Solución I.N.C. - Sistema de Extracción y Estructuración de Datos de Mamografías

Daniel Pareja Franco, Eduen José Flórez Mariño, Fabian Stiv Peña Gonzales, Giovanni Esteban Moreno Urbina, Juan Pablo Mateus Pardo

Supervisores: Daniel Garzón Rodríguez, María Alejandra Escobar Holguín

Índice

1. Introducción

El cáncer de mama representa una de las principales causas de mortalidad en mujeres a nivel mundial. La detección temprana mediante mamografías es crucial para mejorar las tasas de supervivencia y los resultados del tratamiento. La información crítica contenida en los informes de mamografías se encuentra principalmente en formato de texto no estructurado, lo que dificulta significativamente su análisis sistemático y aprovechamiento para investigación, seguimiento clínico y desarrollo de modelos predictivos. El presente documento detalla el desarrollo de un sistema especializado para la extracción y estructuración automatizada de información clínica relevante a partir de bloques de texto no estructurado en informes de mamografías. Esta solución tecnológica busca transformar el proceso actual, altamente dependiente de la revisión y digitalización manual, en un flujo de trabajo automatizado que genere bases de datos estructuradas con alta precisión, consistencia y eficiencia, permitiendo así potenciar la investigación e implementación de modelos de inteligencia artificial para la detección y seguimiento de patologías mamarias. Para abordar este desafío, el equipo de desarrollo evaluó y comparó diversos modelos de lenguaje grandes (LLM) como Deepseek V3, OpenAI (gpt 4o mini) y Phi-4, analizando su capacidad para interpretar correctamente la terminología radiológica especializada, identificar entidades clínicas relevantes y extraer relaciones semánticas complejas en el contexto de informes mamográficos. La evaluación incluyó métricas de rendimiento como precisión en la extracción de variables específicas, tiempo de procesamiento contabilizando el tiempo empleado por el modelo para clasificar los datos contra el tiempo promedio que toma realizarlo manualmente y capacidad de manejo de ambigüedades lingüísticas comparando las salidas del modelo con los datos clasificados manualmente. Esta fase comparativa permitió seleccionar la solución más apropiada para las necesidades particulares del proyecto y ajustar elementos específicos como la estructura de los "prompts", el formato de salida, el número de ejemplos en contexto, y la inclusión de instrucciones específicas de dominio radiológico para maximizar la efectividad en la extracción de información de mamografías. La implementación de este sistema representa un avance significativo en la gestión de información clínica. Según un experimento realizado por el grupo de desarrollo, el proceso 3 manual de extracción y estructuración de estos datos de imágenes mamográficas toma aproximadamente 450 horas de trabajo para procesar 15,000 registros, con un promedio de 6 horas por cada 200 registros. En contraste, el sistema automatizado puede procesar el mismo volumen en aproximadamente 5.3 horas. Esto significa que el sistema es aproximadamente 85 veces más rápido que el método manual tradicional, transformando un proceso que requería 450 horas (56 días laborales) en uno que demanda solo 5.3 horas. Desde la perspectiva económica, el impacto es igualmente significativo. Según el estudio de Hassanpour y Langlotz (2016) publicado en Artificial Intelligence in Medicine [1], el procesamiento manual y estructuración de datos radiológicos representa uno de los mayores costos operativos en departamentos de radiología, estimando un costo promedio entre USD $5.5-8.5 por informe cuando se consideran todos los recursos involucrados (tiempo del personal, capacitación, supervisión y control de calidad). En marcado contraste, el costo computacional utilizando el modelo DeepSeek V3 es de apenas $23,865 pesos colombianos, mientras que con GPT-4o mini el costo estimado asciende a $26,508 pesos colombianos.

2. Definición del Problema

Este problema representa una barrera crítica para la investigación y el desarrollo de sistemas de apoyo diagnóstico basados en aprendizaje automático, ya que la información valiosa permanece atrapada en textos clínicos no estructurados, dificultando su aprovechamiento sistemático. A continuación, se detalla el contexto específico, los objetivos que busca alcanzar este proyecto y las restricciones particulares.

2.1. Contexto y objetivos del proyecto

El contexto actual de la gestión de datos en mamografías se caracteriza por:

- Información crítica en texto no estructurado: Según la información proporcionada por el Instituto Nacional de Cancerología (INC) durante el reto Sabana Hack [2]. Aproximadamente el 70clínica se encuentra en bloques de texto libre dentro de los informes radiológicos.
- Proceso manual ineficiente: Actualmente, profesionales de la salud deben revisar, interpretar y digitalizar manualmente la información relevante en bases de datos paralelas estructuradas, consumiendo tiempo valioso del personal clínico.
- Inconsistencias en la documentación clínica: Los informes presentan frecuentemente errores tipográficos, variaciones en la terminología, formatos inconsistentes y codificación incorrecta, dificultando su procesamiento sistemático.
- Necesidad de estructuración para IA: El desarrollo de modelos avanzados de inteligencia artificial, como los modelos de lenguaje grande (LLM) como ChatGPT 4o Mini o DeepSeek V3, para apoyo diagnóstico, pueden ser usados para aportar datos estructurados de alta calidad que integren tanto las observaciones de las imágenes mamográficas como las variables clínicas asociadas. En este contexto, el proyecto tiene como objetivo Implementar mecanismos de reconocimiento y estandarización para 15 categorías críticas de hallazgos mamográficos (nódulos, calcificaciones, asimetrías, etc.) según criterios BI-RADS [3].

3. Solución Implementada

Descripción de los datos. El INC entrego inicialmente, para el desarrollo del modelo un archivo Excel “BASE_DE_DATOS_2024”, con dos hojas. La primera hoja “archivo_actualizado” 15001 registros, referentes a mamografías de pacientes identificados en la columna “RAPACIENTE” donde la columna “ESTUDIO” contiene lecturas de mamografías para cada paciente, dicha columna es el objeto del modelo, es decir, a partir de esta se debe generar la data estructurada. Adicional hay otras 5 columnas sobre información del paciente. La segunda hoja “Hoja1” contiene data estructurada a partir de la columna “ESTUDIO” el identificador de cada paciente se encuentra en la columna “ID_PACIENTE”. Dichos datos no fueron adecuados para la medición de las métricas del modelo por las siguientes inconsistencias: Identificadores de la hoja “archivo actualizado” no contenidos en la hoja “Hoja1”. En los identificadores coincidentes imposibilidad de hacer correspondencia biunívoca entre el dictamen estructurado y el no estructurado dado la disparidad en el número de observaciones en ambas hojas y la imposibilidad de

ordenarlos, por no tener una variable temporal o similar, además, se encuentra un número muy pequeño de ID coincidentes. El día 26 de marzo a las 14:37 pm, el INC hace entrega de una nueva base de datos para hacer una medición de las métricas, dicha base de datos es un archivo Excel 5 "BASEDatosMamografia" (acompañada de archivo formato JSON) con dos hojas, "base de datosJSON", "Hoja 1", en esta base lo identificadores son las columnas "RAPACIENTE", "RA" respectivamente. La base de datos presenta el mismo inconveniente de la primera base de datos: En los identificadores coincidentes imposibilidad de hacer correspondencia biunívoca entre el dictamen estructurado y el no estructurado dado la disparidad en el número de observaciones en ambas hojas y la imposibilidad de ordenarlos al no haber una variable temporal o similar en ambas hojas. Por las razones anteriormente expuestas entregamos métricas medidas sobre 600 registros extraídos manualmente por personal sin formación médica. Se recomienda una organización adecuada de los datos para tener un insumo pertinente para entrenamiento de otras opciones de modelos, y para calcular las métricas de manera idónea en la solución aportada.

3.1. Metodología utilizada:

Se empleó una metodología de desarrollo basada en servicios web para la extracción de información estructurada desde notas médicas contenidas en archivos CSV. La solución está orientada a la integración de modelos de lenguaje natural (LLMs) mediante APIs RESTful, permitiendo una interacción directa con los modelos para procesar el texto médico de forma automatizada.

3.2. Modelo o algoritmo seleccionado.

La solución ofrece tres alternativas de procesamiento de texto:

- OpenAI GPT (vía API): Utilizado en app.py, que conecta con modelos propietarios de OpenAI (gpt 4o mini) para el análisis del texto.
- Deepseek V3: Utilizado en appd.py, que conecta con modelos propietarios de Deepseek para el análisis del texto.
- Modelo Phi-4 (Ollama local): Utilizado en local.py, que permite ejecutar modelos en entorno local sin necesidad de conexión a servicios externos, usando el modelo Phi-4 hospedado con Ollama.

4. Resultados y Evaluación

4.1. Datos Utilizados

Se utilizaron los siguientes conjuntos de datos para el análisis de los resultados obtenidos al ejecutar los 3 modelos:

Cuadro 1: *Datos utilizados*

Fuente	**Registros**	**Descripción**
Transcripciones.csv	599	Transcripciones originales - referencia (humano)
respuestaDeepSeek.csv	599	Interpretaciones generadas por DeepSeek
respuestaOpenAI.csv	599	Interpretaciones generadas por OpenAI
respuestalocal.csv	599	Interpretaciones generadas por modelo local Phi4

4.2. Metodología

Se evaluaron los modelos utilizando tres métricas principales:

1. Precisión: Exactitud en la clasificación BIRADS y otros campos radiológicos relevantes

2. Tasa de alucinaciones: Porcentaje de campos donde el modelo proporciona información que no coincide con las transcripciones de referencia

3. Precisión por campo: Evaluación detallada de la precisión en la identificación de hallazgos específicos

4.3. Resultados

Para revisar los resultados completos del modelo, incluyendo precisión por clasificación BIRADS, tasa de alucinaciones, análisis por campos radiológicos específicos, y comparativa de precios entre modelos, por favor consulta la documentación detallada en el README del repositorio de GitHub. Además, puedes explorar los resultados y visualizaciones de forma interactiva a través de la siguiente página web: Sitio Web resultados del proyecto

5. Implementación y Reproducibilidad

El instructivo detallado de la implementación paso a paso de la solución planeteada se encuentra documentada en el README del repositorio de GitHub.

6. Trabajo futuro

6.1. Posibles mejoras en el futuro:

Como continuación del proyecto, además de mejorar la eficiencia en el manejo de las historias clínicas, el modelo propuesto se podría integrar con los sistemas existentes en los hospitales o clínicas, para permitir que, a medida que los profesionales redactan la historia clínica, el modelo extraiga y procese la información en tiempo real.

6.2. Implementación de otras metodologías:

A partir de una base de datos lo suficientemente grande y depurada se pueden entrenar modelos de reconocimiento de entidades nombradas, que según la literatura tienen rendimientos similares a los de un LLM, sin los requerimientos (salvo en el entrenamiento)

de hardware que los últimos necesitan, además que dichos modelos pueden ser entrenados y usados en servidores locales, de esta manera, se garantiza el adecuado manejo de las historias clínicas, aunado a lo anterior se puede diseñar 9 software para el acceso y el manejo de los datos obtenidos a partir de la implementación de los modelos.

7. Conclusiones

Se desarrollaron tres modelos para el análisis y estructuración de la información:

- GPT-4o mini (OpenAI)
- DeepSeek V3: Alternativa con menor costo.
- Phi-4 Local (Ollama): Garantiza una mayor privacidad y control total sobre la información.

El desarrollo de estos modelos permite una reducción significativa tanto en costos operativos como en el tiempo de procesamiento de la información.

8. Bibliografías

1. Hassanpour, S., & Langlotz, C. P. (2016). *Information extraction from multi-institutional radiology reports.* Artificial intelligence in medicine.
2. Instituto Nacional de Cancerología. (2025). Reto 2 Instituto Nacional Cancerología...
3. D'Orsi, C. J., et al. (2016). *BI-RADS (5ª ed.).*