



CLIAS

CENTRO DE INTELIGENCIA
ARTIFICIAL Y SALUD
PARA AMÉRICA LATINA
Y EL CARIBE

REPORTE FINAL

Estudio de factibilidad para el desarrollo de un sistema de diagnóstico a distancia y en tiempo real de muestras citológicas de extendidos cervicovaginales fijados (PAPs) asistido por Inteligencia Artificial (IA).

INSTITUTO

Facultad de Cs. Exactas y Naturales,
Universidad de Buenos Aires.

PERÍODO DEL INFORME

20/11/23 - 20/11/24





INTEGRANTES DEL EQUIPO

Dra. Luciana Bruno (IR) (FCEN-UBA)
Dr. Emmanuel Iarussi (UTdT)
Dra. Viviana Siless (UTdT)
Dra. Jimena Medel (Htal. Rivadavia)
Dra. Estefanía Uelf (Htal. Rivadavia)
Dra. Alicia Dos Santos (Htal. Rivadavia)
Dra. Noemí Buosi
Citot. Rosana Vargas (Htal. Rivadavia)
Lic. Juliana Reves (FCEN-UBA)
Est. Hugo Massarolli (FCEN-UBA y UTdT)
Est. Paula Perez Bianchi (FCEN-UBA)
Est. Malena Vázquez Currié (ITBA) Est. Sol Anselmo (ITBA).

COLABORADORES

Ing. Bruno Volcovinsky (ITBA)
Dra. Soledad Fernández (FCEN-UBA)
Dra. Inés Caridi (FCEN-UBA)
Dr. José Ignacio Orlando (UNICEN)
Dr. Diego Fernández Slezack (FCEN-UBA)
Tec. Diego Brunetti (FCEN-UBA)

FCEN-UBA: Facultad de Cs. Exactas y Naturales, Universidad de Buenos Aires.

UTdT: Universidad Torcuato Di Tella.

ITBA: Instituto Tecnológico de Buenos Aires.

UNICEN: Universidad Nacional del Centro de la Provincia de Buenos Aires.



CONTENIDO

01. INTRODUCCIÓN	3
02. OBJETIVOS	4
03. ACTIVIDADES REALIZADAS	4
3.1 CONSTRUCCIÓN DEL DATASET:	5
3.2 DESARROLLO DE LA PÁGINA WEB	5
3.3 EVALUACIÓN DEL CONSENSO ENTRE ANOTADORES	6
3.4 APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE IMÁGENES:	6
04. RESULTADOS	7



01. INTRODUCCIÓN

Este proyecto evaluó la factibilidad del uso de herramientas de Inteligencia Artificial (IA) para apoyar el diagnóstico del cáncer cervicouterino y mejorar el acceso a la salud en poblaciones vulnerables, utilizando datos nacionales. En ese sentido, el proyecto constituye un primer paso hacia el desarrollo de un sistema automatizado de análisis de muestras de extendidos citológicos de pruebas de Papanicolaou basado en IA. Este enfoque ha sido aplicado en pruebas de concepto a nivel clínico en países como China y Kenia.

Actualmente, el tiempo necesario para el análisis manual de las muestras de PAPs es variable. En la práctica clínica diaria, la evaluación de los extendidos cervicovaginales representa una de las principales cargas de trabajo para el personal especializado.

02. OBJETIVOS

El objetivo general fue desarrollar un sistema automatizado de análisis de muestras de extendidos citológicos de pruebas de Papanicolaou basado en IA a partir de imágenes digitalizadas de las muestras, para asistir en el diagnóstico y contribuir al acceso a la salud de población vulnerable.

Los objetivos específicos planteados fueron:

1. Establecer un protocolo de digitalización de muestras citológicas existentes en el Hospital Bernardino Rivadavia.
2. Proponer criterios de selección y anonimización de muestras para entrenar el sistema de IA.
3. Entrenar y validar un sistema de IA para detectar patología cervical.
4. Evaluar la factibilidad de implementación en un hospital público y desarrollar un prototipo funcional.
5. Generar un banco de datos nacional de imágenes histológicas.



03. ACTIVIDADES REALIZADAS

3.1 CONSTRUCCIÓN DEL DATASET:

Se llevó a cabo la digitalización de muestras de PAPs provenientes del archivo del Servicio de Anatomía Patológica del Hospital B. Rivadavia. Antes del escaneo, las muestras fueron inspeccionadas y restauradas, solo en casos necesarios. El proceso de digitalización se realizó con un escáner Grundium OCUS 40 de la empresa *Biomakers*. En total, se digitalizaron muestras de PAPs de 197 pacientes. Para optimizar el procedimiento, sólo se escanearon las regiones de interés de cada muestra (parches), previamente identificadas por médicas patólogas.

La anotación de las muestras se realizó utilizando *Label Studio*, implementado en un servidor del ITBA. Inicialmente, la identificación de células patológicas y normales se efectuaba sobre parches completos, utilizando polígonos creados con múltiples clics del mouse. Aunque este método resultaba laborioso, permitió segmentar y etiquetar un total de 1690 células, conformando el denominado “Set 1”.

Posteriormente, se desarrolló un segundo protocolo de anotación más ágil e intuitivo, en el que solo era necesario un clic para seleccionar la categoría y otro para marcar el núcleo de la célula de interés. Para este proceso, se generaron mini-parches de tamaño uniforme a partir de los parches originales. Se seleccionaron 1000 mini-parches para la anotación. Esta mejora permitió un aumento significativo en la cantidad de células anotadas, alcanzando un total de 15019 células y 26158 anotaciones en total.

3.2 DESARROLLO DE LA PÁGINA WEB

(<https://beta-digitalpapsdb.exactas.uba.ar/>):

La versión final de la página web incluye seis vistas principales: *Home*, *About*, *Classification*, *Download*, *Terms & Conditions*, y *Contact Us*. La vista *Classification* permite explorar y filtrar mini-parches basados en su tipo de muestra, facilitando la navegación del usuario. Además, se han implementado funcionalidades para descargar tanto las imágenes como los archivos con anotaciones, asegurando el acceso y la reutilización del dataset.

Para garantizar la privacidad y organización de los datos en la página web RIVA, se reorganizaron los archivos originales. Dado que la galería de la página está dividida por imagen, se agruparon todas las anotaciones correspondientes a una imagen específica. Por otro lado, los nombres de los mini-parches se asignaron siguiendo un formato que



permite identificar cuáles mini-parches provienen de una misma muestra y su tipo (diagnóstico global), garantizando al mismo tiempo la integridad y anonimato de la información original.

3.3 EVALUACIÓN DEL CONSENSO ENTRE ANOTADORES

Para determinar la identidad de células individuales anotadas por más de una anotadora, se llevó a cabo una clusterización que agrupa las anotaciones por su proximidad en el espacio. De este modo, un cluster representa una célula. Cada una de estas células tiene anotada entre 1 y 4 clases, que pueden coincidir o no entre sí. Este enfoque permitió identificar células individuales anotadas por más de un experto, recuperando las coordenadas de sus núcleos y permitiendo evaluar el consenso inter-anotador. El análisis de consenso fue fundamental para entender el error humano y asignar una confianza a cada anotación. Este proceso valida la consistencia y precisión de las anotaciones realizadas.

3.4 APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE IMÁGENES:

Se optó por YOLOv8 como modelo principal debido a su facilidad de uso, eficiencia y alto rendimiento en visión computacional, entrenándolo sin modificaciones para mantener su velocidad y precisión en la inferencia. El modelo se entrenó con una base de datos pública (CRIC - *Center for Recognition and Inspection of Cells*) y con la base de datos generada en este proyecto (RIVA), evaluándose mediante métricas de rendimiento para modelos de detección de objetos. Se realizaron exhaustivos experimentos que permitieron evaluar el alcance del modelo en la detección y clasificación de células, obteniendo una *accuracy* similar al desempeño humano. No obstante, la capacidad del modelo para generalizar a distintos conjuntos de datos sigue siendo un reto, lo que resalta la importancia de contar con imágenes digitalizadas estandarizadas para optimizar el rendimiento y la robustez de los modelos.



04. RESULTADOS

El logro más destacado de este proyecto es la creación y publicación del **primer dataset** de imágenes anotadas de muestras de Papanicolaou obtenidas en un Hospital Público argentino (RIVA). Este dataset, es el más grande en cuanto a número de células anotadas según el sistema de Bethesda (15019 frente a 11534 de CRIC). La técnica de digitalización utilizada (scanner) y los algoritmos automáticos para la selección de mini-parches permitieron generar imágenes estandarizadas y más representativas de las muestras reales. Además, se ha explorado en forma exhaustiva el modelo YOLOv8 para la detección y clasificación de células.

El proyecto, aprobado por el comité de ética hospitalario, garantizó la anonimización de las muestras para proteger la privacidad de las pacientes, respetando así los principios de IA responsable. La disponibilidad pública del dataset y la página web bajo una licencia *Creative Commons* promueve la reutilización en otros proyectos, fomentando la colaboración y el avance científico. Este enfoque transparente y ético no solo facilita la replicación y validación de estudios, sino que también establece un estándar para futuros proyectos de IA en medicina.



CLIAS

CENTRO DE INTELIGENCIA
ARTIFICIAL Y SALUD
PARA AMÉRICA LATINA
Y EL CARIBE

