****

**Deep learning classification of breast, lung and colon cancer using histological measures**

**(Prototipo)**

**Autores:**

Martín Antúnez García

Luis Acero Morata

Carlos Arcas Perez

Javier Harto Vallejo

**Supervisor:**

Teno González Dos Santos

**Máster en Inteligencia Artificial y Data Science**

**ÍNDICE**

[**1. Descripción del prototipo 4**](#_6h8zrwvu5goh)

[1.1. Objetivos 4](#_28l70vwgft5n)

[1.2. Alcance 4](#_b2yly95jk4f3)

[1.3. Metodología 4](#_q5ucko42zpw8)

[**2. Arquitectura del prototipo 5**](#_2kqh0svhzn9z)

[**3. Conjunto de datos 6**](#_54u5svpvo193)

[3.1. Fuente de los datos 6](#_adoztmy1zi1f)

[3.2. Tamaño y características del conjunto de datos 6](#_5znug136ebmx)

[3.3. Preprocesamiento de los datos 7](#_8znzm0rdvmes)

[**4. Funcionalidades del prototipo 8**](#_y5obdsw39dy4)

[**5. Evaluaciones 9**](#_wesvuljtunp6)

[5.1. Métricas de rendimiento utilizadas 9](#_ov7y8o4co4jb)

[5.2. Resultados obtenidos en la evaluación 9](#_fs1woliqp2yn)

[**6. Casos de uso y ejemplos 10**](#_18c7w332kqyf)

[6.1. Limitaciones del prototipo actual 10](#_tebsylkt65jo)

[6.2. Uso del prototipo en la detección de células tumorales 10](#_h6kbda5j1zdn)

[**7. Bibliografía 11**](#_ftdd3svpl1wg)

# Descripción del prototipo

## 1.1. Objetivos

Desarrollar un modelo de clasificación de imágenes histológicas en células de pulmón, colon y pecho, así como, la distinción entre células cancerígenas y no cancerígenas. Además, para el caso del pulmón, también se diferenciarán las células cancerígenas entre carcinomas y adenocarcinomas.

## 1.2. Alcance

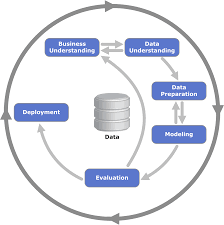
El alcance del prototipo se centra en desarrollar y evaluar cuatro modelos de inteligencia artificial para la predicción de células tumorales. El objetivo principal es demostrar la viabilidad y efectividad de los modelos en la clasificación de células tumorales en diferentes partes del cuerpo, específicamente en el pulmón, el colon y el pecho.

El prototipo se enfocará en las siguientes funcionalidades clave:

* Preprocesamiento de imágenes
* Modelo de clasificación de partes del cuerpo
* Modelos de detección de células cancerígenas
* Evaluación del rendimiento

## 1.3. Metodología

La metodología utilizada para el desarrollo de este prototipo se basa en el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM es un enfoque estructurado y ampliamente reconocido para proyectos de minería de datos e inteligencia artificial, que consta de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.



# Arquitectura del prototipo

El prototipo principalmente se compone de cuatro modelos de redes neuronales:

* body\_part\_model.h5
* breast\_model.h5
* colon\_model.h5
* lung\_model.h5

Todos los modelos han sido generados a partir de redes convolucionales, utilizando la librería TensorFlow. Su creación no requirió de una compleja red neuronal, ya que con una más simple se logró obtener resultados casí perfectos con un menor costo y tiempo computacional.

El prototipo se divide en dos partes principales. En la primera parte, se utiliza el modelo neuronal "body parts model.h5" para realizar la predicción de la parte del cuerpo a la que pertenece una imagen de célula (colon, pecho o pulmón). El objetivo de esta parte del prototipo es clasificar correctamente las imágenes de células en función de la parte del cuerpo a la que corresponden. Para lograr esto, se ha entrenado el modelo "body\_part\_model.h5" utilizando conjuntos de datos etiquetados previamente.

La segunda parte del prototipo consiste en utilizar modelos adicionales específicos para cada parte del cuerpo (pulmón, cólon y pecho) con el fin de distinguir si una célula es cancerígena o no. Estos modelos se aplican una vez que se ha determinado la parte del cuerpo a la que pertenece la célula utilizando el primer modelo "body\_part\_model.h5".

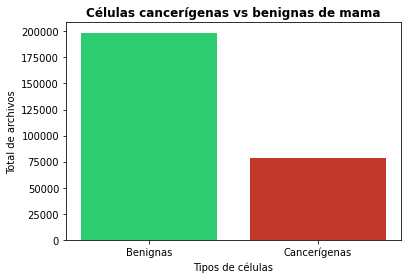
# Conjunto de datos

## 3.1. Fuente de los datos

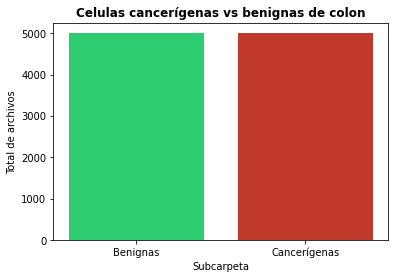
Los datos se han obtenido de dos fuentes Academic Torrents [1] y Kaggle [2]. De la primera se extrajeron los datos referentes a la células de pecho y de la segunda los de colon y pulmón

## 3.2. Tamaño y características del conjunto de datos

El primer dataset utilizado en el proyecto consta de 277.524 imágenes de células de dimensiones 50x50. De estas imágenes, 198.738 corresponden a células benignas y 78.786 a células cancerígenas. El dataset está organizado en carpetas que llevan el nombre de los identificadores de cada paciente. A su vez, cada carpeta se subdivide en dos subcarpetas, una para las imágenes de células benignas y otra para las imágenes de células cancerígenas.



El segundo de los dataset,está compuesto por dos conjunto de datos, colon y pulmones, contiene un total de 25.000 archivos, que se dividen en 10.000 imágenes de colon (5000 benignas y 5000 carcinoma) y 15.000 de pulmón (5000 benignas, 5000 carcinoma y 5000 adenocarcinoma). Las imágenes de este dataset tienen una escala de 768x768 .



## 3.3. Preprocesamiento de los datos

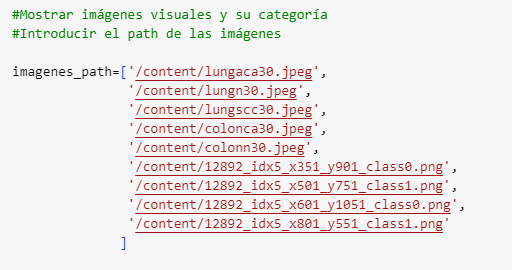
El primer paso en el procesamiento de datos consistió en balancear las clases, para ello, se redujo el número de imágenes de pecho a solo 10.000, seleccionadas de manera aleatoria y uniformemente, de tal manera que hubiese la misma cantidad de ejemplos con células benignas y cancerígenas.

Una vez realizado el balanceamiento de clases, se redimensionó las imágenes de colon y pulmón a 50x50, garantizando que todas las imágenes tuvieran el mismo tamaño y reduciendo considerablemente el tiempo de cómputo necesario.

Por último, para lograr mayor precisión a la hora de hacer el modelo, se optó por hacer un reescalado, dejando los píxeles de las imágenes en un valor comprendido entre [0’5, -0’5].

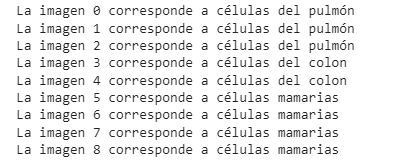
# Funcionalidades del prototipo

El prototipo se puede probar accediendo al archivo “ModeloFinalDetección\_(prototipo).ipynb”. Una vez que se esté en el archivo simplemente bastará con introducir en el código las direcciones a las imágenes de las células.

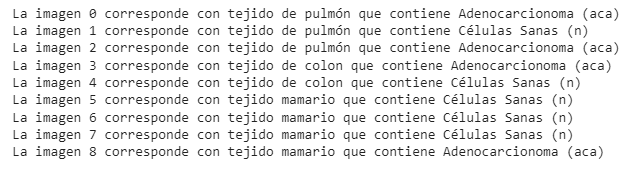


A continuación, bastaría con ejecutar el resto del código para visualizar los resultados. No haría falta ningún preprocesado manual de la imágen, ya que el propio código se encargará de reescalarla y redimensionarla.

El primer resultado que se mostrará, será la parte del cuerpo a la que corresponde cada tipo de célula.



En el segundo se visualizará tanto la parte del cuerpo involucrada como el tipo de célula que se visualiza en la imagen.



# Evaluaciones

## 5.1. Métricas de rendimiento utilizadas

A continuación se explicarán detalladamente algunas de las métricas que se usarán para evaluar el rendimiento de cada uno de los 4 modelos.

* **Matriz de Confusión**: La matriz de confusión es una tabla que muestra la precisión de un modelo de clasificación. Es una representación visual que muestra la cantidad de predicciones correctas e incorrectas realizadas por el modelo para cada clase de interés. La matriz se compone de cuatro elementos: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).
* **Accuracy (Exactitud)**: La exactitud es una métrica que mide la proporción de predicciones correctas realizadas por un modelo en comparación con el total de predicciones. Se calcula dividiendo la suma de los verdaderos positivos y los verdaderos negativos entre el total de predicciones realizadas.
* **Precision (Precisión)**: La precisión es una métrica que cuantifica la proporción de predicciones positivas correctas realizadas por un modelo en relación con el total de predicciones positivas. Se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos positivos.
* **Recall (Recuperación o Sensibilidad)**: El recall es una métrica que mide la proporción de instancias positivas correctamente identificadas por un modelo en relación con el total de instancias positivas. Se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos.
* **F1-Score**: El F1-Score es una métrica que combina la precisión y el recall en un solo valor. Proporciona una medida equilibrada entre ambas métricas y es útil cuando las clases están desequilibradas. Se calcula como la media armónica de la precisión y el recall, y se utiliza como una medida general del rendimiento del modelo.

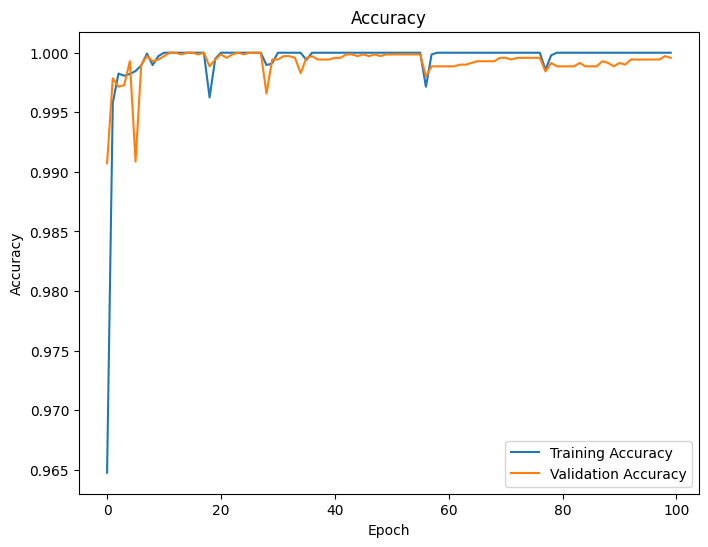
## 5.2. Resultados obtenidos en la evaluación

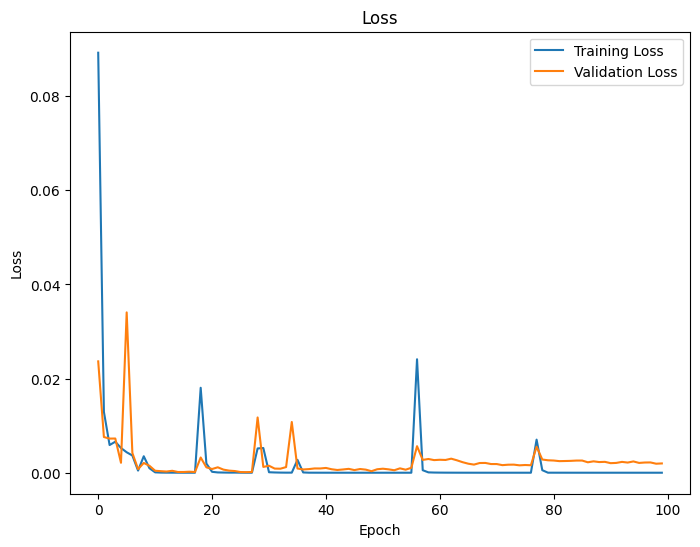
* **Métricas obtenidas**

| **Modelos** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| body\_part\_model.h5 | **1.0** | **1.0** | **1.0** | **1.0** |
| breast\_model.h5 | **0.8355** | **0.7867** | **0.8127** | **0.8460** |
| colon\_model.h5 | **0.9585** | **0.9849** | **0.9301** | **0.9567** |
| lung\_model.h5 | **0.9950** | **0.9959** | **0.9939** | **0.9949** |

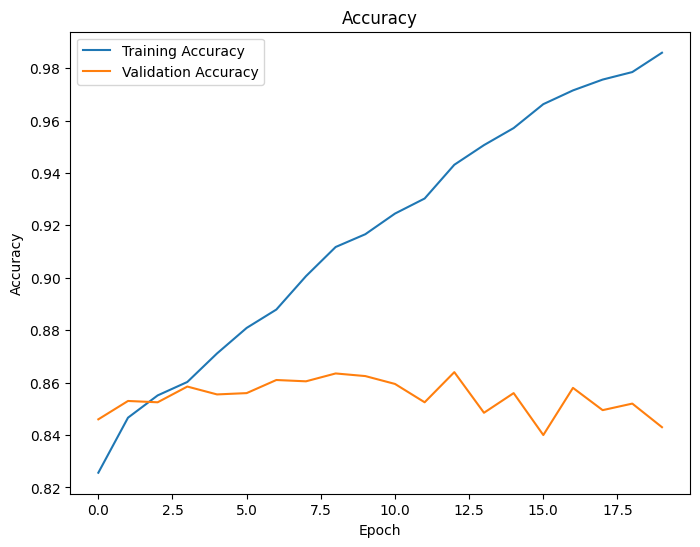
* **Gráficas de la evolución del accuracy y loss por ciclo**

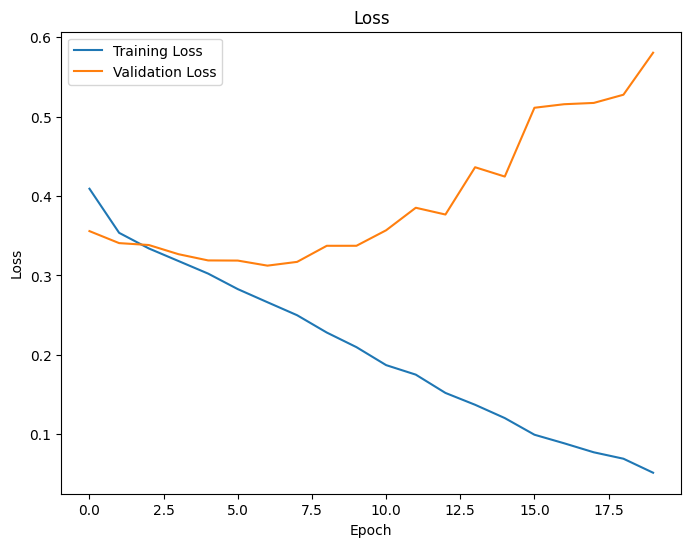
1. Modelo división por partes



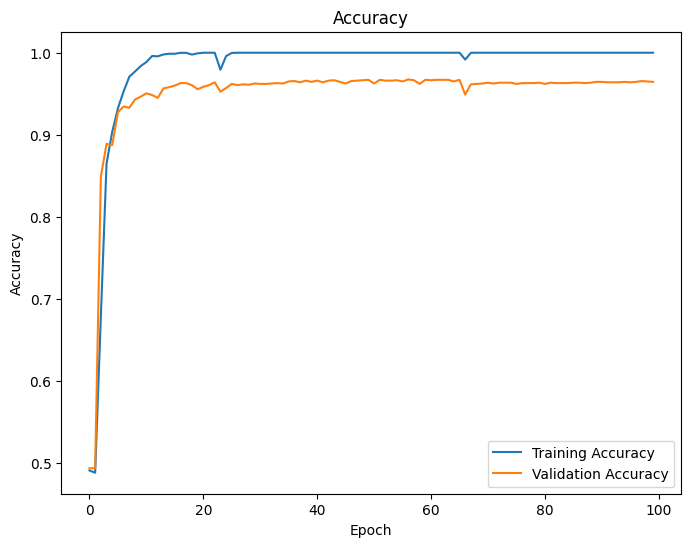


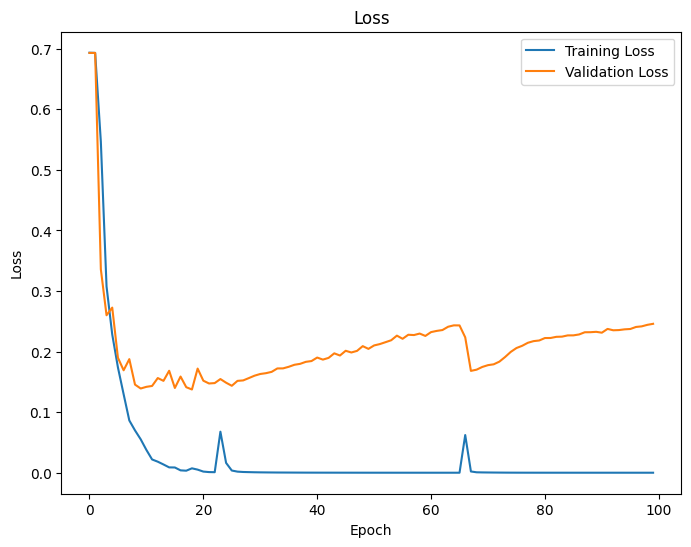
1. Modelo células mamarias



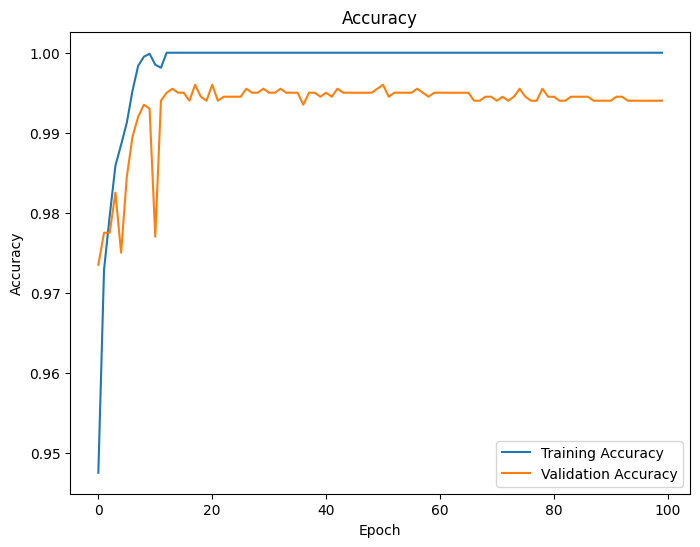


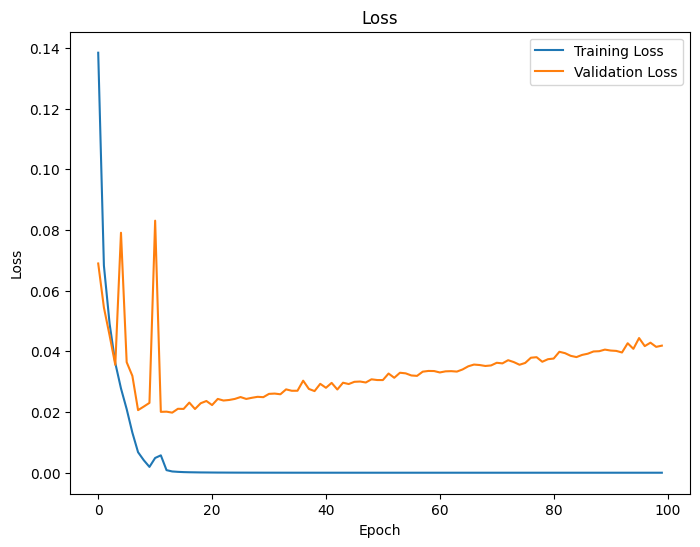
1. Modelo células colon





1. Modelo células pulmón





## 5.3. Futuras mejoras para el proyecto final

El proyecto final se enfocará en implementar diversas mejoras para abordar dos aspectos clave: reducir el nivel de sobreentrenamiento de los modelos y mejorar el resultado en términos de accuracy, este último más enfocado en mejorar el modelo de células mamarias, que actualmente presentan los resultados más bajos. Para ello, se buscará reformar los modelos, reduciendo el número de epochs, si hiciera falta, y el uso de técnicas de data augmentation.

# **Casos de uso y ejemplos**

## 6.1. Limitaciones del prototipo actual

Entendido, una de las limitaciones del prototipo actual es que al redimensionar imágenes de grandes dimensiones, existe la posibilidad de que se pierdan algunas características clave de la imagen. Este problema surge debido a la reducción de resolución y detalles durante el proceso de redimensionado.Una opción para solventar este problema sería trabajar directamente con imágenes de mayor tamaño sin necesidad de redimensionarlas. Sin embargo, esto requeriría una mayor capacidad computacional y recursos, ya que las imágenes más grandes ocupan más espacio de almacenamiento y requieren más potencia de procesamiento.

Otra de sus limitaciones, podría ser el número de ejemplos usados, ya que a pesar de que para las imágenes de células mamarias se disponían de más de 250.000 ejemplos, se tuvo que limitar para que estuviesen nivelados con los datos de colon y pulmón. Una posible solución a este dilema,sería implementar técnicas de “data augmentation” para conseguir un dataset más diversos, sin requerir más imágenes.

## 6.2. Uso del prototipo en la detección de células tumorales

En el video adjunto a esta memoria se puede observar un ejemplo de caso de uso para la detección de células tumorales de las tres zonas del cuerpo (colon, pulmón y pecho).

## 

## 

# Bibliografía

[1] None. (n.d.). *Invasive ductal carcinoma (IDC) histology image dataset*. Academic Torrents.

<https://academictorrents.com/details/e40bd59ab08861329ce3c418be191651f35e2ffa>

[2] Larxel, “Lung and Colon Cancer Histopathological Images.” 14-Apr-2020.

<https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images>