,91	0,53–1,55
DOR	1,18	0,47–2,97

LR: Likehood ratio (razón de probabilidad). Nota: AUC reportada: **0,5332** (sin IC por no disponer de puntajes continuos).

Respecto a los likelihood ratios y el Differential Odds Ratio (DOR), junto con sus intervalos de confianza del 95% (IC95%) calculados mediante el método logarítmico, se observa que el LR+ es de 1,08 (IC95%: 0,73–1,59). Por su parte, el LR− se sitúa en 0,91 (IC95%: 0,53–1,55). El DOR es de 1,18 (IC95%: 0,47-2,97). Adicionalmente, se reporta un Área bajo la Curva (AUC) de 0,5332, aunque no se incluye un IC para este valor debido a la ausencia de puntajes continuos.

Tabla 7. Acuerdo con el radiólogo (Cohen's κ). Cuadro que mide el grado de acuerdo o concordancia entre el evaluador o instrumentos de medida, en este caso el medico radiólogo y los resultados del IA.

Indicador	Valor
Acuerdo observado (P_o)	46,49%
Acuerdo esperado (P_e)	46,31%
Kappa de Cohen (κ)	0,003
IC95% κ (aprox.)	−0,059 a 0,066

Fuente: Base de datos interpretados por la IA y porcentaje de concordancia con el medico radiólogo

Al evaluar el acuerdo diagnóstico con el radiólogo, se obtiene un acuerdo observado (P_o) del 46,49%, en contraste con un acuerdo esperado por azar (P_e) del 46,31%. El coeficiente Kappa de Cohen (κ) resultante es de 0,003, con un intervalo de confianza del 95% (IC95%) que abarca desde −0,059 hasta 0,066.

Tabla 8. Prueba de McNemar (discordancias pareadas). Métrica para comparar dos resultados en diferentes momentos. (Resultados del radiólogo vs resultados de la IA)

b = FN	c = FP	χ^2 (corrección continuidad)	p-valor
8	449	423,63	<0,001

Fuente: base de datos reales y los resultados de la IA. Resultados cruzados. Se estima el cambio en una característica de las mismas imágenes mamográficas antes y después de la intervención de la IA.

La prueba de McNemar, diseñada para evaluar las discordancias pareadas, detalla que el número de falsos negativos (b = FN) fue de 8, mientras que el de falsos positivos (c = FP) ascendió a 449. La prueba chi-cuadrado (χ^2) con corrección de continuidad arrojó un valor de 423,63, asociado a un p-valor de <0,001.

Tabla 9. Evaluación mediante clasificación binaria, diferenciando casos benignos (BI-RADS 0-3) de malignos (BI-RADS ≥4).

Métrica	Valor
Exactitud	0.4649
Precisión	0.0239
F1-Score	0.0459
AUC	0.5332

Fuente: Base de datos combinado con resultados de la IA

Los resultados obtenidos muestran un rendimiento limitado del modelo GMIC para la clasificación BI-RADS en el contexto estudiado. La accuracy de 46.49% indica que menos de la mitad de las clasificaciones fueron correctas, mientras que el AUC de 0.5332 sugiere una capacidad discriminativa apenas superior al azar.

Tabla 10. Matriz de Confusión - Modelo GMIC (La magia de GREYC para la computación de imágenes)

Real \ Predicho	Negativo	Positivo
Negativo	386	449
Positivo	8	11

Fuente: Resultados de la IA

El modelo GMIC clasificó correctamente 386 casos negativos y 11 positivos, mientras que se equivocó en 449 negativos que marcó como positivos y en 8 positivos que consideró negativos, mostrando así un alto número de falsos positivos y una baja capacidad para identificar los casos positivos reales.

Tabla 11. Análisis de sensibilidad y especificidad

Indicador	Resultado	Detalle
Sensibilidad	57.89 %	11 de 19 casos malignos detectados
Especificidad	46.23 %	386 de 835 casos benignos correctamente identificados
Valor Predictivo Positivo	2.39 %	11 de 460 predicciones positivas fueron correctas
Valor Predictivo Negativo	97.97 %	386 de 394 predicciones negativas fueron correctas

Fuente: Datos recopilados por el presente estudio (Hospital Regional de Loreto)

Los indicadores diagnósticos muestran que el modelo GMIC alcanza una sensibilidad de 57.89 %, lo que refleja que identifica poco más de la mitad de los casos positivos reales, mientras que la especificidad de 46.23 % indica una capacidad limitada para reconocer correctamente a los negativos. El valor predictivo positivo es muy bajo (2.39 %), evidenciando que casi todas las predicciones positivas son erróneas, en contraste con el valor predictivo negativo de 97.97 %, que demuestra un buen desempeño al confirmar los casos negativos.

Tabla 12. Evaluación de falsos positivos y negativos

Error diagnóstico	Resultado	Detalle
Falsos positivos	449 casos	53.77 % de los casos negativos
Falsos negativos	8 casos	42.11 % de los casos positivos
Índice Kappa Cohen	0.0033	Concordancia prácticamente nula

Fuente: Datos recopilados por el presente estudio (Hospital Regional de Loreto)

La evaluación de errores diagnósticos revela que el modelo produce un elevado número de falsos positivos (449 casos, 53.77 % de los negativos) y mantiene también un nivel considerable de falsos negativos (8 casos, 42.11 % de los positivos). El índice Kappa de Cohen es de 0.0033, lo que indica que la concordancia entre lo predicho y lo real es prácticamente nula.

CAPITULO V: DISCUSIÓN

El análisis de los resultados de este estudio se centró en evaluar la eficacia de una inteligencia artificial (IA), implementada a través de un chatbot (IA), para la clasificación BI-RADS en el tamizaje de cáncer de mama en el Hospital Regional de Loreto. Los hallazgos obtenidos proporcionan una base para comprender su desempeño y potenciales implicaciones clínicas.

Al comparar la clasificación BI-RADS generada por el chatbot (IA) con la establecida por los médicos radiólogos, se observa un rendimiento limitado del modelo GMIC (GREYC's Magic for Image Computing) en esta tarea. La matriz de confusión reveló 449 falsos positivos y 8 falsos negativos, lo que se traduce en un 53,77% de los casos negativos clasificados incorrectamente como positivos por el chatbot y un 42,11% de los casos positivos reales que fueron pasados por alto. Estos números evidencian una diferencia significativa en las clasificaciones del sistema automatizado respecto al criterio clínico. Al contrastar estos resultados con la literatura internacional, donde diversos estudios han demostrado capacidades notablemente superiores, como la precisión diagnóstica del 96% en un estudio por Escalante M. (2023) (18), o las precisiones del 90% y 93% por Rodríguez et al. (2021) en la clasificación histopatológica (16). Incluso a nivel nacional, trabajos como el de Garrafa Olea (2023) informaron precisiones de hasta 99% en la clasificación de densidad mamaria (12), y un sistema en Ramón J (2021) alcanzó un 88,5% de exactitud (9). Nuestro bajo Valor Predictivo Positivo (VPP) de 2,39% refleja esta limitación crítica.

En la evaluación de métricas de desempeño, la sensibilidad del modelo se situó en un 57,89%, indicando que el chatbot es capaz de identificar poco más de la

mitad de los casos malignos reales. Por su parte, la especificidad fue del 46,23%, lo cual sugiere una capacidad limitada para identificar correctamente los casos negativos. Estos valores, obtenidos en nuestra investigación, distan considerablemente de las sensibilidades y especificidades alcanzadas en otros modelos evaluados en el contexto internacional. Por ejemplo, el modelo D5_v1 de Sáenz JA (2023) obtuvo una sensibilidad del 87% y una especificidad del 92% (17). La baja especificidad en nuestro caso conlleva una alta Tasa de Falsos Positivos (FPR) del 53,77%. La concordancia diagnóstica, cuantificada mediante el coeficiente Kappa de Cohen, arrojó un valor de 0,003, lo cual indica una concordancia prácticamente nula entre las predicciones del chatbot y la clasificación experta del radiólogo. Esta falta de acuerdo se refuerza con el elevado número de discordancias detectadas a través de la prueba de McNemar, la cual mostró una diferencia estadísticamente muy significativa entre la proporción de falsos positivos y falsos negativos. Estos hallazgos de nuestra investigación contrastan con estudios donde los sistemas de IA, como el de Carrilero M (2022) con una precisión de detección del 96.5% y niveles de concordancia equivalentes a la intercorrelación entre profesionales, han demostrado un desempeño superior (20).

Asimismo, se reveló una Tasa de Falsos Positivos (FPR) del 53,77%, lo que significa que más de la mitad de los casos negativos fueron incorrectamente clasificados como positivos. La Tasa de Falsos Negativos (FNR) fue del 42,11%, indicando que un porcentaje considerable de casos positivos reales fueron clasificados incorrectamente como negativos. El Valor Predictivo Positivo (VPP) de apenas 2,39% implica que una proporción considerable de las predicciones catalogadas como positivas resultan ser erróneas, lo que se traduce en una

elevada Tasa de Falsos Descubrimientos (FDR) del 97,61%. En contraste, el Valor Predictivo Negativo (VPN) alcanzó un 97,97%, con una Tasa de Omisión de Falsos (FOR) del 2,03%.

Las limitaciones inherentes a este estudio incluyen la necesidad de disponer de conjuntos de datos más amplios y heterogéneos para optimizar las fases de entrenamiento y validación del modelo. Una base de datos más robusta podría contribuir a mejorar la capacidad de generalización y el desempeño predictivo. Asimismo, la optimización de los parámetros del modelo, tal como se ha sugerido en investigaciones como la de Montoya J (2022) que obtuvo un 66% de precisión con SVM (Maquinas de Soporte Vectorial), podría ser un factor determinante (19). A pesar de que el rendimiento actual del modelo para la clasificación de alto riesgo presenta restricciones, el potencial intrínseco de la inteligencia artificial como herramienta de apoyo en el diagnóstico precoz del cáncer de mama, con la capacidad de agilizar los tiempos de respuesta y mitigar la carga laboral de los especialistas, tal como ha sido demostrado en diversas investigaciones (ej., el estudio en Chiclayo de 2020 con 91.4% de sensibilidad y 85.7% de especificidad), continúa siendo un campo de investigación de gran relevancia y promesa.

CAPITULO VI: CONCLUSIONES

Se validó la eficacia de la inteligencia artificial implementada mediante un chatbot para la clasificación BI-RADS en el tamizaje de cáncer de mama en el Hospital Regional de Loreto, encontrándose una exactitud de 46,49% y un área bajo la curva de 0,5332. Estos resultados reflejaron un rendimiento limitado, condicionado por factores como la calidad de las imágenes y la ausencia de un estándar diagnóstico definitivo mediante biopsia, lo que restringió la posibilidad de confirmar de manera absoluta la validez de las predicciones.

Se comparó la clasificación generada por el chatbot con la de los médicos radiólogos, identificándose una elevada discordancia diagnóstica. El coeficiente kappa de Cohen fue de 0,0033, lo que evidenció concordancia prácticamente nula, mientras que la prueba de McNemar registró diferencias significativas. Este comportamiento se relaciona con la dependencia del modelo respecto a la calidad de los datos de entrenamiento y con la variabilidad en las características de la población local, como la densidad y volumen mamario, que pueden influir en la lectura de las imágenes.

Se evaluó la sensibilidad, especificidad y el grado de concordancia del chatbot frente al criterio del radiólogo, encontrándose una sensibilidad de 57,89% y una especificidad de 46,23%. El valor predictivo positivo alcanzó 2,39% y el valor predictivo negativo 97,97%, con 449 falsos positivos y 8 falsos negativos. Aunque estas cifras no fueron elevadas, ponen en evidencia la capacidad del modelo para descartar casos negativos, aspecto que puede considerarse útil en contextos donde la sobrecarga diagnóstica constituye un problema.

Se determinó la tasa de falsos positivos y falsos negativos, registrándose un 53,77% y 42,11%, respectivamente. Si bien estas proporciones señalaron un desempeño limitado, también permiten plantear la necesidad de estudios posteriores que integren confirmación diagnóstica mediante biopsia y bases de datos más amplias y heterogéneas, de modo que se incremente la capacidad predictiva y la aplicabilidad clínica de la herramienta en escenarios locales.

CAPITULO VII: RECOMENDACIONES

Optimizar la calidad de las imágenes mamográficas en el Hospital Regional de Loreto mediante protocolos estandarizados de captura y almacenamiento digital, de manera que se reduzca la variabilidad en la lectura y se favorezca un mejor desempeño de modelos de inteligencia artificial.

Incorporar la confirmación diagnóstica mediante biopsia en estudios posteriores desarrollados en la Universidad Nacional de la Amazonía Peruana (UNAP), a fin de contar con un estándar de referencia sólido que permita validar con mayor certeza la eficacia de las herramientas de inteligencia artificial aplicadas a la clasificación BI-RADS.

Ampliar las bases de datos locales utilizadas en futuros proyectos de investigación en el Hospital Regional de Loreto, integrando casos con diferentes características de densidad y volumen mamario, para generar un entrenamiento más representativo de la población y mejorar la capacidad de generalización de los algoritmos.

Implementar alianzas de cooperación académica entre la UNAP y centros de investigación internacionales, con el propósito de contrastar el desempeño de modelos de inteligencia artificial en contextos poblacionales diversos y así establecer comparaciones más realistas y adaptadas a la región.

Fortalecer las líneas de investigación en salud digital en el Hospital Regional de Loreto, promoviendo el desarrollo de estudios que combinen inteligencia artificial con criterios clínicos y radiológicos integrados, de modo que la herramienta pueda evolucionar como un apoyo complementario para los especialistas y no como sustituto directo.

CAPITULO VIII: REFERENCIAS BIBLIOGRÁFICAS

1. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, et al. GLOBOCAN 2022 (version 1.1) - 08.02.2024: World Fact Sheet [Internet]. Lyon (FR): International Agency for Research on Cancer; 2024 [citado 2023 Feb 13]. Disponible en: https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf
2. Bustamante Coronado RI. Brechas de acceso al tratamiento oncológico en el Hospital María Auxiliadora durante el año 2019 [Tesis de maestría]. Lima (PE): Universidad Científica del Sur; 2022. Disponible en: TM-Bustamante R-Ext.pdf
3. García KJ, Ocampo JD, Pardo M del P, Aguilar T, Ruiz CA, Castaño A. Calidad de la imagen, la lectura y el servicio de la mamografía en cuatro centros de imágenes diagnósticas de Manizales, Colombia. Biomédica [Internet]. 2021;41:52-64. Disponible en: https://doi.org/10.7705/biomedica.5135
4. Dang LA, Chazard E, Poncelet E, Serb T, Rusu A, Pauwels X, et al. Impact of artificial intelligence in breast cancer screening with mammography. Breast Cancer [Internet]. 2022;29(6):967–77. Disponible en: http://dx.doi.org/10.1007/s12282-022-01375-9
5. Boumaraf S, Liu X, Ferkous C, Ma X. A new computer-aided diagnosis system with modified genetic feature selection for BI-RADS classification of breast masses in mammograms. Biomed Res Int [Internet]. 2020;2020:7695207. Disponible en: https://doi.org/10.1155/2020/7695207
6. Fernandez De Freitas MA, Capecchi A. Inteligencia artificial en la detección del cáncer de mama por tomosíntesis, ¿hacia dónde vamos? Revisión narrativa. Rev Cient CMDLT [Internet]. 2021;15(2):e-211066. Disponible en: https://doi.org/10.55361/cmdlt.v15i2.66
7. del Castillo Carrera S. Métodos computacionales para la interpretabilidad de los resultados bioinformáticos en el ámbito clínico [Trabajo Fin de Grado]. Málaga (ES): Universidad de Málaga; 2022.
8. Quesquén R. Utilización de algoritmos para la identificación automática de microcalcificaciones en imágenes digitales de mamografía [Tesis de

licenciatura]. Pimentel (PE): Universidad Señor de Sipán; 2020.

9. Ramón J. Sistema inteligente para apoyar al análisis mamográfico en la detección de tumores de mama femenino entre las edades de 40 a 60 años en el Hospital “Las Mercedes” de Chiclayo [Tesis de licenciatura]. Chiclayo (PE): Universidad Católica Santo Toribio de Mogrovejo; 2021.
10. Bunnell A, Glaser Y, Valdez D, Wolfgruber T, Altamirano A, González CZ, et al. Learning a clinically-relevant concept bottleneck for lesion detection in breast ultrasound [Internet]. arXiv [cs.CV]. 2024. Disponible en: http://arxiv.org/abs/2407.00267
11. Cortes D, Pérez K. Validación de software para la detección de regiones anormales en imágenes de mamografía [Tesis de licenciatura]. Bogotá (CO): Universidad ECCI; 2023.
12. Garrafa S, Olea Z. Modelos de clasificación de densidad mamaria utilizando redes neuronales convolucionales [Tesis de licenciatura]. Lima (PE): Universidad Peruana de Ciencias Aplicadas; 2023.
13. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature [Internet]. 2020;577(7788):89-94. Disponible en: http://dx.doi.org/10.1038/s41586-019-1799-6
14. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology [Internet]. 2019;292(1):60-6. Disponible en: http://dx.doi.org/10.1148/radiol.2019182716
15. Hill H, Roadevin C, Duffy S, Mandrik O, Brentnall A. Cost-effectiveness of AI for risk-stratified breast cancer screening. JAMA Netw Open [Internet]. 2024;7(9):e2431715. Disponible en: http://dx.doi.org/10.1001/jamanetworkopen.2024.31715
16. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. J Natl Cancer Inst [Internet]. 2019;111(9):916-22. Disponible en: http://dx.doi.org/10.1093/jnci/djy222
17. Sáenz JA. Ensamble de redes neuronales convolucionales para la clasificación BI-RADS de tumores de mama en ultrasonido [Tesis de

maestría]. México (MX): Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional; 2023.

18. Escalante M. Aplicación de la inteligencia artificial para la detección del cáncer de mama. Rev Méd Sinergia [Internet]. 2023;8(12). Disponible en: https://doi.org/10.31434/rms.v8i12.1113
19. Montoya J, Briñez J, Fonnegra R. Desarrollo de un algoritmo para clasificación de lesiones benignas y malignas en imágenes de resonancia magnética de mama usando inteligencia artificial. Rev Cintex. 2022;27(2):69–78.
20. Carrilero M. A deep neural network for describing breast ultrasound images in natural language [Tesis de licenciatura]. Madrid (ES): Universidad Nacional de Educación a Distancia; 2022.
21. Hernández J. Clasificación de lesiones mamográficas a partir del modelado cuantitativo del léxico BI-RADS para masas [Tesis de maestría]. México (MX): Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional; 2021.
22. Navarro PB, Guzmán M. Mama femenina [Internet]. Kenhub; 2021 [citado 2024 Dic 4]. Disponible en: https://www.kenhub.com/es/library/anatomia-es/mama-femenina
23. Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: A summary of changes. Diagn Interv Imaging [Internet]. 2017;98(3):179–90. Disponible en: http://dx.doi.org/10.1016/j.diii.2017.01.001
24. Aibar L, Santalla A, Criado MSL, González-Pérez I, Calderón MA, Gallo JL, et al. Clasificación radiológica y manejo de las lesiones mamarias. Clin Invest Ginecol Obstet [Internet]. 2011;38(4):141–9. Disponible en: http://dx.doi.org/10.1016/j.gine.2010.10.016
25. Centers for Disease Control and Prevention (CDC). Diagnóstico del cáncer de mama [Internet]. Atlanta (US): CDC; 2024 [citado 2024 Dic 4]. Disponible en: https://www.cdc.gov/breast-cancer/es/screening/diagnosis.html
26. Radiopaedia. Breast Imaging Reporting and Data System (BI-RADS) [Internet]. 2024 [citado 2025 Ago 30]. Disponible en: https://radiopaedia.org/articles/breast-imaging-reporting-and-data-system-bi-rads-2?lang=us

27. Rodríguez-Ruiz A, Mann RM, Sechopoulos I. Mammography-based artificial intelligence for breast cancer detection and diagnosis: clinical implementation challenges and opportunities. Insights Imaging [Internet]. 2025 [citado 2025 Ago 30];16(1):46. Disponible en: https://insightsimaging.springeropen.com/articles/10.1186/s13244-025-01983-x
28. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature [Internet]. 2020 [citado 2025 Ago 30];577(7788):89-94. Disponible en: https://www.nature.com/articles/s41591-024-03408-6
29. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med [Internet]. 2005 [citado 2025 Ago 30];37(5):360-3. Disponible en: https://pubmed.ncbi.nlm.nih.gov/15883903/
30. Yu X, Chen J, Zhang Y, Li Z, Wang X, Yang J, et al. A multi-modal AI system for screening mammography: development, validation, and prospective deployment across 18 sites. arXiv [Preprint]. 2024 [citado 2025 Ago 30]. Disponible en: https://arxiv.org/abs/2504.05636

ANEXOS

ANEXO N° 01: MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADOR	ESCALA
General ¿Cuál es la eficacia de una Inteligencia Artificial usando un chatbot para la clasificación BIRADS como herramienta	**General** Validar la eficacia de una inteligencia artificial usando un chatbot para la clasificación BIRADS en el tamizaje de cáncer de mama en el Hospital Regional de Loreto, 2024. **Específicos**	H1: El chatbot basado en inteligencia artificial es eficaz y preciso en la clasificación BI-RADS para el tamizaje de cáncer de	Precisión del diagnóstico del chatbot	Porcentaje de coincidencia	Racional

para el tamizaje de Cáncer de Mama, en Loreto 2024?	1. Comparar la clasificación BIRADS realizada por el chatbot con la de los médicos radiólogos en el tamizaje de cáncer de mama. 2. Evaluar la sensibilidad, especificidad y el grado de concordancia diagnóstica del chatbot frente al criterio del radiólogo.	mama en pacientes atendidas en el Hospital Regional de Loreto, durante el año 2024.			
		H0: El chatbot basado en inteligencia artificial no es	Sensibilidad del Chatbot	Proporción de verdaderos positivos entre el total de casos positivos reales	Razón

	3. Determinar la tasa de falsos positivos y falsos negativos generados por el chatbot en la clasificación BIRADS.	eficaz ni preciso en la clasificación BI-RADS para el tamizaje de cáncer de mama en pacientes atendidas en el Hospital Regional de Loreto, durante el año 2024		[VP / (VP + FN)]	
			Especificidad del Chatbot	Proporción de verdaderos negativosentre el total de casos negativos reales [VN / (VN + FP)]	Razón
			Tasa de falsos positivos/negativos	Porcentaje de falsos positivos/negativos	Racional
			Tiempo de respuesta del	Tiempo medido en segundos durante el proceso de	Intervalo

			Chatbot	interacción.	
			Capacidad de detección por categorías BIRADS	Porcentaje de clasificaciones correctas por categoría	Intervalo

Anexo 2: Ficha de recolección de datos

UNAP

Estudio: VALIDACIÓN DE UNA INTELIGENCIA ARTIFICIAL USANDO CHATBOT PARA LA CLASIFICACIÓN BIRADS EN EL TAMIZAJE DE CÁNCER DE MAMA EN LORETO, 2024

Hospital: Hospital Regional de Loreto

Año: 2025

Datos Generales del Paciente

- **Código del paciente:** _____
- **Edad:** _____años
- **Fecha de mamografía:** ____/ ____/ _____

Resultados diagnósticos

- **Clasificación BIRADS asignada por el radiólogo:**
  ☐ BIRADS 1 ☐ BIRADS 2 ☐ BIRADS 3
  ☐ BIRADS 4 ☐ BIRADS 5 ☐ BIRADS 6
- **Clasificación BIRADS generada por el chatbot:**
  ☐ BIRADS 1 ☐ BIRADS 2 ☐ BIRADS 3
  ☐ BIRADS 4 ☐ BIRADS 5 ☐ BIRADS 6
- **¿Coinciden ambas clasificaciones?**
  ☐ Sí
  ☐ No

Variables analíticas derivadas

- **Clasificación final del radiólogo:**
  ☐ Sospechoso (BIRADS 4, 5, 6)
  ☐ No sospechoso (BIRADS 1, 2, 3)
- **Clasificación final del chatbot:**
  ☐ Sospechoso (BIRADS 4, 5, 6)
  ☐ No sospechoso (BIRADS 1, 2, 3)

- **Resultado analítico del chatbot respecto al radiólogo (marcar solo una):**
  ☐ Verdadero Positivo (TP)
  ☐ Verdadero Negativo (TN)
  ☐ Falso Positivo (FP)
  ☐ Falso Negativo (FN)

Observaciones adicionales (si aplica):
