****

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

**(Memoria)**

**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**

**Indice:**

[**1. Introducción 2**](#_6h8zrwvu5goh)

[1.1 Antecedentes 2](#_1xskxjt3wwnp)

[1.2. Objetivos 2](#_28l70vwgft5n)

[1.3. Alcance 3](#_b2yly95jk4f3)

[1.4. Metodología 3](#_q5ucko42zpw8)

[**2. State of art 6**](#_sgymkvgca3dm)

[**3. Arquitectura del proyecto 9**](#_owy6akb93wfq)

[**4. DataSets 10**](#_n15imt90r8wl)

[4.1. Fuente de los datos 10](#_adoztmy1zi1f)

[4.2. Tamaño y características del conjunto de datos 10](#_5znug136ebmx)

[4.3. Preprocesamiento de los datos 11](#_8znzm0rdvmes)

[**5. Funcionalidades del proyecto 12**](#_y5obdsw39dy4)

[**6. Evaluaciones 13**](#_wesvuljtunp6)

[6.1. Métricas de rendimiento utilizadas 13](#_ov7y8o4co4jb)

[6.1.1. Resultados preliminares y primeros obstáculos 14](#_xemioduyrarv)

[6.2. Resultados obtenidos en la evaluación 17](#_fs1woliqp2yn)

[**7. Conclusiones 28**](#_dr9s0e4od)

[**8. Trabajos futuros 31**](#_hewe3xozk8w0)

[**9. Anexos 33**](#_h4qmuumxcn9)

[**Bibliografía 36**](#_se0enq4stn60)

# Introducción

## 1.1 Antecedentes

Recientemente ha habido un importante auge del uso de CNN en la detección de cáncer, lo cual ha tenido una notable contribución a la reducción de la mortalidad. Sin ir más lejos, son varios los estudios publicados en la revista Nature que proporcionan varios modelos con un nivel de precisión más que aceptable. Por ejemplo, Zhoul et al (2022) proporcionaron un modelo llamado HCCANet para la detección del cáncer colorrectal superando muchos otros modelos de machine learning clásicos. Por otro lado, Hameed et al (2022), reportaron una precisión del 98% en detección de cáncer de mama, siendo capaces de detectar si el tejido era normal, una lesión benigna, in situ carcinoma o carcinoma invasivo, proporcionando una válida herramienta para el diagnóstico temprano en cáncer de mama. Por otro lado, Chaunzwa et al., (2021), usó tejidos para la detección de cáncer de pulmón entrenados mediante CNN , pudiendo detectar la presencia tumoral, aunque algunos clasificadores de machine learning como k-nearest neighbors y support vector machine mostraron resultados similares.

Actualmente, aunque hay varios modelos orientados a la detección de presencia tumoral en imágenes de un tejido específico, aún es preciso el desarrollo de una herramienta que implemente una mayor flexibilidad en su labor de detección. Por ello, crearemos una herramienta de clasificación dividida en varios pasos. Primero, el clasificador será capaz de indicar si el tejido histológico pertenece a una muestra de pulmón, colon o mama, seguido de un clasificador que indicará la presencia o no de tumor. Finalmente, si un tumor es detectado, el clasificador indicará el tipo de tumor, con el fin de estimar el grado de peligrosidad y urgencia de tratamiento que el paciente pudiera necesitar.

## 1.2. 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. Más concretamente y de forma más esquemática los objetivos concretos desde el anteproyecto son:

* **Estimación del mejor sistema de preprocesamiento de imágenes:** comprender sus características y posibles desafíos para que estén listas para su entrada a la CNN.
* **Sistema de etiquetado de las imágenes** para responder a los objetivos planteados, realización de aprendizaje supervisado.
* **Diseño y construcción de los modelos de CNN, así como su entrenamiento.**
  + Modelo que clasifique entre células de tejido mamario, de pulmón o de colon.
  + Modelo que clasifique entre células tumorales o sanas
  + Módelo que clasifique entre tipos de células tumorales.
* **Evaluación de los modelos CNN**. Para ello estimaremos medidas tales como: accuracy, recall, AUC. Intentaremos conseguir el mayor grado de precisión, sensibilidad y especificidad posible para nuestro modelo. También buscaremos que el modelo sea robusto y pueda extrapolarse a otras imágenes externas al dataset usado para el entrenamiento.
* **Desplegado de los modelos**
* **Presentación y discusión de los resultados** (objetivos cumplidos, limitaciones del trabajo, futuras líneas de trabajo).

## 1.3. Alcance

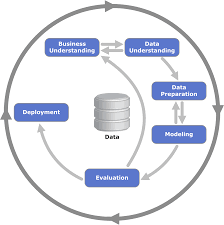
El alcance de este proyecto se centra en desarrollar y evaluar cuatro modelos de inteligencia artificial para la predicción de células tumorales. El objetivo principal es demostrar la viabilidad y efectividad de los modelos en la clasificación de células tumorales en diferentes partes del cuerpo, específicamente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

El trabajo 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

## 1.4. 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.



**Business Understanding:**

-Entender las necesidades clínicas y la relevancia de la detección y clasificación de los tipos de cáncer de mama, colon y pulmón.

-Identificar los indicadores de éxito del proyecto (por ejemplo, una cierta precisión en la clasificación de los tipos de cáncer, la mejora de la velocidad de procesamiento de las imágenes, etc.)

**Data Understanding:**

-Analizar los tres conjuntos de datos de imágenes para cada tipo de cáncer.

-Realizar un análisis exploratorio de los datos para entender sus características y posibles desafíos (por ejemplo, sesgos en los datos, calidad y resolución de las imágenes, etc.)

**Data Preparation:**

-Preparar los datos para la entrada a la red neuronal convolucional.

-Aplicar técnicas de preparación y aumento de datos, como centrar los datos restando los valores medios de los píxeles por canal, realizar aumentos aleatorios como recortes, volteos horizontales, perturbaciones de brillo, contraste y color, y cambios de escala​​.

**Modeling:**

-Diseñar y entrenar una red neuronal convolucional para cada tipo de cáncer utilizando los datos preparados.

-Ajustar los hiper parámetros de las redes para optimizar su rendimiento.

**Evaluation:**

-Evaluar el rendimiento de las redes neuronales en función de las métricas adecuadas (por ejemplo, precisión, sensibilidad, especificidad, AUC-ROC, etc.)

-Realizar una validación cruzada para asegurar que el modelo es robusto y generaliza bien a nuevos datos.

**Deployment:**

-Implementar los modelos en un entorno de producción, de manera que se puedan utilizar para clasificar nuevas imágenes de los tipos de cáncer de mama, colon y pulmón.

# 

# 

# State of art

En este apartado reflejaremos el “state of art”, que se refiere a las áreas de conocimiento involucradas con la solución del problema propuesto, así como las tecnologías relacionadas con dicha solución. Para ello, hablaremos de las tecnologías y soluciones desde una perspectiva más amplia a más específicas, intentando centrarnos principalmente en aquellas relacionadas con la clasificación de imágenes.

1. **Machine Learning**

Una de las tecnologías más empleadas para la clasificación de imágenes es el machine learning, aunque esta era principalmente utilizada antes de la llegada del deep learning. Las técnicas más tradicionales de machine learning involucran algoritmos tales como:  *Support Vector Machine* (SVM por sus siglas en inglés), *K-Nearest Neighbors* (KNN), *Decision Trees* (Árboles de Decisión) y *Random Forests*. Todas estas técnicas pueden ser utilizadas para extraer y seleccionar features de las imágenes histopatológicas, lo cual puede ayudar a detectar la presencia de cáncer en dichas imágenes.

1. **Deep learning**

El Deep Learning es un paradigma de aprendizaje automático bastante similar al aprendizaje humano donde se emplean redes neuronales artificiales para el procesamiento de la información. Una importante diferencia con los algoritmos tradicionales de machine learning es que utiliza tres o más layers, teniendo una mayor complejidad que le permite manejar y aprender de un mayor tamaño de datos. Sin embargo, cabe destacar que, aunque este mayor número de layers puede generar otro problema, que es que actuaría como una “caja negra” donde no se sabe exactamente qué features son los más útiles para clasificar las imágenes. Por tanto, sacrificamos transparencia en pos de obtener un clasificador más práctico y efectivo.

Dentro del Deep Learning hay distintas arquitecturas y modelos, por ejemplo podemos encontrar el *Feedforward Neural Networks* (FNNs) que son una de las formas más simples de redes neuronales, donde la información se mueven en una sola dirección desde el input hasta el output, pasando por los diferentes hidden layers. Sin embargo, para el procesamiento de imágenes, uno de los modelos más extendidos son los Convolutional Neural Networks (CNNs).

1. **Convolutional Neural Networks (CNNs)**

Estas redes son principalmente usadas para procesamiento de imágenes así como su clasificación. Convencionalmente ha sido una gran innovación capaz de identificar rostros y objetos entre otros. La estructura básica de CNNs se basa en un Convolutional Layer que realiza una serie de operaciones convolucionales en la imagen o input, lo cual permite extraer los features de la imagen, manteniendo la relación espacial entre los píxeles.

* **Proyectos similares**
* Toward robust mammography-based models for breast cancer risk [[7]](#_uhm5nyf8kunb) : Este proyecto se enfocó en desarrollar un modelo de riesgo de cáncer de mama basado en aprendizaje profundo (**deep learning**) que fuera más preciso que los modelos de riesgo clínico establecidos.

El estudio utilizó una muestra retrospectiva de 88,994 mamografías de cribado consecutivas en 39,571 mujeres. Se desarrollaron tres modelos de riesgo de cáncer de mama para predecir el riesgo dentro de los próximos 5 años: un modelo de regresión logística basado en factores de riesgo tradicionales (RF-LR), un modelo de aprendizaje profundo basado solo en imágenes (image-only DL) y un modelo de aprendizaje profundo híbrido que combinaba factores de riesgo tradicionales con mamografías.

Los resultados mostraron que el modelo híbrido de aprendizaje profundo y el modelo de aprendizaje profundo basado solo en imágenes obtuvieron áreas bajo la curva característica de operación del receptor (AUC) de 0.70 y 0.68, respectivamente. Estos valores de AUC fueron significativamente mayores que los obtenidos por el modelo RF-LR (AUC: 0.67) y el modelo de riesgo clínico establecido Tyrer-Cuzick (AUC: 0.62).

En conclusión, el estudio demostró que los modelos de aprendizaje profundo que utilizan mamografías completas ofrecen una discriminación de riesgo sustancialmente mejorada en comparación con el modelo Tyrer-Cuzick. Estos hallazgos respaldan el uso de modelos de aprendizaje profundo en la mejora de la predicción del riesgo de cáncer de mama.

* Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning [[8]](#_uhm5nyf8kunb): El artículo presenta una investigación sobre la detección de cáncer de piel utilizando **aprendizaje profundo** (deep learning) y **redes neuronales convolucionales** (convolutional neural network, CNN).

En este estudio, se utilizó el conjunto de datos ISIC2018, que contiene 3533 imágenes de lesiones cutáneas, incluyendo tumores benignos, malignos, no melanocíticos y melanocíticos. Para mejorar las imágenes, se aplicó el método ESRGAN (Enhanced Super-Resolution Generative Adversarial Networks) en una etapa de preprocesamiento. Luego, se empleó una red neuronal convolucional (CNN) para clasificar las imágenes de las lesiones cutáneas en tumores malignos y benignos. Además, se realizaron ajustes finos utilizando modelos de transferencia de aprendizaje como Resnet50, InceptionV3 e Inception Resnet.

El estudio demostró que el modelo diseñado alcanzó una tasa de precisión del 83.2% en la clasificación de las imágenes de lesiones cutáneas, en comparación con otros modelos pre-entrenados como Resnet50 (83.7%), InceptionV3 (85.8%) y Inception Resnet (84%) .

# Arquitectura del proyecto

El proyecto 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 casi perfectos con un menor costo y tiempo computacional.

El proyecto 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 proyecto 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 trabajo consiste en utilizar modelos adicionales específicos para cada parte del cuerpo (pulmón, colon 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".

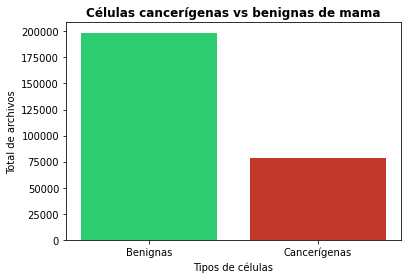
# DataSets

## 4.1. Fuente de los datos

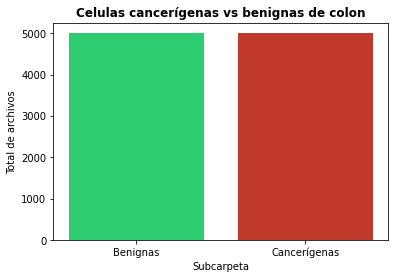
Los datos se han obtenido de dos fuentes Academic Torrents [1] y Kaggle [2]. De la primera se extrajeron los datos referentes a la células de pecho y de la segunda los de colon y pulmón

## 4.2. 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 colón (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 .



## 4.3. Preprocesamiento de los datos

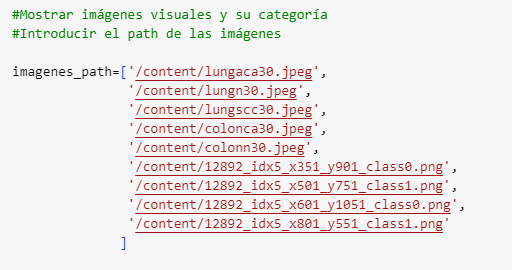
El primer paso en el procesamiento de datos consistió en balancear las clases, para ello, se redujo el número de imágenes de pecho a solo 10.000, seleccionadas de manera aleatoria y uniformemente, de tal manera que hubiese la misma cantidad de ejemplos con células benignas y cancerígenas.

Una vez realizado el balanceamiento de clases, se redimensiona 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].

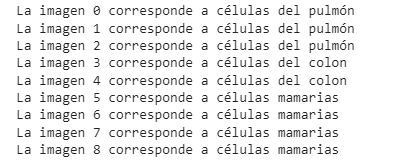
# 5. Funcionalidades del proyecto

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

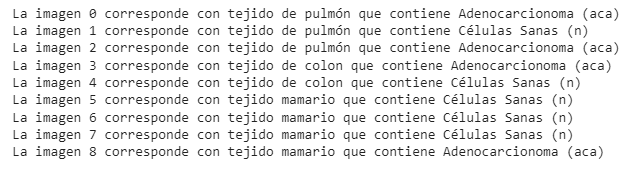


A continuación, bastaría con ejecutar el resto del código para visualizar los resultados. No haría falta ningún preprocesado manual de la imágen, ya que el propio código se encargará de reescalarla y redimensionarla.

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.



# 6. Evaluaciones

## 6.1. Métricas de rendimiento utilizadas

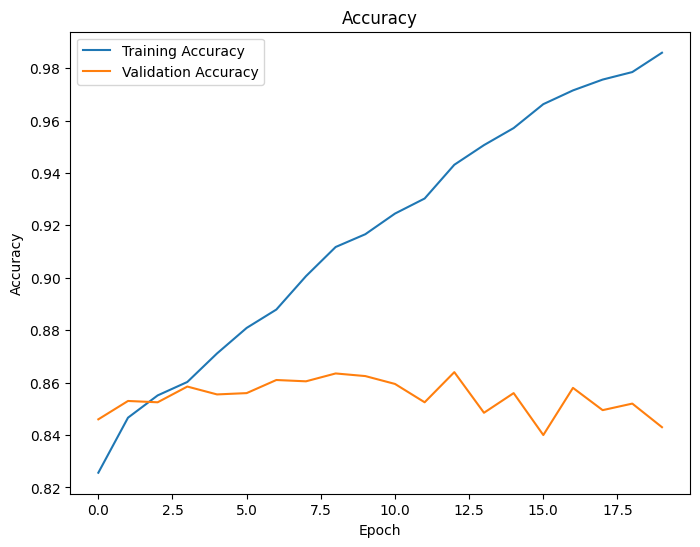
A continuación se explicarán detalladamente algunas de las métricas que se usarán para evaluar el rendimiento de cada uno de los 4 modelos.

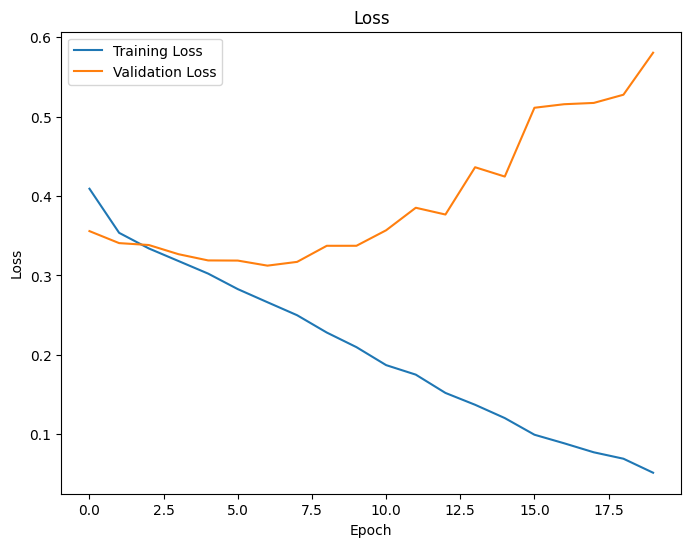
* **Matriz de Confusión**: La matriz de confusión es una tabla que muestra la precisión de un modelo de clasificación. Es una representación visual que muestra la cantidad de predicciones correctas e incorrectas realizadas por el modelo para cada clase de interés. La matriz se compone de cuatro elementos: 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.

### 6.1.1. Resultados preliminares y primeros obstáculos

Preliminarmente, obtuvimos un modelo no satisfactorio a la hora de clasificar el cáncer de mama:

| **Modelos** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| breast\_model.h5 | **0.8355** | **0.7867** | **0.8127** | **0.8460** |





Existen varias posibilidades por las que el modelo inicial pudo haber tenido un rendimiento inferior:

* **Información de contexto faltante:** Las imágenes fragmentadas pueden perder información de contexto sobre la ubicación y relación espacial de las células cancerosas. En la patología del cáncer, a veces la disposición espacial y la interrelación de las células y las características puede ser tan importante como las células y las características en sí.
* **Variabilidad inter-paciente:** Al tratar cada imagen fragmentada como una instancia de datos separada, el modelo podría haberse confundido por la variabilidad entre pacientes. Por ejemplo, las imágenes de un mismo paciente podrían ser más similares entre sí que las imágenes de diferentes pacientes. Al reunir todas las imágenes de un paciente en una sola imagen, estamos agrupando la variabilidad en una sola instancia, lo que podría ayudar al modelo a aprender mejor.
* **Desequilibrio de clases:** En el conjunto de datos original tenía un desequilibrio de clases, es decir, la cantidad de imágenes en las clases (0 y 1) era diferente. Esto puede hacer que el modelo sea sesgado hacia la clase con más ejemplos. Al usar técnicas de aumento de datos, hemos equilibrado las clases, lo que podría haber mejorado el rendimiento del modelo.

Debido a estos resultados insatisfactorios con este modelo, decidimos cambiar los pasos de preprocesamiento específicamente para detectar y clasificar el cáncer de mama, estimando el mejor sistema de preprocesamiento de dichas imágenes. De esta forma decidimos reorganizar el dataset y modificar el enfoque que se estaba utilizando para manejar las imágenes. En lugar de tener múltiples fragmentos de imágenes por cada paciente, optamos por consolidar todos los fragmentos de cada paciente en una sola imagen final compuesta para cada una de las clases (0 y 1).

Para lograr esto, cada fragmento de imagen fue ubicado en su respectiva coordenada de la imagen total, correspondiendo al lugar original que ocupaba en la imagen del tejido mamario. Se generó una imagen completa para cada clase (0: células benignas, 1: cáncer maligno de tipo carcinoma ductal invasivo) de cada paciente. Este proceso resultó en un nuevo conjunto de datos que contenía poco más de 558 imágenes, procedentes de los 279 pacientes en el estudio.

Sin embargo, nos encontramos con el desafío de tener un conjunto de datos desequilibrado, ya que el otro modelo (body\_part\_model.h5) que predecía el tipo de cáncer (colon, mama o pulmón) tenía muchos más ejemplos en el caso del cáncer de colon y pulmón. Para evitar sesgos en el aprendizaje y mejorar la robustez de nuestro modelo, decidimos aplicar técnicas de aumento de datos (data augmentation). Esta técnica nos permitió generar nuevas imágenes a partir de las originales, mediante la aplicación de transformaciones como rotaciones, desplazamientos y cambios de escala. Así, logramos nivelar la cantidad de ejemplos en las distintas clases.

Para aplicar la técnica de aumento de datos (data augmentations) hemos aplicado una serie de transformaciones aleatorias a nuestras imágenes. Estas transformaciones incluyen volteos horizontales y verticales, rotación aleatoria y ajustes aleatorios de brillo y contraste:

* **El volteo horizontal y vertical de las imágenes** ayuda a que nuestro modelo no sea sensible a la orientación de las características en las imágenes. Así, puede reconocer patrones de cáncer independientemente de si están al revés o reflejados.
* **La rotación de las imágenes** proporciona un beneficio similar, permitiendo al modelo identificar características clave independientemente de su ángulo.
* **Los ajustes de brillo y contraste** hacen que nuestro modelo sea robusto frente a las variaciones de iluminación en las imágenes.

Al implementar este nuevo conjunto de datos y evaluar el rendimiento de nuestro nuevo modelo, logramos resultados considerablemente superiores en todas las métricas evaluadas. Nuestro modelo ahora puede predecir de manera confiable y precisa si una imagen representa un carcinoma ductal invasivo o un caso benigno.

En conclusión, esta nueva estrategia nos ha permitido mejorar la capacidad de clasificación de nuestro modelo de detección de cáncer de mama, lo que pone de manifiesto que el modo en que preparamos y manejamos las imágenes puede tener un impacto enorme en la eficacia del modelo.

## 

## 6.2. Resultados obtenidos en la evaluación

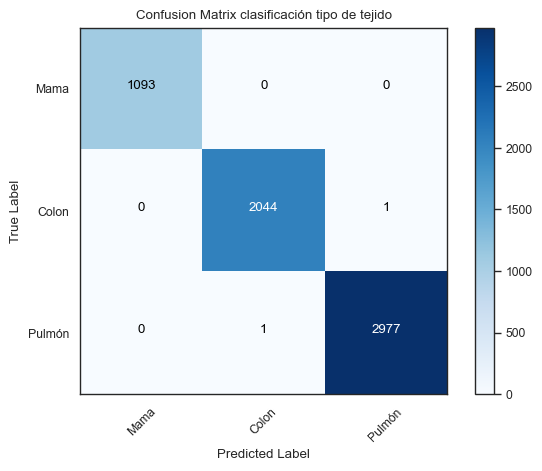
Tras la ejecución de los cuatro modelos se obtuvieron las siguientes métricas y resultados.

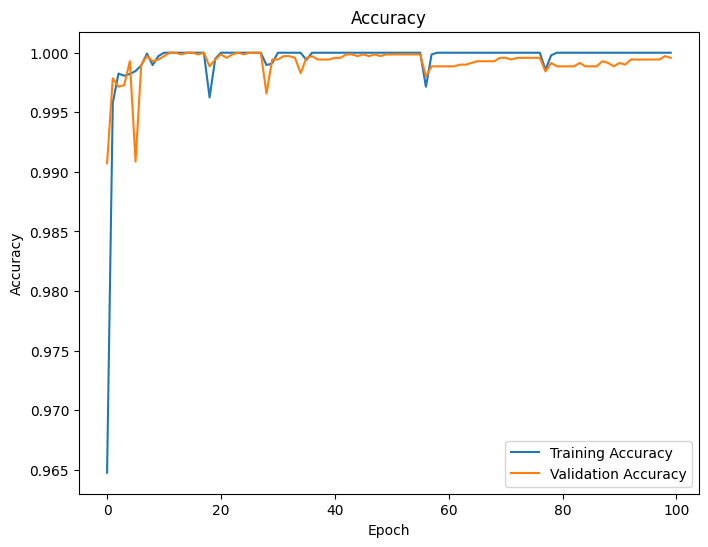
* **Métricas obtenidas**

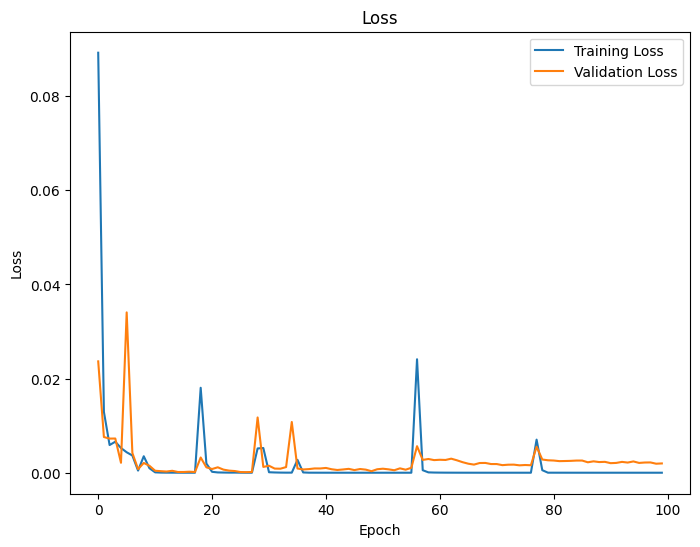
| **Modelos** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| body\_part\_model.h5 | **1.0** | **1.0** | **1.0** | **1.0** |
| breast\_model.h5 | **0.9731** | **0.9688** | **0.9789** | **0.9738** |
| colon\_model.h5 | **0.9585** | **0.9849** | **0.9301** | **0.9567** |
| lung\_model.h5 | **0.9520** | **0.9958** | **0.9907** | **0.9932** |

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

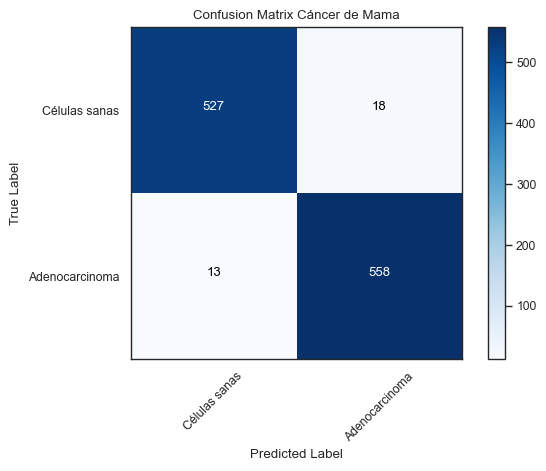
**Modelo división por partes**

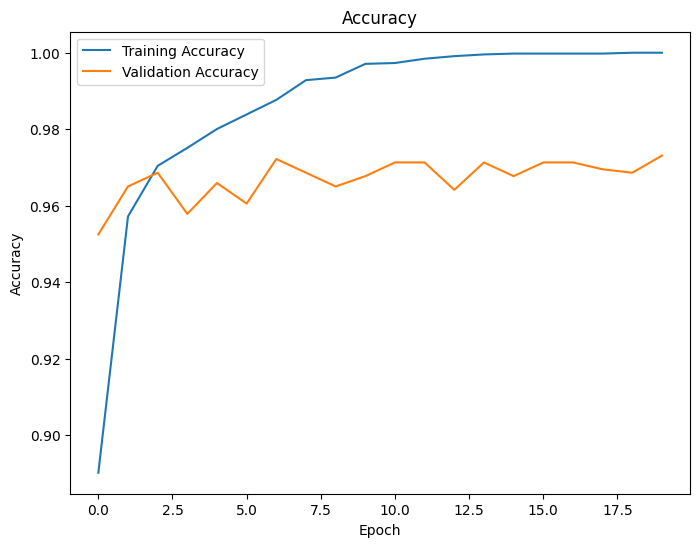


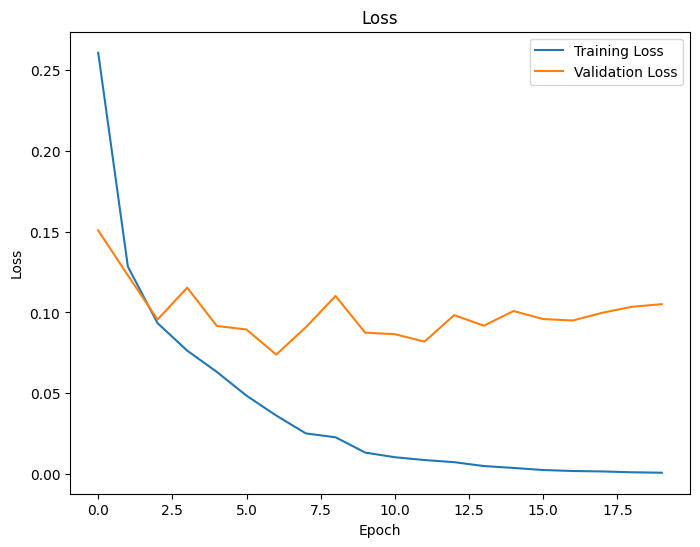




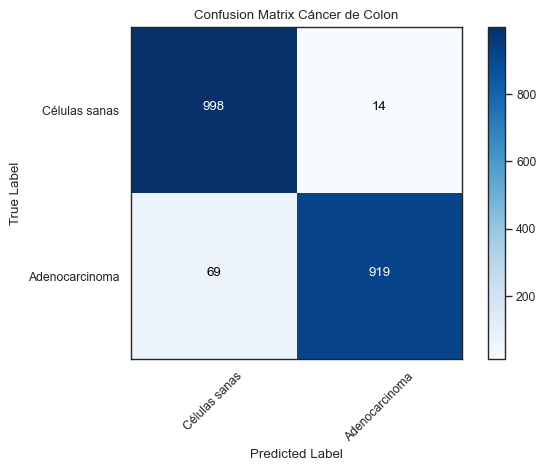
**Modelo células mamarias**

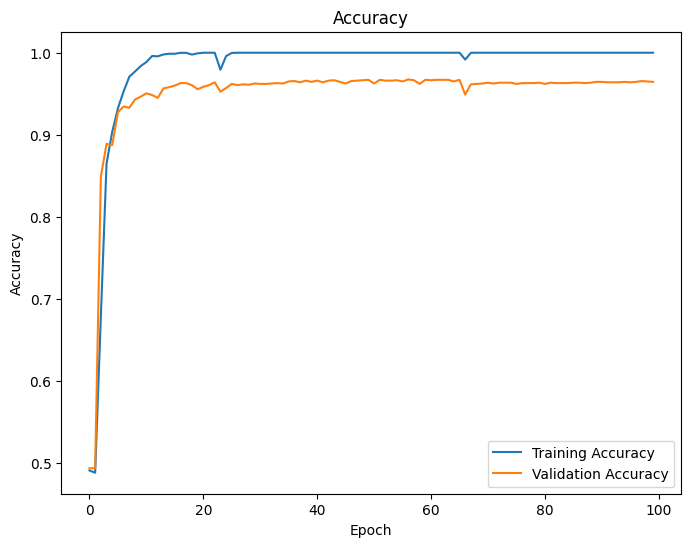


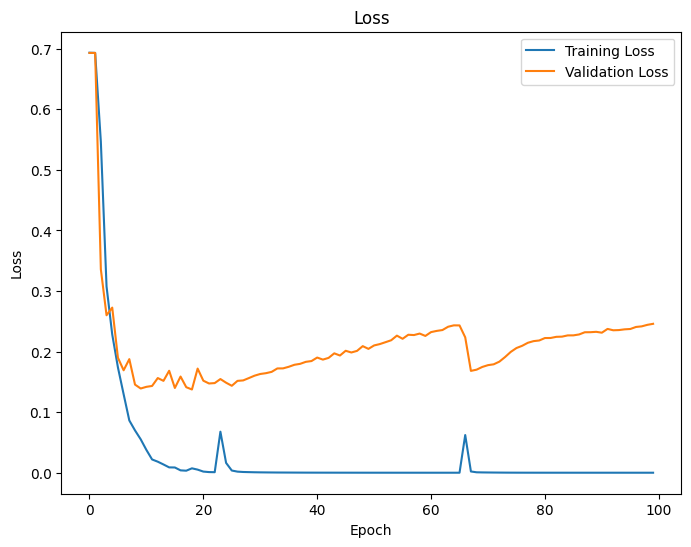




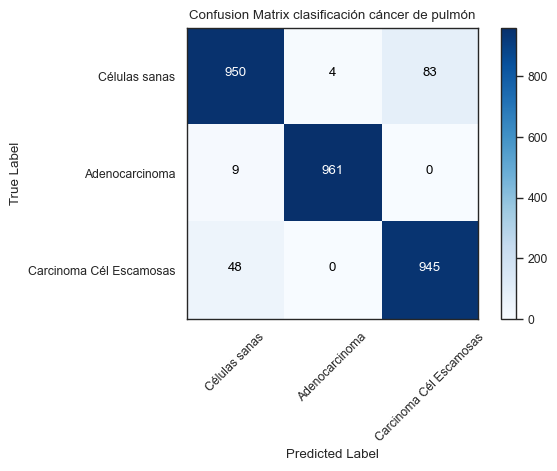
**Modelo células colon**

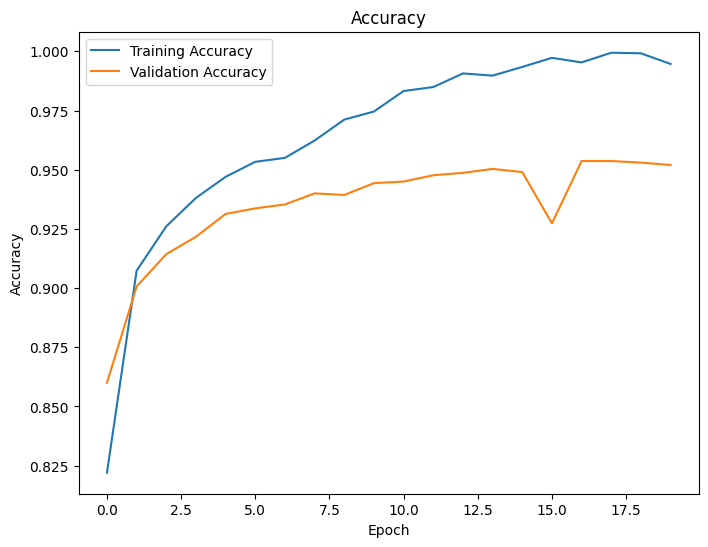


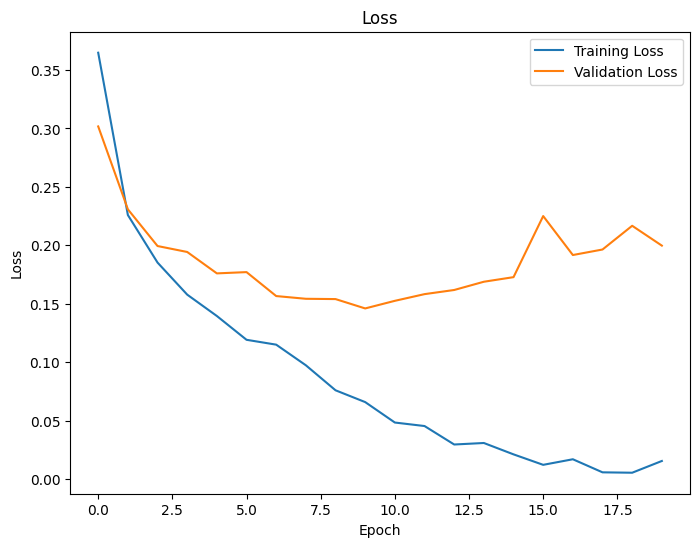




**Modelo células pulmón**







# 7. Conclusiones

En este trabajo se ha identificado el problema de crear un clasificador de imágenes de tejidos histológicos de distintas áreas del cuerpo y de clasificar si dicho tejido posee células sanas o células cancerígenas. De esta forma, hemos creado un modelo preciso para la clasificación y detección de cáncer de pulmón, mama y colon. Nuestro modelo pretende ser utilizado en el ámbito sanitario facilitando la toma de decisiones y el plan de tratamiento ante un diagnóstico de cáncer.

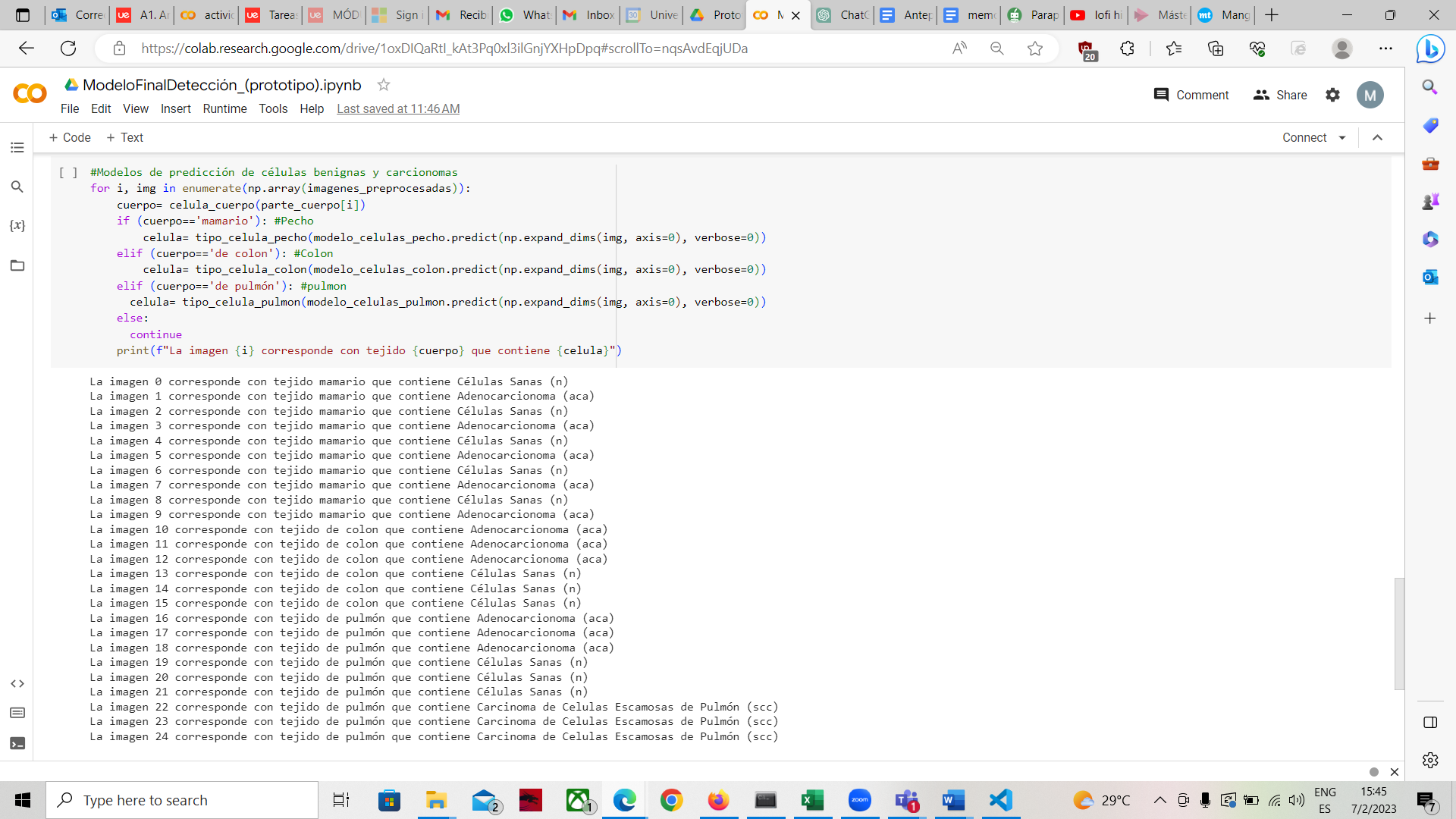
Observando las métricas obtenidas, podemos decir que el modelo muestra un desempeño adecuado a la hora de detectar el tipo de tejido, es decir, si pertenece a una muestra de pulmón, mama y colon, así como detectar la presencia de células cancerígenas en dichas áreas, cumpliendo con el objetivo general del proyecto. Así mismo, en el caso del cáncer de pulmón el modelo es capaz de determinar si el tejido cancerígeno es específicamente un adenocarcinoma o carcinoma de células escamosas de pulmón, siendo aún de mayor utilidad en el apoyo al diagnóstico.

Así mismo, prestando atención a los objetivos concretos, podemos decir que hemos cumplido satisfactoriamente los siguientes:

* **Estimación del mejor sistema de preprocesamiento de imágenes:** Hemos conseguido preprocesar las imágenes adecuadamente para obtener un modelo adecuado. Por ejemplo, en las imágenes de cáncer de mama hemos transformado los múltiples fragmentos de imágenes por paciente en una imagen consolidada para cada clase de carcinoma ductal invasivo, lo que ha mejorado notablemente la precisión y el rendimiento del modelo en esta área. El tratamiento y reorganización de las imágenes ha demostrado ser un factor determinante para la eficacia de la clasificación.
* **Sistema de etiquetado de las imágenes.** Con el fin de responder a los objetivos planteados y la realización de aprendizaje supervisado. Por ello, indicamos que cada imagen podría pertenecer a tejido de: colon, pulmón y mama. Así mismo, también etiquetamos las imágenes, que podían pertenecer a “células sanas” o “adenocarcinoma”. En el caso del cáncer de pulmón, añadimos la etiqueta adicional de “carcinoma de células escamosas de pulmón”.
* **Diseño y construcción de los modelos de CNN, así como su entrenamiento.**
  + Modelo que clasifique entre células de tejido mamario, de pulmón o de colon.
  + Modelo que clasifique entre células tumorales o sanas
  + Módelo que clasifique entre tipos de células tumorales.

Hemos sido capaces de diseñar y construir dichos modelos satisfactoriamente, así como entrenarlos y validarlos.

* **Evaluación de los modelos CNN**. En todos los modelos, obtuvimos un accuracy, precision, recall y F1-Score superior a 0.95 en la validación. En el caso del modelo que clasificaba entre células de tejido mamario, de pulmón o de colon, estos valores subían a 1.
* **Desplegado de los modelos.** Hemos conseguido desplegar los modelos y emplearlos en varias imágenes aleatorias para comprobar su efectividad y utilidad. Cada imagen pasa por los modelos y el código muestra si la imagen corresponde con un tejido específico y el tipo de células que contiene, como ilustra la siguiente imagen:



* **Presentación y discusión de los resultados**. Finalmente, hemos podido concluir que hemos cumplido los objetivos fijados, como se ha comentado al inicio de esta sección. Sin embargo es importante también reflejar ciertas limitaciones del trabajo aquí expuesto, así como futuras líneas de trabajo.

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.

Finalmente, cabe destacar que el resultado de este trabajo es solo un primer paso en otras lineas y futuros proyectos más ambiciosos. Nuestro modelo debería ser expandido para clasificar tejido histológico en más tipos específicos de tejido (por ejemplo, de estómago, de prostata, etc) y con distintos tipos de imágenes, con el fin de generar un modelo automatizado dotado de una mayor flexibilidad. Además, también se puede mejorar la clasificación del subtipo de cancer al obtener mayores muestras de otros tipos de cáncer (por ejemplo, además de adenocarcinoma, podemos encontrar cáncer de tipo inflamatorio o carcinoma lobular).

Así mismo, también se puede intentar entrenar el modelo para, no solo detectar el tipo de cancer, si no el estado en el que dicho cancer se encuentro. Esto facilitaría la selección del tratamiento adecuado y especializado a las situaciones concretas y particulares del paciente empleando el sistema de salud.

Además, futuros proyectos podrían aspirar a la integración de los resultados de modelos muy flexibles con datos clínicos, lo cual podría mantenerse en el historial del paciente para obtener una vista más comprensiva y completa de la situación y trayectoria del paciente para futuros diagnósticos de cáncer o asignación de tratamiento.

# 

# 8. Trabajos futuros

Tras haber completado el proyecto, una de las primeras mejoras que se propuso para mejorar la simplicidad, es la unificación de los cuatro modelos de inteligencia artificial en un único modelo. Esta unificación permitiría simplificar la estructura del proyecto al eliminar la necesidad de mantener y gestionar múltiples modelos por separado. Al combinarlos en uno solo, se lograría una mayor eficiencia y coherencia en el proceso de implementación y mantenimiento. Además, se reducirían los costos asociados con la infraestructura requerida para ejecutar y almacenar varios modelos.

Como siguiente propuesta, se sugiere la expansión del modelo para que pueda reconocer células tumorales de más partes del cuerpo. Esto permitiría aumentar la capacidad del modelo para detectar y clasificar células cancerosas en diferentes regiones anatómicas, lo que tendría un impacto significativo en la detección temprana y el tratamiento del cáncer. Al ampliar el alcance del modelo, se podría incluir el reconocimiento de células tumorales en órganos como la piel, el páncreas, el hígado, el estómago y otros sistemas del cuerpo, con el fin de generar un modelo automatizado dotado de una mayor flexibilidad. Además, también se puede mejorar la clasificación del subtipo de cáncer al obtener mayores muestras de otros tipos de cáncer (por ejemplo, además de adenocarcinoma, podemos encontrar cáncer de tipo inflamatorio o carcinoma lobular).

Así mismo, también se puede intentar entrenar el modelo para, no solo detectar el tipo de cáncer, si no el estado en el que dicho cáncer se encuentro. Esto facilitaría la selección del tratamiento adecuado y especializado a las situaciones concretas y particulares del paciente empleando el sistema de salud.

Además, futuros proyectos podrían aspirar a la integración de los resultados de modelos muy flexibles con datos clínicos, lo cual podría mantenerse en el historial del paciente para obtener una vista más comprensiva y completa de la situación y trayectoria del paciente para futuros diagnósticos de cáncer o asignación de tratamiento.

Por último, para obtener unos resultados más precisos y concluyentes, se propone la utilización de imágenes de mayor resolución para el análisis y el entrenamiento de modelos. Aunque todo esto requería de equipos de mayor calidad y mucho mayor tiempo de cómputo, este trabajo nos permitiría obtener un modelo mucho más efectivo que serviría como herramienta más completa y precisa para el diagnóstico.

En resumen, la suma de todos estos enfoques brindarán a los profesionales de la salud una herramienta más completa y precisa para el diagnóstico y la planificación de tratamientos personalizados. Al aprovechar esta tecnología basada en Deep Learning tan popular en los últimos años, podremos obtener una herramienta más precisa y completa para el diagnóstico de cáncer y la planificación de tratamientos personalizados, mejorando así la calidad de la atención médica y el bienestar de los pacientes.

# 9. Anexos

* 1. **Estructura de carpeta del proyecto**

Cáncer de mama grupo 1

01\_AnteProyecto

Preprocesado\_Analisis

dataset\_puros

pkl\_Finales

Pkl\_intermedios

imagenes\_muestras

modelos\_preprocesado

02\_Proyecto\_Intermedio

03\_Proyecto\_Final

Archivos\_Finales

Fotos\_Prueba

Colon

Mama

Pulmón

Modelos\_Entrenados(.h5)

Modelos\_Finales

* 1. **Diario del proyecto**

Empezamos a crear los grupos de trabajo desde el grupo de whatsapp en común. Expusimos nuestras ideas de los proyectos posibles y creamos un grupo de Teams para poder hacer reuniones y ponernos a trabajar en conjunto.

Una vez consensuado el proyecto que queríamos, empezamos a realizar el anteproyecto .

**Anteproyecto**

Durante el anteproyecto, los miembros del grupo se reunieron con el fin de identificar una propuesta válida y el proyecto. Así, tras observar el auge de la creación de modelos de inteligencia artificial para la detección y diagnóstico de cáncer, decidimos aportar nuestro grano de arena al crear un clasificador que además tuviera en cuenta distintos tipos de tejido. Así, propusimos la metodología para el desarrollo de un modelo adecuado y preciso, capaz de clasificar y detectar diferentes tipos de cáncer de mama, de pulmón y colon, empleando para ello entrenamiento de redes neuronales convolucionales con aprendizaje supervisado usando datos de imágenes histológicas, lo cual permitirá adquirir una herramienta de diagnóstico que contribuirá positivamente al tratamiento de casos de cáncer de colon y pulmón. Adicionalmente propusimos utilizar la metodología de trabajo CRISP-DM para definir todas las etapas del marco de trabajo, así como establecimos el diagrama de Gantt para organizar temporalmente el curso del trabajo durante el mes para finalizar satisfactoriamente el proyecto.

**Entrega intermedia (Prototipo)**

En este punto comenzamos a buscar toda la información necesaria para darle forma al proyecto. Empezamos creando una descripción del prototipo y determinando su arquitectura.

Con el fin de crear los modelos, obtuvimos el conjunto de datos de diferentes datasets con imágenes de tejidos histológicos de Mama, Colon y Pulmón. Tras determinar el tamaño y características de los datos, seleccionamos el mejor plan de acción para el preprocesamiento de los datos. Entre los pasos elegidos, se incluía la redimensión y el reescalado de las imágenes. Creamos un modelo aplicando red neuronal convolucional (CNN) que utiliza capas de convolución y capas densas para clasificar el tipo de tejido, así como un modelo específico para cada tejido para determinar si las células eran sanas o cancerígenas (y, en el caso del tejido de pulmón, determinar el tipo de cáncer).

Finalmente, una vez que los modelos fueron creados y comprobamos que obtuvimos unas métricas aceptables, testeamos su funcionalidad con imágenes histológicas seleccionadas aleatoriamente. Sin embargo, al final determinamos los siguientes objetivos en reducir el nivel de sobreentrenamiento de los modelos y mejorar el resultado en términos de accuracy.

A modo de resumen, al final del anteproyecto nos propusimos en pulir un modelo final que sea capaz, cuando añades una imagen de diferenciarla por tipo de cáncer y su estado. Adicionalmente, creamos la memoria intermedia, el video, el github y lo subimos para su validación como prototipo.

**Entrega final (Memoria)**

Durante las últimas etapas del trabajo, cambiamos el sistema de preprocesamiento en ciertos sets de datos con el fin de pulir la eficiencia de los modelos, especialmente al manejar las fotos pertenecientes al dataset de cáncer de mama. Tras aplicar la mejor, obtuvimos mejores resultados en todas las métricas de cáncer de mama. Finalmente en la memoria final hemos desplegado nuevamente los modelos finales y comprobado su efectividad y utilidad, probando imágenes seleccionadas aleatoriamente. En cada imagen aleatoria, el modelo indicó correctamente el tipo de tejido al que pertenecía y si las células eran categorizadas como sanas o no. Finalmente, se presentaron y discutieron los resultados, concluyendo que cumplimos los objetivos fijados aunque planteando las limitaciones encontradas así como futuras líneas y abordajes de trabajo.

# 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>

[3] Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: a cancer journal for clinicians, 73(1), 17-48.

[4] Chaunzwa, T. L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., ... & Aerts, H. J. (2021). Deep learning classification of lung cancer histology using CT images. Scientific reports, 11(1), 1-12.

[5] Hameed, Z., Garcia-Zapirain, B., Aguirre, J. J., & Isaza-Ruget, M. A. (2022). Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network. Scientific Reports, 12(1), 15600.

[6] Zhou, P., Cao, Y., Li, M., Ma, Y., Chen, C., Gan, X., ... & Chen, C. (2022). HCCANet: histopathological image grading of colorectal cancer using CNN based on multichannel fusion attention mechanism. Scientific Reports, 12(1), 15103.

[7] Adam Yala et al. "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction." Radiology, 2019 Jul. PMID: 31063083.

[8] Artículo: Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. Disponible en: https://www.mdpi.com/2227-9032/10/7/1183