****

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

**(Anteproyecto)**

**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. Identificación de la propuesta y del proyecto 2**](#_hbgrxolqnwym)

[**2. Intervinientes del proyecto y resumen (Introducción) 3**](#_vvvbtjbwowoo)

[**3. Antecedentes y estado actual 4**](#_prmajf91kkkl)

[**4. Objetivos del proyecto 5**](#_fr3zh7e7p5bc)

[4.1. Objetivo general 5](#_q7bp86hwrk0v)

[4.2. Objetivos concretos 5](#_h9dihjoswkui)

[**5. Plan de trabajo y objetivos del proyecto 6**](#_opmt13nsa8me)

[5.1. Metodología 6](#_c4nlkvdemab3)

[5.2. Diagrama de Gantt 8](#_ip8vx2bcxqm7)

[5.3. Tecnologías propuestas 9](#_pu3zhdc3fsrd)

[**6. Experiencia previa en el tema 10**](#_gn4fuweimilr)

[**7. Citas 11**](#_qe4ltfa2cgd1)

# 

# **Identificación de la propuesta y del proyecto**

Detección de cáncer de pulmón y colon

Este proyecto tiene como objetivo desarrollar un modelo preciso para la clasificación y detección de cáncer de pulmón, mama y colon utilizando técnicas avanzadas de inteligencia artificial y clasificación de imágenes. Además, el modelo será capaz de detectar el tipo de cáncer específico detectado en dichas áreas histológicas. La metodología propuesta implica el preprocesamiento del conjunto de datos, la extracción de características y el entrenamiento de redes neuronales convolucionales (CNN) para tareas de clasificación. Finalmente, el modelo será evaluado mediante métricas de rendimiento para encontrar el modelo con el mayor potencial para funcionar como herramienta rápida de diagnóstico de cáncer basadas en imágenes histológicas, lo cual podría tener grandes implicaciones en el ámbito sanitario al facilitar la toma de decisiones y el plan de tratamiento ante un diagnóstico de cáncer.

# 

# **Intervinientes del proyecto y resumen (Introducción)**

El equipo desarrollador de este proyecto se compone de Martín Antúnez García, Luis Acero Morata, Carlos Arcas Perez y Javier Harto Vallejo. El director-cliente del proyecto es Teno González Dos Santos.

En la actualidad, es bien sabido que una de las enfermedades más graves y con mayor tasa de mortalidad es el cáncer, el cual se cobra la vida de cientos de miles de personas cada año. Afortunadamente, datos más esperanzadores como los extraídas del National Center for health Statistics (Siegel, Miller, Wagle & jemal, 2023), indican que los índices de mortalidad se están reduciendo en los últimos años, debido a la inversión en sanidad y, quizás más importante, a la aplicación de nuevas tecnológicas en el diagnóstico y tratamiento del cáncer. Esto es particularmente relevante en tipos de cánceres muy agresivos, tales como el Carcinoma Ductal Invasivo (IDC) que se presenta como uno de los subtipos de cáncer más comunes. Este proyecto quiere aportar su grano de arena a esta causa al proporcionar un modelo preciso para la clasificación y detección de cáncer de pulmón y colon, utilizando para ello técnicas avanzadas de inteligencia artificial y clasificación de imágenes.

Para alcanzar dicho objetivo, proponemos la siguiente metodología. El primer paso es realizar técnicas de aprendizaje supervisado, buscando una base de datos que estén previamente etiquetados para entrenar el modelo. Una vez adquiridos dichos datos, para comenzar, realizaremos un exhaustivo preprocesamiento de los datos de imágenes histológicas, que serán de pulmón, mama y colon. Tras el preprocesamiento de las imágenes, realizaremos un entrenamiento de redes neuronales convolucionales (CNN, por sus siglas en inglés), con el fin de llevar a cabo la tarea de clasificación. Mediante el uso de CNN seremos capaces de captar aquellos patrones y características de las imágenes a distintos niveles de abstracción, lo que permitirá entrenar un modelo adecuado para semejante tipo de datos.

A modo de resumen, este proyecto se centrará en el desarrollo de 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.

# 

# **Antecedentes y estado actual**

Como se ha comentado anteriormente, 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 neghbors 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.

# 

# **Objetivos del proyecto**

## 4.1. Objetivo general

Desarrollar un modelo de clasificación de imágenes histológicas en cáncer de pulmón y cáncer de colon, así como el tipo de cáncer específico.

## 4.2. Objetivos concretos

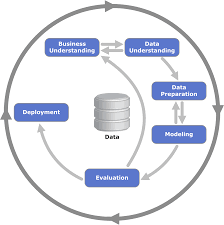
* **Estimación del mejor sistema de preprocesamiento de imágenes:** comprender sus características y posibles desafíos para que estén listas para su entrada a la CNN.
* **Sistema de etiquetado de las imágenes** para responder a los objetivos planteados, realización de aprendizaje supervisado.
* **Diseño 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).

# 

# **Plan de trabajo y objetivos del proyecto**

## 5.1. Metodología

En este proyecto usaremos la metodología de trabajo *CRISP-DM* para definir todas las etapas del marco de trabajo, empezando desde la definición y análisis de los datos, pasando por el desarrollo de los modelos y la evaluación de sus métricas, hasta la implementación final y su 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.

## 5.2. Diagrama de Gantt

En este apartado se presenta el diagrama de Gantt correspondiente al desarrollo de la investigación, desde la definición del problema hasta la presentación de los resultados

| Actividades | SEMANA 1  (1-4 Junio) | SEMANA 2  (5-11 Junio) | SEMANA 3  (12-18 Junio) | SEMANA 4  (19-25 Junio) | SEMANA 5  (29-30 Junio) |
| --- | --- | --- | --- | --- | --- |
| Business Understanding |  |  |  |  |  |
| Data Understanding |  |  |  |  |  |
| Data Preparation |  |  |  |  |  |
| Modeling |  |  |  |  |  |
| Evaluation |  |  |  |  |  |
| Deployment |  |  |  |  |  |

# 

# 

## 5.3. Tecnologías propuestas

En este proyecto emplearemos una serie de tecnologías estandarizadas en el campo de la Inteligencia Artificial y el procesamiento de imágenes. Estas tecnologías comprenden:

* **Python:** Optamos por Python porque es fácil de usar y está ampliamente respaldado por la comunidad de ciencia de datos. En el campo del aprendizaje automático, es el lenguaje de programación preferido.
* **NumPy:** Para manipular datos en forma de matrices, que es un formato común en el procesamiento de imágenes, esta biblioteca fundamental de Python será esencial.
* **TensorFlow y Keras:** Para diseñar, entrenar y evaluar nuestras Redes Neuronales Convolucionales (CNNs), usaremos estos recursos. Las CNNs tienen una capacidad para operar directamente sobre una matriz de píxeles de una imagen y aprender características jerárquicas, además de que son especialmente útiles para el procesamiento de imágenes.
* **Jupyter Notebooks**: Utilizaremos Jupyter Notebooks para experimentar con el código y documentarlo de manera interactiva.
* **Scikit-learn:** Las tareas de aprendizaje automático y análisis de datos serán beneficiosas con esta herramienta.

# 

# **Experiencia previa en el tema**

Para este proyecto, hemos elegido un equipo multidisciplinar que colaborará de manera igualitaria en cada una de las etapas del proyecto. El equipo está compuesto por un ingeniero informático (Luis), desarrollador de aplicaciones web y multiplataforma (Javi), psicólogo (Martín) y Técnico Electrónico (Carlos).

El equipo tiene varias habilidades complementarias que les permite: -Tener un entendimiento profundo del manejo de datos. -Creación y despliegue de modelos de inteligencia artificial. -Entendimiento de la literatura y de las necesidades del ámbito sanitario. Los miembros pertenecen a distintas empresas relacionadas con la tecnología y el uso de proyectos de este tipo.

# 

# **Citas**

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.

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.

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.

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.