UNIVERSIDAD COMPLUTENSE DE MADRID

CLASIFICACIÓN DE IMÁGENES HISTOLÓGICAS DE CÁNCER DE MAMA UTILIZANDO UN MODELO DE REDES NEURONALES CONVOLUCIONALES

MASTER DATA SCIENCE , BIG DATA & BUSINESS ANALYTICS

PRESENTA
JUAN MANUEL RAMIREZ MELO

SEPTIEMBRE 2024

RESUMEN

Este proyecto tiene como objetivo desarrollar un modelo de Machine Learning para la clasificación de imágenes histopatológicas, con el propósito de detectar la presencia de carcinoma ductal invasivo (IDC) en tejidos mamarios. El carcinoma ductal invasivo es un tipo común de cáncer de mama, y la detección temprana es clave para mejorar las tasas de supervivencia. Utilizando un conjunto de datos de imágenes de histopatología descargado de Kaggle, que contiene imágenes etiquetadas como "IDC" (cáncer presente) y "No IDC" (sin cáncer), se procesaron las imágenes para su posterior análisis y modelado.

El proyecto aborda varios desafíos clave, como el desequilibrio de clases en el conjunto de datos, ya que hay más del doble de imágenes "No IDC" en comparación con "IDC". Las imágenes, uniformemente dimensionadas a 50x50 píxeles con tres canales de color, fueron analizadas en términos de sus propiedades estadísticas, como la intensidad de color. Esta información proporcionó una visión importante sobre las características de las imágenes, lo que permitirá ajustar el enfoque de modelado. La metodología se enfoca en el uso de redes neuronales convolucionales (CNNs) para realizar la clasificación automática de estas imágenes, buscando maximizar la precisión en la detección de IDC.

ÍNDICE

CONJUNTO DE DATOS

El conjunto de datos utilizado en este proyecto consiste en imágenes histopatológicas de tejido mamario, descargadas del dataset de Kaggle "Breast Histopathology Images". Este conjunto de datos es de acceso público y es comúnmente utilizado en investigaciones para la detección de carcinoma ductal invasivo (IDC).
El dataset contiene un total de 277,357 imágenes, cada una de las cuales ha sido categorizada en una de dos clases:
- Clase 0: No hay presencia de IDC (162,203 imágenes)
- Clase 1: Presencia de IDC (65,154 imágenes)

Cada imagen en el conjunto de datos es una imagen a color de 50x50 píxeles (RGB), lo que la hace pequeña en tamaño pero considerable en volumen, dado el número de muestras. Las imágenes están etiquetadas de manera uniforme, lo que permite realizar tareas de clasificación binaria. El conjunto de datos está desequilibrado, ya que hay una mayor cantidad de imágenes de "No IDC" en comparación con "IDC". Este desequilibrio es importante tenerlo en cuenta al entrenar un modelo de aprendizaje profundo, ya que puede introducir un sesgo hacia la clase mayoritaria.
Las imágenes están organizadas en directorios nombrados de acuerdo a su clasificación (0 para No IDC y 1 para IDC), y cada imagen corresponde a una pequeña porción de una muestra de tejido mamario más grande. Las imágenes fueron preprocesadas y aumentadas para garantizar que el modelo pudiera generalizar correctamente a datos no vistos, como se describe en capítulos posteriores.
Fuente del conjunto de datos: [Kaggle - Breast Histopathology Images Dataset](https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images)
Este conjunto de datos es fundamental para este proyecto, ya que ofrece un escenario real en el que el análisis automatizado de imágenes histopatológicas podría asistir a los profesionales médicos en la identificación del cáncer de mama de manera más eficiente.

METODOLOGÍA

Este proyecto involucra varios pasos que van desde el procesamiento de las imágenes originales hasta el entrenamiento del modelo. La metodología utilizada se puede dividir en dos partes principales:

i) Aumento de datos
ii) Entrenamiento del modelo

1 Aumento de datos

Para evitar el sobreajuste del modelo debido al bajo número de muestras comparado con otros problemas de clasificación de redes neuronales convolucionales, se aplicaron técnicas de aumento de datos. Esto incluye dividir las imágenes en parches y mejorar el conjunto de datos mediante espejado y rotación de los parches. Estas técnicas permiten que las imágenes de cáncer de mama se estudian desde diferentes orientaciones sin cambiar el diagnóstico, ayudando así a incrementar el tamaño del conjunto de datos. Los pasos de aumento de datos incluyen:

I) División en parches
II) Rotación
III) Espejado

Las 250 imágenes originales se convirtieron en 70,000 imágenes, manteniendo la misma etiqueta de clase que la imagen original.

2 Entrenamiento del modelo

Para el entrenamiento se utilizó una red neuronal convolucional (CNN). En concreto, se empleó la arquitectura DenseNet-121. Las imágenes, originalmente de 512x512 píxeles, se redimensionan a 224x224 píxeles antes de pasarlas por la primera capa convolucional, que reduce el tamaño a 112x112 píxeles. Posteriormente, las imágenes pasan por una capa de agrupamiento (pooling), que las reduce a 56x56 píxeles.

Esta arquitectura cuenta con cuatro bloques densos y tres capas de transición. Para mejorar la compacidad del modelo, se reduce el número de mapas de características (feature maps) en cada capa de transición. Los bloques densos aplican filtros de diferentes tamaños a cada imagen, reduciendo el tamaño de salida a la mitad en cada paso. El modelo final se optimiza para la clasificación de las imágenes aumentadas.

REDIMENSIONAMIENTO CON MATLAB

Para facilitar el procesamiento y adaptarse a las limitaciones de memoria RAM en el entorno de Google Colab, se utilizó MATLAB para redimensionar las imágenes del conjunto de datos original. Inicialmente, las imágenes se procesaron a mayor resolución, pero debido a que la RAM disponible no era suficiente para manejar este tamaño en Colab, se decidió aumentarlas a 224 x 224 píxeles.

Este redimensionamiento fue fundamental para permitir que el modelo se pudiera entrenar sin problemas de memoria. MATLAB fue la herramienta elegida para realizar este ajuste, ya que garantizó una gestión eficiente de los recursos y conservó la calidad de las imágenes. Aunque las imágenes fueron reducidas en tamaño, se mantuvieron las características histológicas clave necesarias para la clasificación. De este modo, se optimiza tanto el procesamiento de datos como el rendimiento del modelo de red neuronal convolucional.

AUMENTO DE DATOS (DATA ARGUMENTATION)

Antes de aplicar las técnicas de aumento de datos, el conjunto de datos presentaba un desequilibrio significativo entre las clases. La distribución de clases era la siguiente:

- Clase 0 (No IDC): 162,203 imágenes
- Clase 1 (IDC): 65,154 imágenes

Para mejorar la capacidad de generalización del modelo y evitar el sobreajuste hacia la clase mayoritaria, se aplicó la técnica (Data Augmentation). Esta técnica permite generar nuevas imágenes a partir de las ya existentes, aplicando transformaciones como rotaciones, espejado y escalado. Esto ayuda a aumentar el número de muestras en la clase minoritaria y, en última instancia, a equilibrar la distribución de clases en el conjunto de datos.

El proceso de aumento de datos incluyó los siguientes pasos:
1. Filtrar y seleccionar las imágenes de la clase minoritaria con la forma correcta (224x224x3).
2. Generar imágenes aumentadas aplicando rotaciones y espejado.
3. Combinar las imágenes originales con las aumentadas para crear un conjunto de datos más equilibrado.

Después de aplicar (Data Augmentation), la distribución de clases fue la siguiente:

- Clase 1 (IDC): 321,804 imágenes
- Clase 0 (No IDC): 53,634 imágenes

Gracias a este proceso, se consiguió un conjunto de datos más equilibrado, lo que facilita el entrenamiento del modelo sin sesgos hacia la clase mayoritaria. Finalmente, para optimizar el tiempo de procesamiento, se guardó el conjunto de datos equilibrado en un archivo (.pkl), permitiendo una carga más rápida en los experimentos posteriores.

MODELO DE RED NEURONAL CONVOLUCIONAL (CNN)

En este proyecto, se ha seleccionado un modelo de red neuronal convolucional (CNN) debido a su eficacia en la clasificación de imágenes. Las CNN están diseñadas para reconocer patrones espaciales y características en los datos visuales, lo que las hace ideales para tareas de detección de objetos o, en este caso, para la detección de cáncer en imágenes histopatológicas.

- Preparación del Conjunto de Datos

Se realizó un muestreo aleatorio para seleccionar 20,000 muestras balanceadas de cada clase, correspondientes a imágenes con y sin presencia de carcinoma ductal invasivo (IDC). Estas imágenes fueron convertidas en arrays de NumPy y divididas en conjuntos de entrenamiento y prueba con una proporción de 80% para entrenamiento y 20% para prueba.

- Construcción del Modelo CNN

El modelo CNN se construyó utilizando una arquitectura secuencial con varias capas convolucionales y de agrupamiento. La configuración específica fue la siguiente:
- Capas Convolucionales: Se aplicaron capas convolucionales con filtros de diferentes tamaños (32, 64 y 128) y funciones de activación `relu` para extraer características clave de las imágenes.
- Pooling: Después de cada capa convolucional, se aplicaron capas de pooling (submuestreo) para reducir las dimensiones de las imágenes, lo que ayuda a mejorar la eficiencia del modelo.
- Capa Densa: Finalmente, se aplanaron las características extraídas y se pasó por una capa densa (fully connected) con una función de activación `sigmoid` para la clasificación binaria (IDC o No IDC).

El modelo fue compilado usando el optimizador **Adam** y la función de pérdida binary crossentropy, con precisión como métrica principal.

- Entrenamiento del Modelo

El modelo fue entrenado en un total de 20 épocas, con una tasa de abandono (dropout) del 0.8% para prevenir el sobreajuste. Durante el proceso de entrenamiento, se alcanzó una precisión de entrenamiento del 87.59% y una precisión de validación del 84.91%, mostrando una mejora constante en cada época. El uso de técnicas como el escalado y la rotación de imágenes ayudó a aumentar la robustez del modelo.

- Evaluación del Modelo

Una vez entrenado, se evaluó el desempeño del modelo utilizando una matriz de confusión, lo que permitió analizar el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). El modelo alcanzó los siguientes resultados:
- TP (Verdaderos Positivos): 2830 imágenes correctamente clasificadas como IDC.
- TN (Verdaderos Negativos): 3933 imágenes correctamente clasificadas como No IDC.
- FP (Falsos Positivos): 53 imágenes incorrectamente clasificadas como IDC.
- FN (Falsos Negativos): 1184 imágenes incorrectamente clasificadas como No IDC.

La precisión del modelo fue calculada en aproximadamente 84.54%, lo que refleja una alta tasa de acierto en la clasificación de imágenes de cáncer de mama.

ENTRENAMIENTO DEL MODELO

Para el entrenamiento del modelo se utilizó un generador de datos (`ImageData Generator`) con el fin de aplicar transformaciones en tiempo real durante el proceso de entrenamiento, lo que ayuda a mejorar la robustez del modelo y a prevenir el sobreajuste. Las transformaciones aplicadas incluyeron:

- Rescale: Reescalar las imágenes (1/255) para normalizarlas.
- Rotación: Aplicar rotaciones aleatorias de hasta 20 grados.
- Zoom: Un rango de zoom de 15%.
- Desplazamientos: Desplazamientos en el ancho y la altura de hasta 20%.
- Cizallamiento: Aplicar una deformación con un rango de cizallamiento de 15%.
- Espejado: Reflejar las imágenes horizontalmente.

El modelo fue entrenado durante 20 épocas utilizando un tamaño de lote de 32 imágenes. Durante el entrenamiento, se monitoriza la precisión tanto en el conjunto de entrenamiento como en el de validación, así como la función de pérdida (loss).

- Resultados del Entrenamiento

A lo largo de las 20 épocas de entrenamiento, se observó un incremento continuo en la precisión del modelo, alcanzando una **precisión final de entrenamiento del 88.47%** y una **precisión en validación de 84.54%**. Los valores de pérdida (loss) también se redujeron significativamente, reflejando una mejora en la capacidad del modelo para generalizar sobre el conjunto de validación.

Los resultados del entrenamiento por época fueron los siguientes:

- Epoch 1/20: Precisión = 80.21%, Precisión de validación = 81.36%, Pérdida = 0.4583
- Epoch 10/20: Precisión = 86.74%, Precisión de validación = 88.88%, Pérdida = 0.3285
- Epoch 20/20: Precisión = 88.47%, Precisión de validación = 84.54%, Pérdida = 0.2814

Este aumento de precisión indica que el modelo fue mejorando su capacidad para identificar correctamente las imágenes con y sin carcinoma ductal invasivo (IDC) a lo largo del tiempo. Además, la reducción en la pérdida muestra que el modelo fue ajustando sus predicciones de manera más precisa.

CÁLCULO DE LA PRECISIÓN DEL MODELO

Este cálculo muestra cómo se obtuvo la precisión del modelo utilizando los valores de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).

Cálculo de la Precisión

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

- **TP (True Positives)** = 2830
- **TN (True Negatives)** = 3933
- **FP (False Positives)** = 53
- **FN (False Negatives)** = 1184

Sustituyendo los valores en la fórmula:

$$\text{Precisión} = \frac{2830 + 3933}{2830 + 3933 + 53 + 1184} = \frac{6763}{8000} = 0.845375$$

Convertido a porcentaje, la precisión es aproximadamente **84.54%**.

Sustituyendo estos valores en la fórmula, se obtiene una precisión de aproximadamente 84.54%.

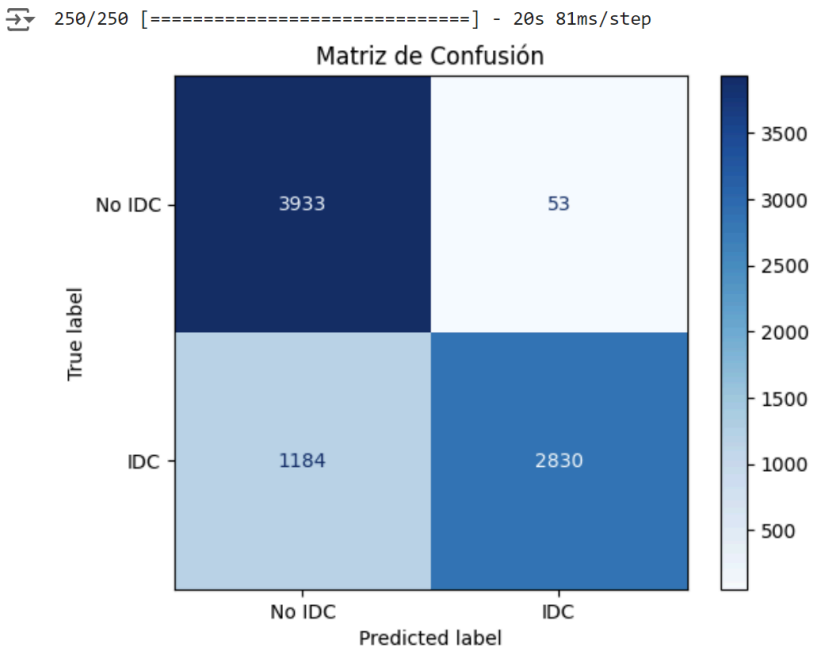
RESULTADOS

La matriz de confusión muestra el desempeño del modelo al clasificar imágenes con y sin carcinoma ductal invasivo (IDC). En total, se probaron 8,000 imágenes, de las cuales 3,933 fueron correctamente clasificadas como "No IDC" (Verdaderos Negativos), y 2,830 fueron correctamente clasificadas como "IDC" (Verdaderos Positivos).

Sin embargo, también se presentaron errores de clasificación:
- Falsos Positivos (FP): 53 imágenes fueron clasificadas incorrectamente como "IDC" cuando no tenían cáncer.

- Falsos Negativos (FN): 1,184 imágenes que sí tenían cáncer fueron clasificadas incorrectamente como "No IDC".

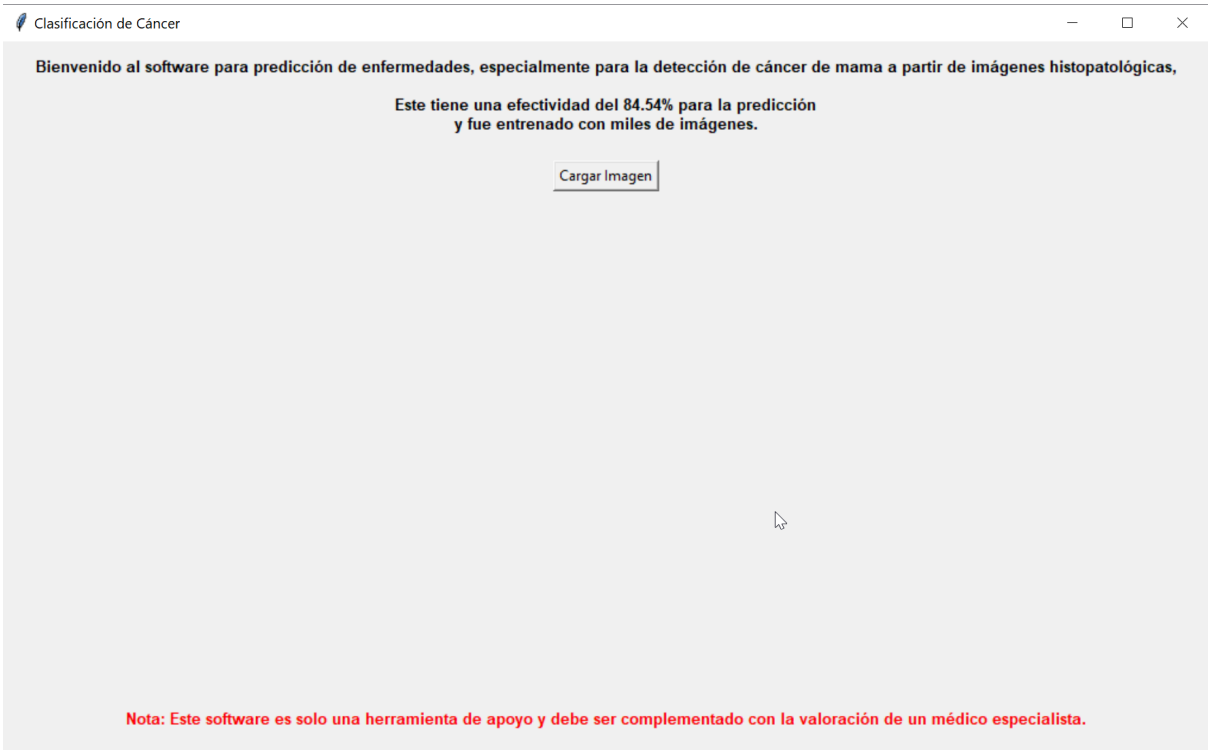
Estos resultados llevaron a una precisión general de 84.54%, lo que indica que el modelo tiene un buen desempeño en la clasificación de imágenes histopatológicas, aunque sigue habiendo margen para mejorar en la reducción de falsos negativos.



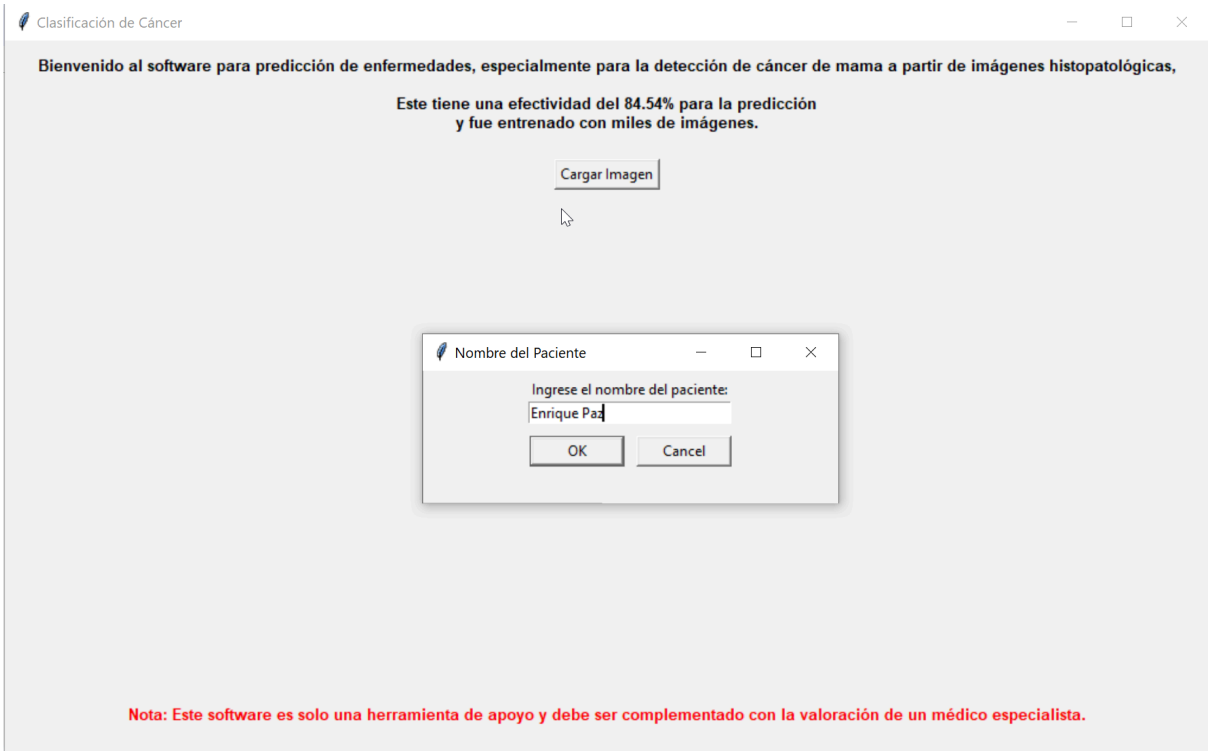
APLICACIÓN DE ESCRITORIO (EXTENSIÓN .EXE)

La aplicación de escritorio fue desarrollada para facilitar la predicción de la presencia de cáncer de mama utilizando el modelo previamente entrenado y guardado en formato `.h5`. El flujo de la aplicación es el siguiente:

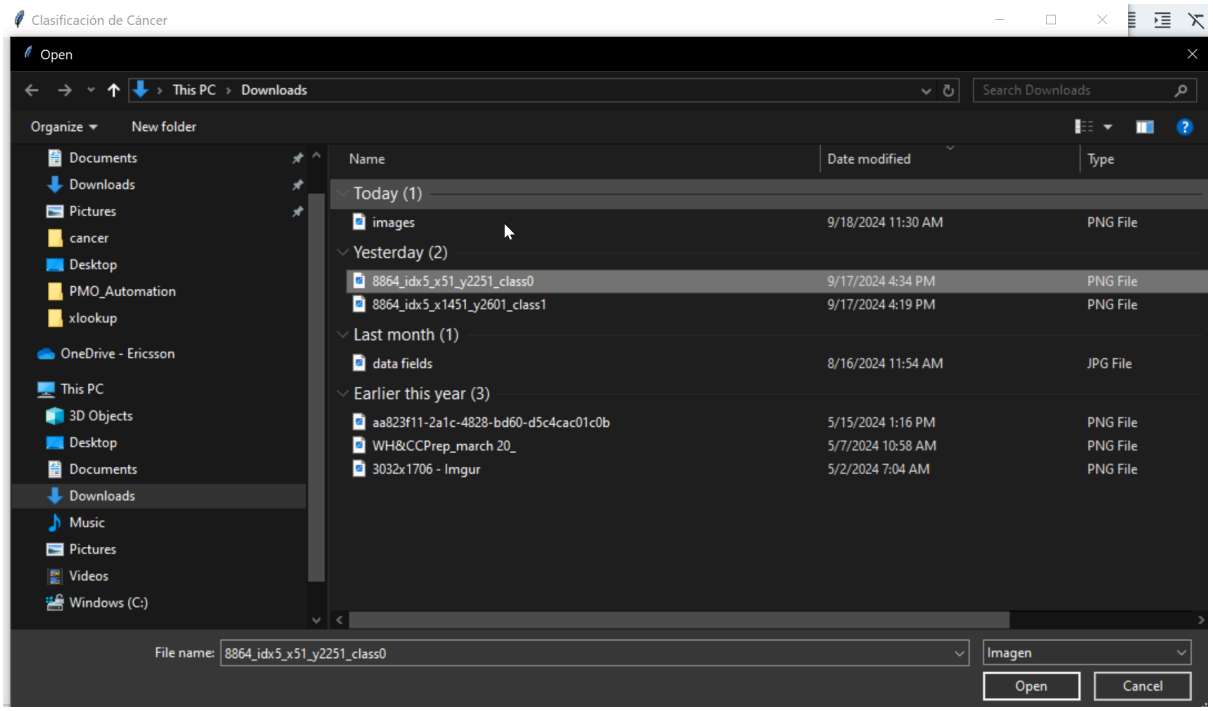
1. **El usuario abre la aplicación: La interfaz de la aplicación da la bienvenida al usuario y explica brevemente la finalidad de la herramienta, resaltando que tiene una precisión del 84.54%.**



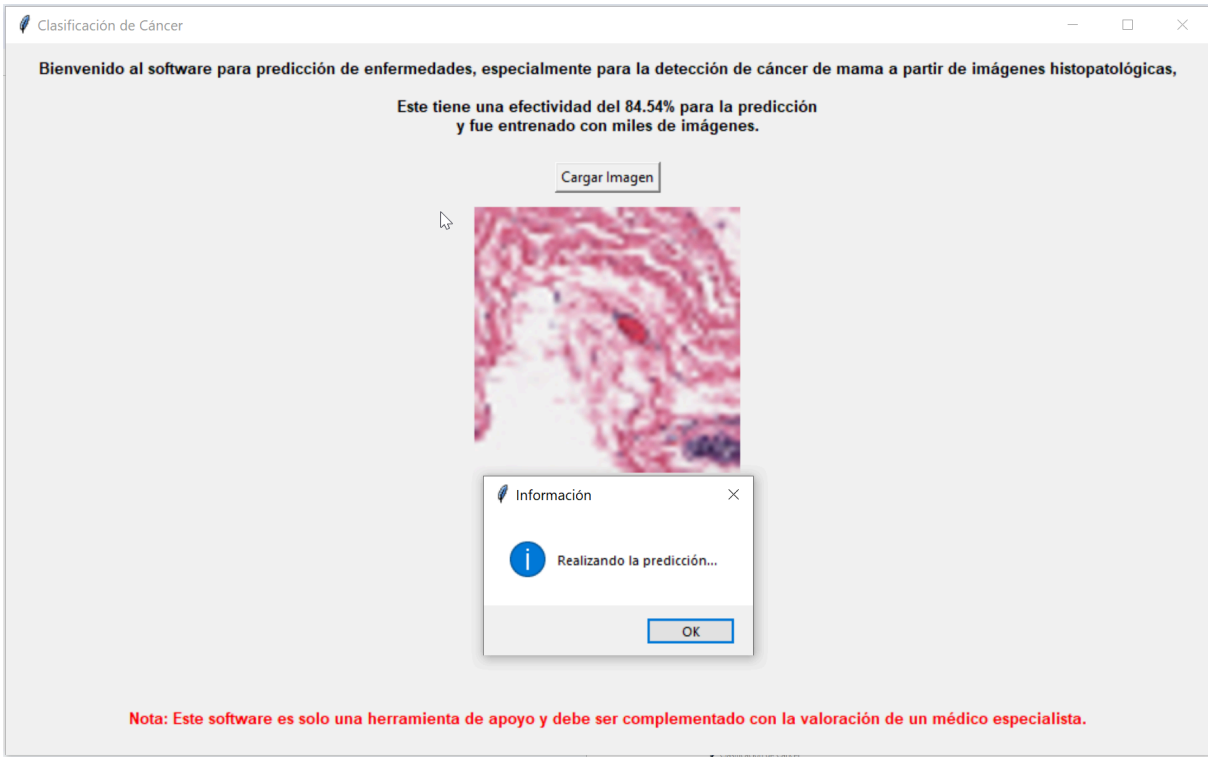
2. **Ingreso del nombre del paciente: El usuario debe ingresar el nombre del paciente que será analizado antes de cargar la imagen para el diagnóstico.**

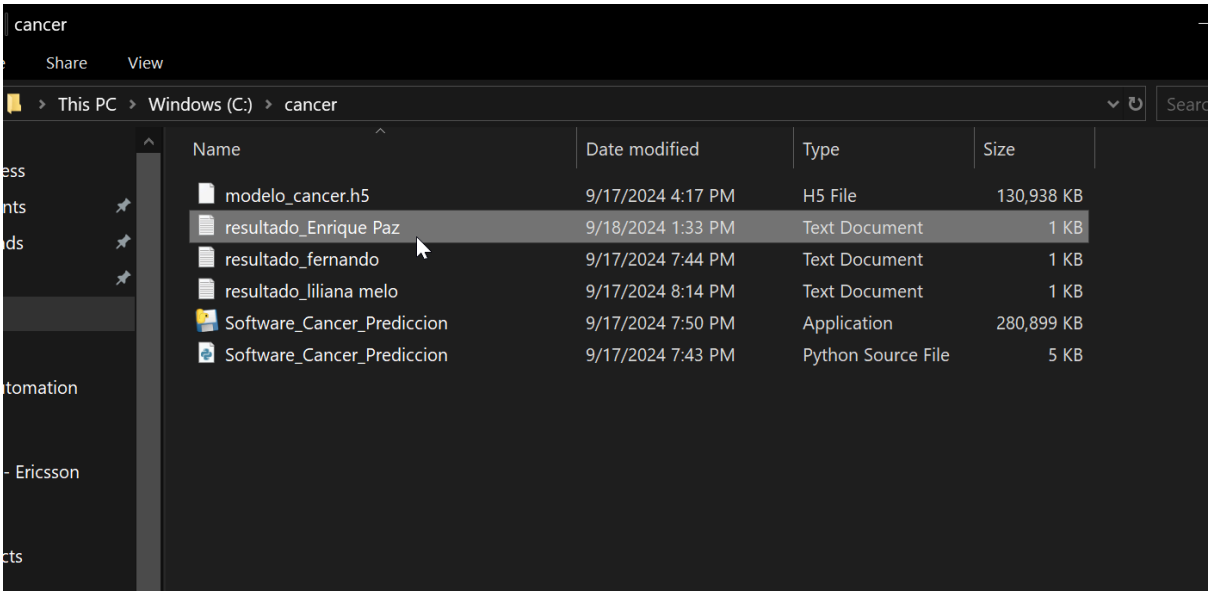


3. **Análisis de la imagen: El usuario selecciona una imagen histopatológica desde su sistema. La aplicación carga la imagen y procede a realizar el análisis mediante el modelo entrenado.**



4. **Resultado en texto: Una vez finalizado el análisis, la aplicación guarda el resultado en un archivo de texto en el sistema, indicando si se ha detectado o no carcinoma ductal invasivo (IDC) en la imagen. El archivo tiene como nombre el del paciente para facilitar su identificación.**





CONCLUSIÓN

Este proyecto ha demostrado la viabilidad de utilizar redes neuronales convolucionales (CNNs) para la detección automatizada de carcinoma ductal invasivo (IDC) en imágenes histopatológicas de cáncer de mama. A lo largo del desarrollo, se han aplicado diversas técnicas y metodologías para mejorar la precisión del modelo y superar los desafíos inherentes al desequilibrio de clases y el manejo de grandes volúmenes de datos.

Uno de los primeros pasos clave fue la preparación del conjunto de datos, que incluía tanto imágenes con presencia de IDC como sin ella. Debido al marcado desequilibrio entre ambas clases, se aplicaron técnicas de **Data Augmentation** para generar nuevas imágenes y equilibrar el conjunto de datos. Esto permitió entrenar un modelo con mayor capacidad de generalización y evitar sesgos hacia la clase mayoritaria.

El modelo se construyó utilizando la arquitectura DenseNet-121, seleccionada por su capacidad para extraer características detalladas de las imágenes. Se realizaron varias optimizaciones, incluyendo el uso de **Dropout** y **Batch Normalization** para prevenir el sobreajuste. Durante el entrenamiento, el modelo mostró mejoras continuas en precisión, alcanzando una precisión final de **84.54%** en el conjunto de validación.

La evaluación del modelo se realizó mediante una matriz de confusión, que permitió analizar tanto los aciertos como los errores de clasificación. Los resultados indicaron que, si bien el modelo tiene un buen rendimiento general, es necesario seguir trabajando en la reducción de los falsos negativos, dado que estas clasificaciones incorrectas tienen un impacto significativo en el diagnóstico de enfermedades críticas como el cáncer.

Finalmente, el modelo se encapsuló en una aplicación de escritorio en formato `.exe` para facilitar su uso por parte de médicos y profesionales de la salud. La aplicación permite cargar imágenes, realizar el análisis de manera automática y generar un resultado en texto con el diagnóstico. Esta herramienta tiene el potencial de ser utilizada como un sistema de apoyo en la detección de cáncer, aunque debe ser complementada con la evaluación de un especialista.

En conclusión, este proyecto no solo ha demostrado la capacidad de las CNNs para resolver problemas complejos en el ámbito médico, sino que también ha creado una solución práctica para apoyar a los profesionales de la salud en la detección temprana de cáncer de mama. Sin embargo, futuras mejoras deben centrarse en optimizar aún más la precisión del modelo y reducir el tiempo de procesamiento en entornos de baja capacidad de memoria, como Google Colab.

ANEXO

Link del Dataset Original
https://drive.google.com/drive/folders/1LYHrc5_IHeg6uT70PFIdHhIm4UGroc00?usp=sharing

Link del Dataset con las imágenes redimensionadas 224
https://drive.google.com/drive/folders/1U5OxqbwfsW4eTr7aZW1eFqAmHGmStLUd?usp=sharing

Link del Data frame con las imágenes 224px
https://drive.google.com/file/d/1Df-KXieRh2FbvabregRHl3b94Uf3sWrO/view?usp=sharing

Link al NoteBook de Google Colab
https://colab.research.google.com/drive/12wDiu1bhkRGaIlsW85iNBVe_-3eLX6MC?usp=sharing

Link del Modelo Exportado en extensión h5
https://drive.google.com/file/d/1LlPPSGtAhvy2-ht0HNIYcodqc42cHie1/view?usp=sharing

Link del Codigo para la app de Escritorio .EXE
https://drive.google.com/file/d/1LlPPSGtAhvy2-ht0HNIYcodqc42cHie1/view?usp=sharing

Link al video del Proyecto
https://drive.google.com/file/d/1RNgvwDfhKXJUO8ZN6ugVhLVsFOjWqO4X/view?usp=sharing

Link a la Presentación del Proyecto
https://docs.google.com/presentation/d/1C4aukotpfiLWvqs3zqq0xHlAsy53wo1rt7_2IxCOqb4/edit?usp=sharing