DETECCIÓN DE ADENOCARCINOMA DE COLON MEDIANTE MÉTODOS DE *DEEP LEARNING*

Julen Bohoyo Bengoetxea

TRABAJO FINAL DE GRADO BIOTECNOLOGÍA

Tutor académico: Nombre: Santiago Garcia Vallvé

Departamento: Departamento Bioquímica y Biotecnología, URV

Email: santi.garcia-vallve@urv.cat

En cooperación con: Instituto de Investigación Sanitaria Biodonostia

Supervisor: Nombre: Dr. Marcos J. Araúzo-Bravo

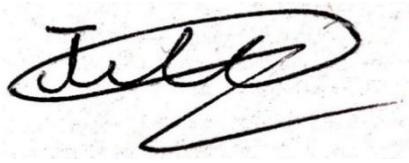
Departamento: Biología Computacional, IIS Biodonostia

Email: mararabra@yahoo.co.uk

Tarragona, junio 2022

Jo, JULEN BOHOYO BENGOETXEA, amb DNI 45197190M, sóc coneixedor de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 4 de junio de 2020



CONTENIDO

DATOS DEL CENTRO

He cursado mis prácticas curriculares en el Departamento de Biología Computacional del Instituto de Investigación Sanitaria Biodonostia (IIS Biodonostia) durante los meses de julio y agosto del año 2021. Además, y tras la finalización de las mismas, actualmente continúo trabajando con la entidad con el fin de seguir avanzando en el proyecto.

El IIS Biodonostia es el primer instituto de investigación sanitaria de Euskadi, fundado en 2008. Se trata de uno de los tres centros de investigación sanitaria a través de los cuales Osakidetza (Servicio Vasco de Salud) lleva a cabo sus principales actividades de Investigación e Innovación.

El mencionado Departamento de Biología Computacional está dirigido por el Dr. Marcos J. Araúzo-Bravo quien, además, ha sido mi tutor profesional durante la realización de las prácticas. La posibilidad que se me ha brindado de colaborar en los proyectos desarrollados en esta institución me ha permitido poner en práctica los conocimientos multidisciplinares que he adquirido durante mis estudios, pudiendo por primera vez combinar la biotecnología con el desarrollo de software.

Dirección: Paseo Dr. Begiristain, s/n, 20014 San Sebastián, Gipuzkoa

Teléfono: 943 00 60 12

RESUMEN Y PALABRAS CLAVE

El adenocarcinoma de colon es uno de los tipos de cáncer con mayor incidencia en la población. Debido a que el diagnóstico y tratamiento temprano son esenciales para lograr una mayor tasa de supervivencia, es necesario implementar programas de cribado que permitan analizar de forma periódica a los sectores de población con mayor riesgo de contraerlo. Uno de los métodos de diagnóstico más utilizados por los sistemas sanitarios es el análisis de muestras histológicas. El análisis visual de dichas muestras requiere de gran cantidad de tiempo por parte de expertos. Por lo tanto, resulta conveniente desarrollar una herramienta que agilice este trabajo a los patólogos. Debido a que se trata de una tarea de visión por computador, y que en los últimos años la inteligencia artificial ha demostrado tener un gran potencial en este ámbito, se ha optado por utilizar las redes neuronales artificiales con el fin de desarrollar dicha herramienta. Sin embargo, uno de los principales escollos que se suelen encontrar en esta clase de desarrollos de ámbito biomédico es la falta de datos de calidad necesarios para el entrenamiento de las redes neuronales. Por esta razón, se han aplicado múltiples técnicas que ayudan a paliar este inconveniente, permitiendo obtener un sistema razonablemente preciso a partir de un conjunto limitado de imágenes.

Palabras clave: adenocarcinoma de colon, red neuronal artificial, visión por computador, clasificación, segmentación.

INTRODUCCIÓN

Cáncer Colorrectal

Según la Organización Mundial de la Salud (OMS), el cáncer colorrectal (CRC del inglés c*olorectal cancer*) es el segundo tipo de cáncer que más muertes provocó (935 000 muertes) y el tercero más diagnosticado (1,93 millones de casos) en 2020 (OMS Cancer, 2021).

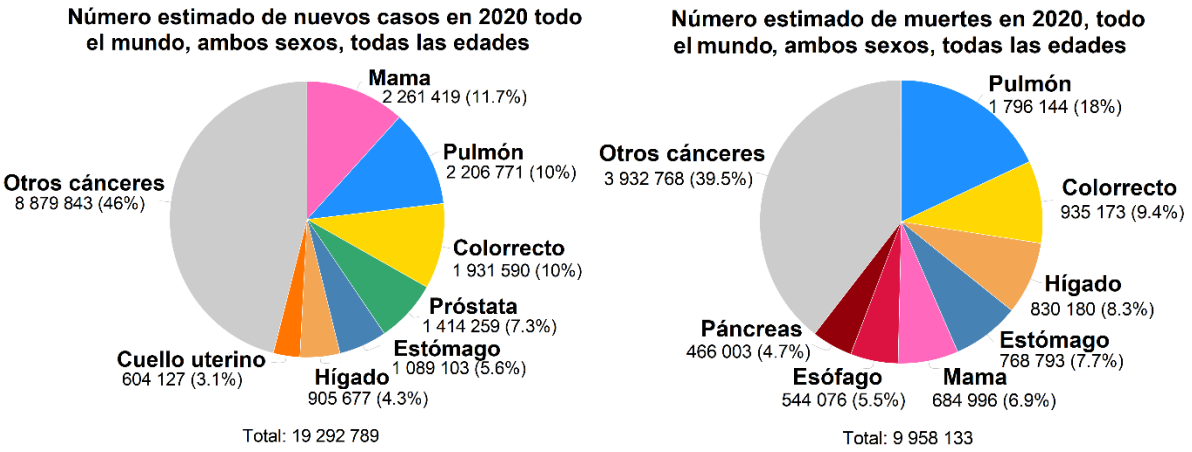


Figura 1: Porcentajes de nuevos casos y muertes por tipos de cáncer en 2020, datos de GLOBOCAN, adaptado de (Cancer Today, 2021).

En la mayoría de los casos, el CRC comienza como un pólipo; es decir, un crecimiento no canceroso que aparece en la mucosa del colon o del recto (American Cancer Society, 2021). Estos pólipos son muy comunes y se pueden encontrar en la mitad de la población mayor de 50 años, aunque la mayoría de ellos no acaban representando ningún riesgo para el individuo (American Cancer Society, 2021). En función del patrón de crecimiento, los pólipos pueden ser clasificados como adenomatosos, cuando se desarrollan masas semejantes a glándulas y que son el precursor de CRC más usual; o serrados, llamados así por su forma de sierra (American Cancer Society, 2021) (National Institutes of Health, 2021).

El cáncer colorrectal se desarrolla, generalmente, por la vía adenoma-carcinoma (Brenner et al., 2014). Un proceso mediante el cual un pólipo adenomatoso pasa a convertirse en adenocarcinoma de colon (Brenner et al., 2014). Los

adenocarcinomas de colon representan alrededor del 90% de los casos de CRC (Munro et al., 2018). Si el adenocarcinoma se propaga y llega a invadir algún tejido de la pared del colon más allá de la mucosa, se conoce como adenocarcinoma infiltrante o invasivo (American Cancer Society, 2021). En este caso el riesgo es mucho mayor, ya que puede derivar en metástasis (Neo et al., 2010). De hecho, el 70% de las muertes de pacientes de CRC son causadas por metástasis al hígado (Neo et al., 2010).

Factores de Riesgo

En cuanto a los factores de riesgo a tener en cuenta, lo más común es que el cáncer suceda de manera esporádica, tal y como se puede apreciar en la *Figura 2*. Los casos en los que se reportan antecedentes de familiares con esta enfermedad representan tan sólo un cuarto de los pacientes. Por último, únicamente un 5% se asocia a un síndrome de cáncer hereditario (Keum et al., 2019).

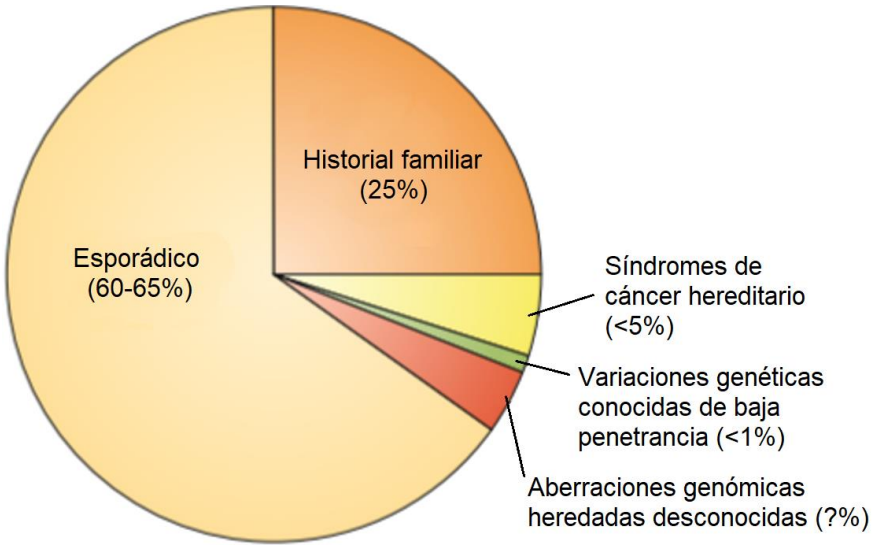


Figura 2: Proporción de casos de CRC asociados a factores genéticos y esporádicos, adaptado de (Keum et al., 2019).

En cuanto a los factores exógenos, se ha demostrado que el más relevante es una dieta inadecuada. En concreto, la ingesta excesiva de carne roja, carnes ultra procesadas, así como el consumo de carnes cocinadas a temperaturas muy altas

pueden aumentar el riesgo de cáncer (Labianca et al., 2010). Por el contrario, existen evidencias de que una dieta alta en calcio, fibra, antioxidantes, productos lácteos y granos integrales puede ayudar a disminuir la probabilidad de padecer CRC (Song et al., 2015).

Asimismo, también se han asociado a este tipo de cáncer algunos factores no relacionados con la dieta, como el uso continuado de antiinflamatorios no esteroideos, el consumo habitual de aspirinas y, principalmente, el sedentarismo y el tabaquismo (Labianca et al., 2010). Los individuos que practican actividad física regularmente tienen un 25% menos de probabilidades frente a la media de sufrir CRC; por el contrario, las personas más sedentarias tienen un 50% más de probabilidades de contraerlo (Rawla et al., 2019).

Por último, como en la mayoría de los tipos de cáncer, la edad es uno de los principales factores de riesgo, ya que a partir de los 50 años la probabilidad de padecer un CRC aumenta drásticamente (Keum et al., 2019)

Tabla 1: Factores de riesgo y prevención, adaptado de (Keum et al., 2019).

Factores de riego	Evidencia
Obesidad	↑↑
Actividad física	↓↓
Dieta occidental	↑↑
Dieta prudente	↓↓
Carne procesada	↑↑
Carne roja	↑
Fibra total	↓
Granos integrales	↓
Alcohol (etanol)	↑↑
Tabaquismo	↑
Consumo de aspirinas	↑↑
Calcio total	↓

↑↑ riesgo convincente, ↑ riesgo probable, ↓↓ protección convincente, ↓ protección probable.

Por todo lo anteriormente mencionado, la mayor incidencia de cáncer colorrectal se asocia a la población de los países más desarrollados, tal y como se puede observar en la *Figura 3*. De hecho, la relación de la incidencia en países industrializados frente a la de países no industrializados es de más de 2:1. Como consecuencia, es en Asia y en Europa Oriental donde actualmente se aprecia un mayor incremento en el número de casos, ya que se trata de zonas en vías de desarrollo (Labianca et al., 2010).

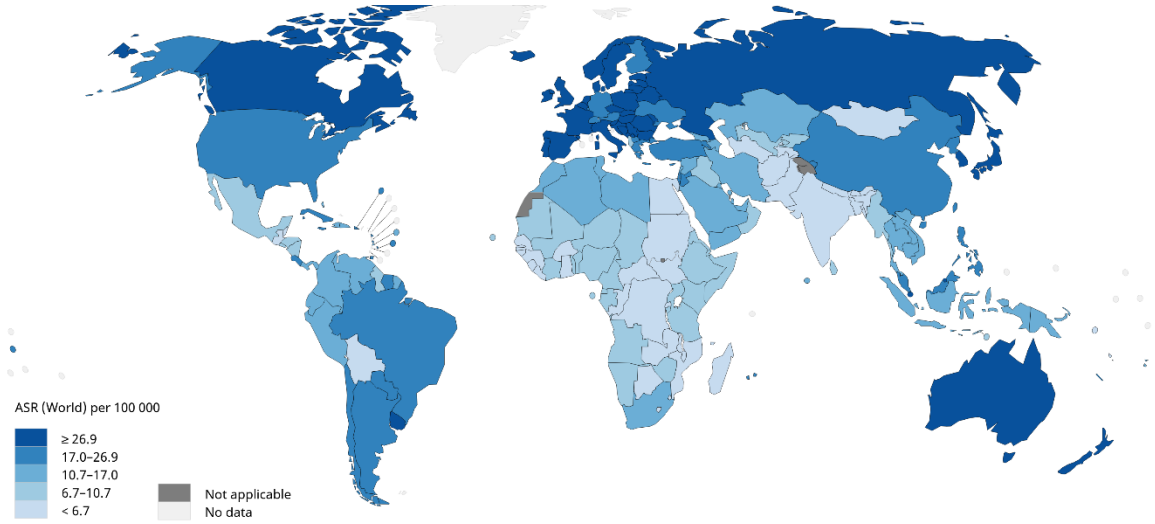


Figura 3: Mapa estimado de tasas de incidencia ASR (age-standardized) en 2020, cáncer colorrectal (Cancer Today, 2021).

Estadificación

La determinación de la etapa en la que se encuentra el cáncer o estadificación se realiza en base a la clasificación TNM (tumor, nódulos y metástasis), propuesta por el *American Joint Committee on Cancer* (AJCC). Esta clasificación indica cómo de avanzado está el cáncer y cuan grave es, por lo que es determinante a la hora llevar a cabo una prognosis, así como de establecer un tratamiento (Sagaert et al., 2018) (Edge et al., 2010).

Tabla 2: Clasificación TNM de las etapas del cáncer, adaptado de (Labianca et al., 2010).

Clasificación TNM
Tumor primario (T)
TX: No puede evaluarse el tumor primario
T0: Sin evidencia de un tumor primario
Tis: Carcinoma in situ: intraepitelial o invasión de la lámina propria
T1: El tumor invade la submucosa
T2: El tumor invade la muscularis propia
T3: El tumor invade la subserosa
T4: El tumor invade otros órganos o estructuras y/o perfora el peritoneo visceral
Nodos linfáticos regionales (N)
NX: No pueden evaluarse los nodos linfáticos cercanos
N0: No hay metástasis en los nodos linfáticos cercanos
N1: Metástasis entre 1 a 3 nodos regionales
N2: Metástasis en 4 o más nodos regionales
Metástasis distante (M)
MX: No puede evaluarse la metástasis distante
M0: No hay metástasis distante
M1: Hay metástasis distante

El pronóstico del paciente está claramente relacionado con el grado de penetración que presenta el tumor en la pared del colon, así como con la presencia o ausencia de nódulos (Labianca et al., 2010) . Por lo tanto, es crucial para los patólogos identificar correctamente qué tejidos se encuentran afectados.

Diagnóstico

Cuando el CRC se detecta en una fase temprana la tasa de supervivencia a 5 años es del 90% (American Cancer Society, 2021), por lo que un diagnóstico y tratamiento temprano es decisivo para mejorar el pronóstico del paciente.

El cáncer colorrectal temprano no produce síntomas apreciables en individuo (Labianca et al., 2010). Además, muchos de los síntomas que pueden aparecer como consecuencia del cáncer colorrectal son muy inespecíficos: cambio en los hábitos intestinales, malestar abdominal, pérdida de peso sin causa aparente o cansancio constante (Labianca et al., 2010). Por estas razones, es esencial realizar grandes esfuerzos en la detección a través de la implementación de programas de cribado en la población (Labianca et al., 2010).

Existen diversos métodos que ayudan a los médicos a detectar precozmente el cáncer colorrectal, desde los menos invasivos, como el análisis de materia fecal o la prueba de ADN en sangre, hasta los más invasivos, como la sigmoidoscopia o la colonoscopia, que a su vez pueden ser acompañados de una biopsia (National Institutes of Health, 2021). En cualquier caso, cuando en una prueba poco invasiva se obtiene un resultado positivo, se recomienda verificar el diagnóstico mediante una colonoscopia (National Institutes of Health, 2021).

Las principales características que valorar en una prueba de cribado son su sensibilidad (tasa de verdaderos positivos) y su especificidad (tasa de falsos positivos). Cuando la no detección del cáncer representa un mayor peligro se prioriza la sensibilidad, mientras que cuando el sobretratamiento representa un mayor peligro se prioriza la especificidad (Simon, 2016). Además, una prueba debe ser sencilla de realizar y fácilmente tolerable por el paciente, ya que en la mayoría de los casos se realizan a pacientes asintomáticos (Simon, 2016).

Tabla 3: Métodos de screening de CRC, adaptado de (Simon, 2016).

Test	Premisa	Sensibilidad para CRC	Intervalo de cribado	Ventajas	Limitaciones
Colonoscopia	Examen endoscópico del colon completo	>95%	Cada 10 años	- Alta sensibilidad. - Visualización completa del colon. - Detección de lesiones distales y proximales. - Se pueden eliminar lesiones.	- Invasivo. - Desagradable. - Requiere instalaciones especializadas y sedación. - Coste. - Accesibilidad. - Riesgo de perforación de intestino.
Sigmoidoscopia	Examen endoscópico del colon distal	>95% (solo colon distal)	Cada 5 años junto a FOBT	- Alta sensibilidad. - No requiere sedación completa. - Se pueden eliminar lesiones	- Semi-invasivo. - Desagradable. - Requiere instalaciones especializadas y sedación. - Coste. - Accesibilidad. - Solo colon distal. Seguridad.
Corografía CT (Computed tomography)	Visualización radiológica del colon	>90%	Cada 5 años	- Alta sensibilidad. - Visualización completa del colon. - No requiere sedación. - Detección de lesiones distales y proximales.	- Semi-invasivo. - Desagradable. - Requiere instalaciones especializadas. - No puede eliminar lesiones. Seguridad radiológica.

FOBT (Fecal occult blood test)	Detección enzimática de hemoglobina en las heces	33%-75%	Anual	- Accesibilidad. - No invasivo. - Bajo coste. - Detección de lesiones distales y proximales.	- Detección pobre de lesiones precancerosas. - No puede eliminar lesiones. - Detecta hemoglobina ingerida.
FIT (Fecal immunochemi-cal test)	Detección enzimática de hemoglobina en las heces	60%-85%	Anual	- Accesibilidad. - No invasivo. - Bajo coste. - Detección de lesiones distales y proximales.	- Detección pobre de lesiones precancerosas. - No puede eliminar lesiones
Test mt-sDNA	Detección molecular de aberraciones del ADN y hemoglobina.	92%	Cada 3 años	- Alta sensibilidad. - Accesibilidad. - No invasivo. - Detección de lesiones distales y proximales.	- Peor detección de lesiones precancerosas. - No puede eliminar lesiones

Aprendizaje profundo

El aprendizaje profundo (DL del inglés *deep learning*) es una subdisciplina del aprendizaje automático o *machine learning,* que pretende imitar el funcionamiento del cerebro humano creando una red neuronal artificial (ANN del inglés *artificial neural network*) que aprenda a partir de ejemplos de funcionamiento. Una vez definida la estructura o arquitectura de la red, la ANN se entrena de forma autónoma a partir de una base de datos (Wen et al., 2020).

Por este motivo, es una herramienta idónea para utilizar en problemas científicos en los que se trabaja con un gran y complejo conjunto de datos (Wen et al., 2020). Algunas aplicaciones en las que el DL se utiliza de forma exitosa son en el reconocimiento de escritura manual, el reconocimiento de imágenes, el reconocimiento de voz, la comprensión de lenguaje natural, el modelado acústico o la biología computacional (Emmert-Streib et al., 2020).

La unidad mínima de una red neuronal artificial es denominada neurona (x) (Emmert-Streib et al., 2020) (véase *Figura 4)*. Esta neurona funciona de forma similar a su homóloga real: a través de conexiones con otras neuronas, la neurona recibe n señales binarias (x_1, x_2, …, x_n), es decir, que pueden adoptar un valor de 0 o 1. Cada una de estas señales recibidas a través de estas “sinapsis” es modulada por un valor llamado peso (w), que puede tener un efecto excitatorio o inhibidor. Por lo que finalmente a la neurona llega una secuencia de n valores modulados por sus

respectivos pesos (x_1*w_1, x_2*w_2, …, x_n*w_n). La suma (Σ) de estos valores constituye lo que se denomina el valor de activación (a) de la neurona (Gurney et al., 1997). La función de activación (θ) es la función que se utiliza para computar esta suma ponderada y decidir si la neurona se activará o no (Enyinna Nwankpa et al., 2018). En algunos casos también se representa un valor *bias* (b), este valor se suma al valor de activación, lo que permite desplazar el resultado para ajustarlo mejor al modelo (Collis, 2017).

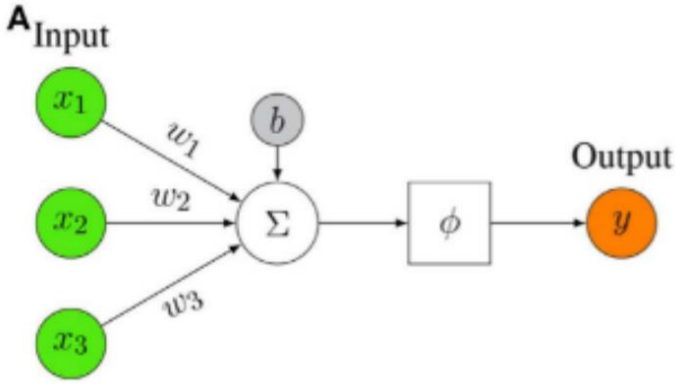


Figura 4: Representación de una neurona artificial (Emmert-Streib et al., 2020).

Las neuronas se agrupan en capas, estas capas se ensamblan de forma interconectada formando una ANN. La cantidad de interconexiones entre capas, o expresado de otra forma, la cantidad total de capas menos uno, se conoce como la profundidad de la ANN. Este valor denota la cantidad de transformaciones no lineales que aplica la red (Emmert-Streib et al., 2020). Por lo tanto, la capacidad computacional de la ANN está almacenada en las interconexiones que la constituyen (Gurney et al., 1997).

La estructura o arquitectura que forman las interconexiones entre neuronas es determinante para el funcionamiento de la ANN. Sin embargo, no existe un procedimiento establecido mediante el cual seleccionar la estructura idónea para cada aplicación, por lo que suele tratarse de una tarea de prueba y error, buscando la optimización de la estructura de la red. (Nowakowski et al., 2018).

Entrenamiento supervisado y validación

Como se ha mencionado previamente, la capacidad que tiene una ANN para realizar una tarea concreta reside en las interconexiones que se dan entre las neuronas que la forman, y estas interconexiones se modulan con los valores W y b. En un principio estos valores se inicializan con valores aleatorios, por lo que cabe esperar que el desempeño de la red será muy bajo (Chollet, 2017). Para mejorar el rendimiento de la red es necesario ajustar gradualmente estos valores mediante lo que se conoce como entrenamiento (Chollet, 2017).

Existen dos tipos de entrenamiento: el entrenamiento supervisado, apropiado para tareas de clasificación, y el no supervisado, indicado para tareas de agrupamiento (Zaharchuk et al., 2018). En el entrenamiento supervisado existe un resultado esperado que se utiliza para guiar la salida de la ANN durante el proceso de entrenamiento, mientras que en el no supervisado no existe ninguna expectativa de un resultado concreto, por lo que la red aprende por criterio propio a identificar grupos de patrones de datos (Zaharchuk et al., 2018). En este proyecto se ha desarrollado una herramienta de clasificación, por lo que se ha optado por un entrenamiento supervisado.

El proceso de entrenamiento supervisado consta de los siguientes pasos (Chollet, 2017):

1) Obtener un lote de muestras (X) y su resultado esperado (Y).
2) Aplicar la ANN sobre X para obtener un resultado real (Y^i).
3) Comparar Y con Y^i y calcular el error.
4) Modular los pesos de la red intentando reducir el error.

Estos pares de muestras y resultados esperados se almacenan en un conjunto de datos (Zaharchuk et al., 2018). Durante el entrenamiento se itera múltiples veces (denominadas épocas) sobre este conjunto de datos hasta que el rendimiento de la red converge (Zaharchuk et al., 2018). Además, el conjunto de datos se suele dividir en diferentes subconjuntos de entrenamiento, validación y test (Zaharchuk et al., 2018).

Evaluación y Sobreajuste

Una vez entrenada la ANN, es necesario evaluar el rendimiento de esta. El objetivo es obtener una red que se ajuste lo mejor posible al conjunto de entrenamiento y, que, a su vez, sea capaz de generalizar; es decir, que funcione correctamente con nuevos datos de entrada que nunca se le hayan mostrado (Chollet, 2017). Para poder valorar la capacidad de generalizar del sistema se utiliza el conjunto de test. Este conjunto no se muestra a la red durante el entrenamiento, por lo que las muestras que contiene son completamente nuevas para la ANN (Chollet, 2017). Al inicio del entrenamiento el rendimiento es bajo tanto en el conjunto de entrenamiento como en el de test, se dice que el modelo está subajustado (Chollet, 2017). A medida que se itera sobre el conjunto de entrenamiento, se puede observar cómo el rendimiento mejora en ambos conjuntos de datos. Sin embargo, a partir de cierto número de iteraciones el rendimiento del conjunto de test comienza a empeorar, dicho de otra manera, el sistema deja de generalizar (Chollet, 2017). Esto se debe a que la red comienza a memorizar patrones específicos del conjunto de entrenamiento que son desfavorables cuando se tratan nuevos datos. Este fenómeno se conoce como sobreajuste, y la mejor solución es proporcionar un conjunto de entrenamiento mayor y más variado durante el entrenamiento (Chollet, 2017).

HIPÓTESIS DE TRABAJO Y OBJETIVOS

En el caso concreto del proyecto en el que colaboro, los oncólogos del Hospital Universitario de Araba analizan muestras histológicas de colon teñidas con tinción hematoxilina-eosina (H&E), lo que les facilita la identificación de los tejidos y de las regiones afectadas por CRC. A continuación, introducen estas imágenes en QuPath (Bankhead et al., 2017), un software de patología digital, análisis y etiquetado de imágenes. Mediante este software examinan las muestras en busca de indicios de CRC y, en caso de encontrarlo, examinan que tejidos podrían estar afectados. Durante este proceso utilizan la herramienta de etiquetado que proporciona QuPath para marcar los tejidos presentes en la muestra y las regiones afectadas por el cáncer en caso de que las haya.

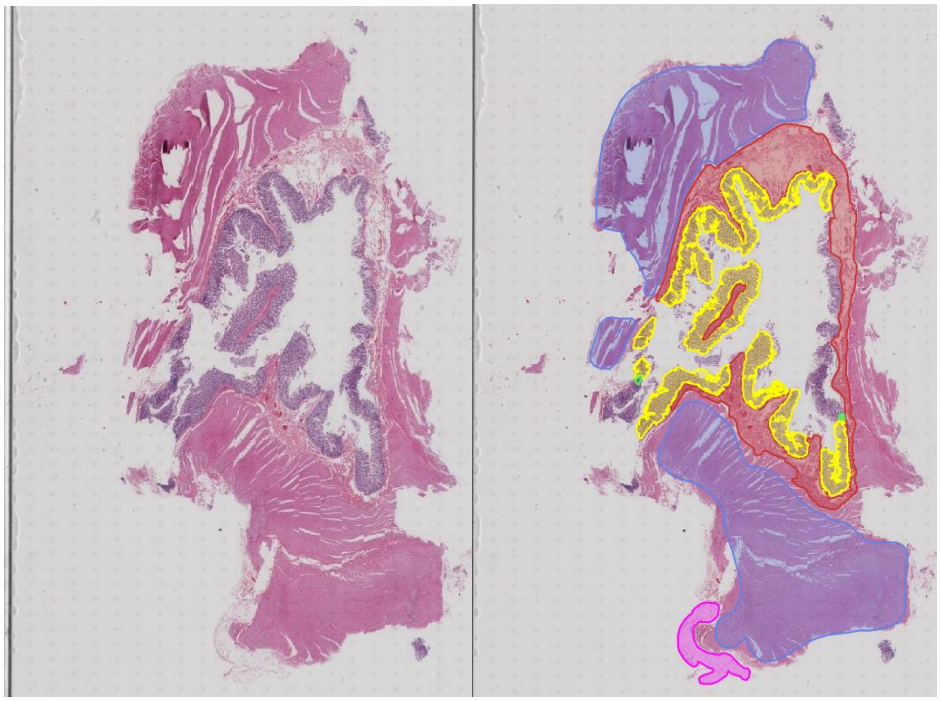


Figura 5 : Imagen histológica teñida con H&E antes y después de etiquetarla con QuPath.

Código cromático: Amarillo: Mucosa; Rojo: Submucosa; Azul: Muscular; Rosa: Subserosa.

Debido a que esta tarea requiere mucho tiempo, el grupo de oncólogos solicitó una herramienta computacional que agilice su trabajo y permita obtener un diagnóstico preliminar automatizado. De esta forma, podrán optimizar su tiempo y prestar mayor atención a los casos que representen un mayor riesgo. En concreto, el sistema deberá identificar regiones afectadas por cáncer en imágenes histológicas de

colon teñidas con H&E. Además, para los oncólogos podría resultar útil saber qué clase de tejido está siendo invadido, por lo que el sistema también deberá clasificar los tejidos presentes en la muestra.

Hipótesis de trabajo

Las redes neuronales artificiales son una herramienta adecuada para desarrollar un sistema de diagnóstico preliminar automatizado de adenocarcinoma de colon.

Objetivos

El objetivo principal de este proyecto es **desarrollar una red neuronal artificial que detecte la presencia de adenocarcinoma de colon e identifique los tejidos que han sido invadidos por éste**. El sistema recibirá imágenes histológicas teñidas con H&E, las procesará y mostrará las imágenes originales con máscaras de colores superpuestas indicando el tipo tejido y la región afectada por el cáncer.

Para desarrollar este proyecto el trabajo se ha dividido en las siguientes tareas:

1) Familiarizarse con el clúster computacional del IIS Biodonostia, así como con el software QuPath y desarrollar un script para exportar las imágenes anotadas con el formato adecuado.
2) Desarrollar una ANN que segmente los tejidos presentes en una imagen histológica teñida con H&E.
3) Desarrollar una ANN que identifique las regiones que puedan estar afectadas por CRC en una imagen histológica teñida con H&E.
4) Unificar ambos sistemas superponiendo sus resultados
5) Desarrollar una interfaz gráfica de usuario para facilitar el uso a usuarios finales sin amplios conocimientos de informática.

METODOLOGÍA

Determinar la tarea a desarrollar

Cuando me incorporé al proyecto me indicaron que los oncólogos necesitaban una herramienta que identificara el CRC y los tejidos afectados por éste. Para esta tarea se había propuesto utilizar la visión por computador mediante inteligencia artificial. Sin embargo, fue necesario establecer de forma más concreta la tarea que deberíamos llevar a cabo.

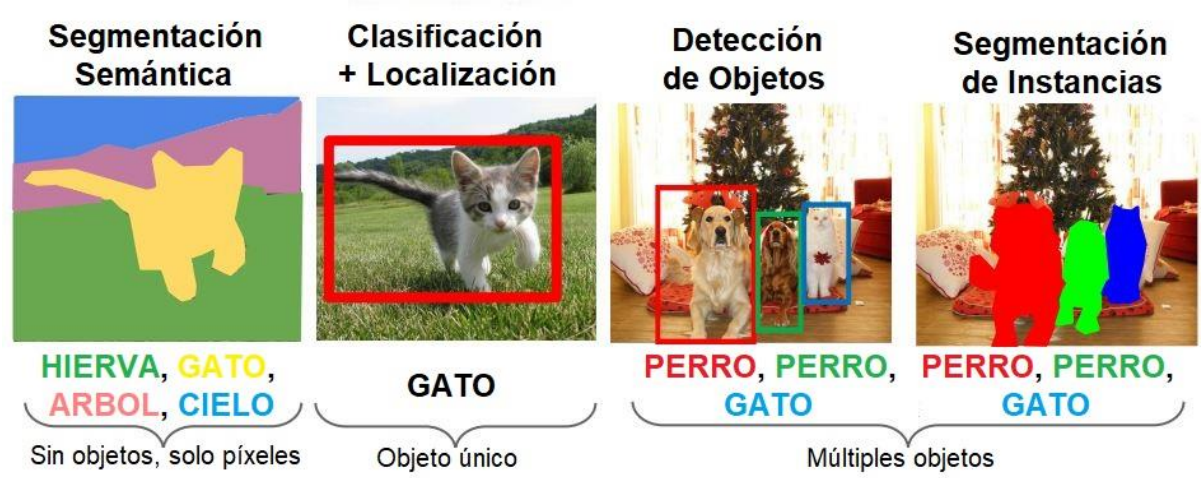


Figura 6: Tareas de visión por computador. Adaptado de (Jabalameli et al., 2018)

Existen cuatro variantes principales de visión por computador, las cuales se muestran en la *Figura 6*. En este caso no es suficiente con identificar la presencia del cáncer, sino que es necesario delimitarlo y conocer qué píxeles lo conforman para detectar que tejidos ha invadido. Por lo tanto, se trata de una tarea de segmentación. Tras consultar con los médicos responsables del proyecto, éstos señalaron que no es relevante para el diagnóstico identificar las posibles áreas afectadas como elementos individuales, por lo que se descartó la segmentación de instancias. Por consiguiente, se concluyó que habría que desarrollar un sistema de segmentación semántica. Así pues, se dedicaron las primeras dos semanas de prácticas a la formación en el desarrollo de esta clase de sistemas y en preparar el entorno informático con todas las herramientas necesarias.

Script para exportar imágenes

Los oncólogos proporcionaron las imágenes etiquetadas guardadas como proyecto de QuPath. Para utilizar dichas imágenes en nuestro flujo de análisis de datos, fue necesario crear un script para exportar las imágenes a un formato adecuado. Es necesario exportar cada imagen de dos formas. Por un lado, hay que exportar las imágenes sin el etiquetado de tejidos. Estas serán las imágenes de entrada de la red neuronal. Por otro lado, para generar la información de supervisión que se utilizara durante el entrenamiento, se exporta la imagen como una máscara de segmentación (véase *Figura 7*). Una máscara de segmentación consiste en generar una imagen en la que cada clase de la imagen (los tejidos etiquetados por los oncólogos en este caso) es representada mediante un número. Para obtener estos dos formatos a partir de las imágenes de QuPath se programó un script en el lenguaje Groovy; el idioma de programación soportado por QuPath. Este script se puede ejecutar sobre un proyecto en el que existen múltiples imágenes etiquetadas y exporta tanto la imagen sin etiquetar como la máscara de segmentación de cada una de estas, ambas en formato jpg.

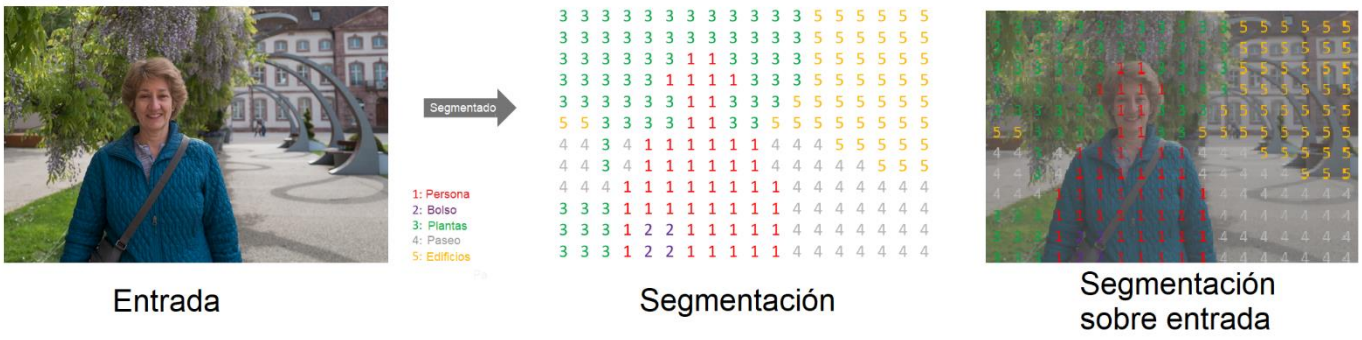


Figura 7: Máscara de segmentación. Adaptado de (An overview of semantic image segmentation, 2022)

Arquitectura de la red: U-net

Ninguno de los involucrados en el proyecto tenía experiencia previa en el diseño de redes neuronales. Por lo tanto, y debido a la complejidad que conlleva desarrollar y optimizar una arquitectura de ANN propia, optamos por utilizar una arquitectura predefinida. En concreto, hemos utilizado U-net. U-net es una arquitectura de ANN ampliamente adoptada en el ámbito biomédico (Siddique et al., 2021). Está específicamente diseñada para la segmentación de imágenes (Siddique

et al., 2021); es decir, para el particionado de imágenes (o video) en múltiples segmentos u objetos (Minaee et al., 2020).

U-Net es una red formada por capas completamente convolucionales. Como se puede ver en la *Figura 8*, la arquitectura de la red es simétrica y consta de un codificador que extrae características espaciales de la imagen y un decodificador que construye el mapa de segmentación a partir de las características codificadas (Ibtehaz et al., 2019). Ambas partes están constituidas por unos bloques con funciones opuestas que se repiten 4 veces. En el caso del codificador, los bloques contienen dos convoluciones ReLU (*Rectified Linear Unit* por sus siglas en inglés) que actúan como filtros, y una operación de *Max Pooling,* la cual disminuye la resolución de la imagen agrupando píxeles (Ibtehaz et al., 2019). En cambio, los bloques del decodificador están formados por 2 operaciones ReLU y una convolución transpuesta, que aumenta la resolución de la imagen (Ibtehaz et al., 2019). La principal característica de U-net reside en conectar cada uno de los bloques del codificador con los bloques a la misma altura del decodificador. Estas conexiones de salto permiten que la red recupere la información espacial perdida por las operaciones de agrupación (Ibtehaz et al., 2019).

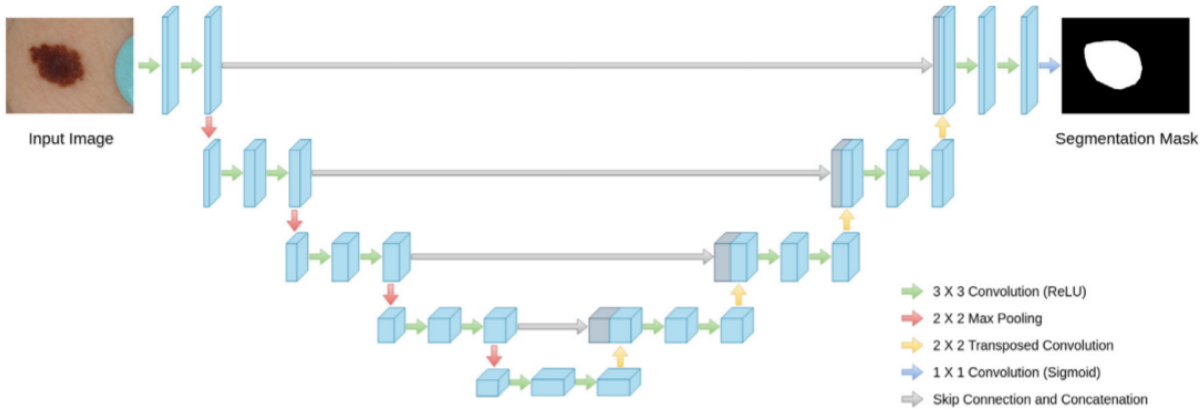


Figura 8: Arquitectura de U-Net, formado por un codificador, un decodificador y conexiones de salto. Adaptado de (Ibtehaz et al., 2019)

Estructura del sistema

Desarrollar una única ANN que clasifique los diferentes tejidos presentes en la muestra e identifique las regiones afectadas por el cáncer es una tarea especialmente

compleja. Si se tiene en cuenta que el sistema debe discernir entre 6 posibles clases (fondo, mucosa, linfocitos, submucosa, muscular y (o) subserosa) y, a su vez, cada una de estas clases puede clasificarse como sano o afectado, nos encontramos con que la red neuronal tendrá que clasificar cada píxel entre 12 posibles clases. Esta nos pareció la forma más apropiada de estructurar el sistema, ya que permitiría a la ANN aprender las características propias de cada tejido, tanto en las regiones sanas como en las afectadas por CRC. Esto es especialmente útil, ya que las células malignas presentan morfologías alteradas respecto a las sanas; principalmente, una mayor proporción de núcleo respecto a citoplasma e irregularidades en la membrana citoplasmática (Fischer, 2020).

Sin embargo, existen ciertas limitaciones técnicas que impiden trabajar de esta forma. En primer lugar, el grupo de oncólogos que proporciona las imágenes de entrenamiento no puede dedicar mucho tiempo a su preparación, por lo que se dispone de un conjunto de datos limitado para el entrenamiento. Como consecuencia, al crear una única ANN con un gran número de clases y una pequeña cantidad de ejemplos para el entrenamiento, nos encontraremos con que hay muy pocos ejemplos para cada clase especifica, por lo que rápidamente el sistema presentara problemas de sobreajuste. En segundo lugar, para los oncólogos es muy laborioso clasificar 12 clases en cada imagen, lo cual agrava más el problema del poco tiempo disponible. Un último condicionante está relacionado con el objetivo final del proyecto, consistente en proporcionar una herramienta accesible que los médicos puedan utilizar habitualmente. Ello implica que la ANN tenga un tamaño limitado, de manera que pueda ser ejecutada en ordenadores personales a disposición de los patólogos en los servicios sanitarios.

A la vista de estos condicionantes y tras estudiar diversas soluciones, se ha optado por diseñar un sistema dividido en dos ANN, tal y como se aprecia en la *Figura 9*. La primera red neuronal, a la que se ha denominado *Healthy Tissue Predictor* (HTP), se encarga de la clasificación de tejidos. La segunda red neuronal, a la que se ha denominado *Cancer Tissue Predictor* (CTP), se especializa en identificar las regiones afectadas por CRC. Una vez obtenidas ambas predicciones, simplemente hay que superponer ambos resultados para identificar los tejidos invadidos por el cáncer.

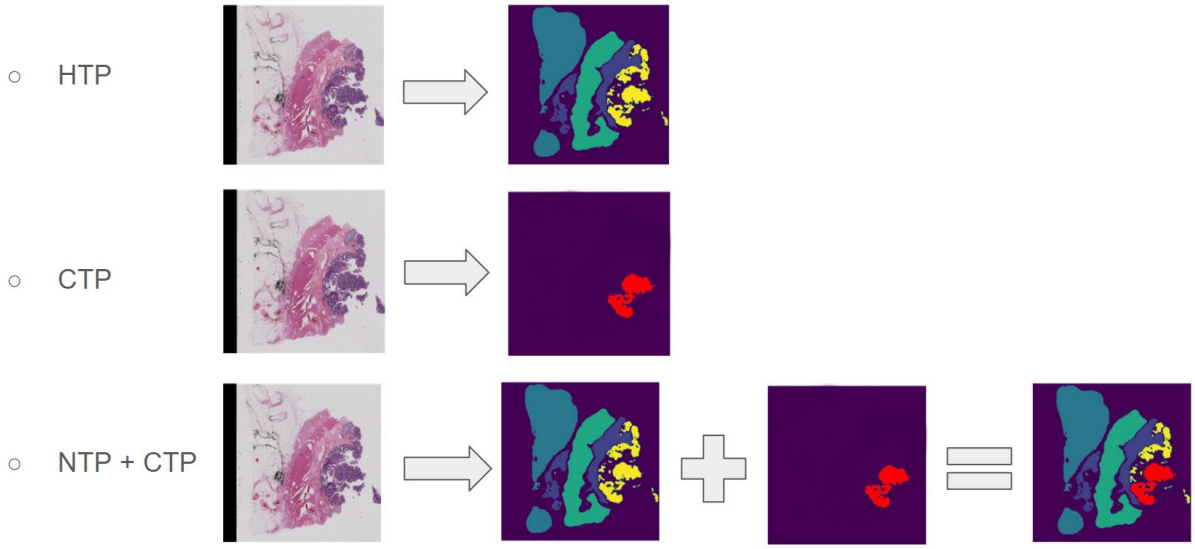


Figura 9: Estructura del sistema dividido en dos redes neuronales: Healthy Tissue Predictor (HTP) y Cancer Tissue Predictor (CTP).

Gracias a esta división del proceso se obtienen dos redes neuronales sencillas, funcionando cada una de ellas como una capa que contiene parte de la información necesaria y que, una vez superpuestas, proporcionan la misma información que el proceso original. Por otro lado, es menos laborioso (o más eficiente respecto al tiempo) clasificar en algunas imágenes 6 clases (para los tejidos) y en otras 2 clases (sano o enfermo), que clasificar 12 clases en cada imagen.

Por último, la segmentación de imágenes es una tarea en la que es necesario explorar un gran número de posibles configuraciones de los diversos parámetros hasta lograr optimizar el resultado. El hecho de dividir el sistema en dos subsistemas aporta el beneficio añadido de permitir modular esas combinaciones de forma independiente, posibilitando observar el efecto de cada modificación en cada sistema concreto y facilitando, con ello, la optimización de cada uno por separado.

Reescalado

Las imágenes de entrenamiento que proporcionan los oncólogos son archivos de entre 1,5 GB y 3 GB e, incluso, mayores; con resoluciones del orden de 10.000 x 10.000 píxeles. La elevada resolución de las imágenes de entrada afecta al tamaño

de la ANN, y por tanto a la memoria RAM requerida para su almacenamiento y ejecución. En vista de que el sistema debe poder ejecutarse en los ordenadores personales disponibles en el hospital, se llegó a la conclusión de que no era viable trabajar con resoluciones tan altas. Sin embargo, cada píxel de una imagen es una unidad de información que será modulada a través de las diversas capas de la ANN e influirá en el resultado obtenido. Por lo tanto, trabajar con resoluciones menores implica trabajar con menos unidades de información, y, por ende, con menos datos.

Debido a esto, se seleccionaron dos resoluciones menores con las que trabajar: en la primera se aplicó un reescalado mediante el cual se disminuye la resolución 10 veces (x0,1), y en la segunda se disminuye 40 veces (x0,025). Se optó por trabajar con dos resoluciones diferentes porque ello permitiría conocer cómo afecta la disminución de la resolución a la calidad de la predicción, y de esta manera, ser conscientes de los compromisos que se asumen a medida que aumenta dicha perdida de información.

Se ha encontrado que la mayor resolución (x0,1) tiene mayor rendimiento al identificar linfocitos, los cuales son difícilmente apreciables a simple vista, mientras que la resolución menor (x0,025) tiene un mejor rendimiento identificando la subserosa, un tejido que se presenta en forma de regiones amplias, difusas y poco distinguibles del fondo. Por lo tanto, se ha llegado a la conclusión de que las resoluciones menores favorecen la identificación de regiones amplias y poco definidas, en las que el contexto es muy relevante y es necesario una visión global de la imagen; mientras que las resoluciones más altas ayudan a discriminar estructuras de menor tamaño gracias al mayor detalle en la muestra, pero repercuten negativamente en la visión global.

Aumentado de datos y transferencia de aprendizaje

Como se ha mencionado previamente, al entrenar una red neuronal lo habitual y aconsejable es utilizar un gran número de imágenes de entrenamiento. Sin embargo, en este caso se cuenta con un pequeño número de muestras. Por lo tanto, se decidió aplicar ciertas técnicas de aumentado de datos que permiten crear una base de datos más extensa a partir de la original. En primer lugar, se aplicaron transformaciones a las imágenes, como voltearlas vertical y horizontalmente. La

cantidad o calidad de la nueva información que estas imágenes creadas mediante transformaciones aportan a la ANN no es equiparable a la aportada por imágenes genuinas. Sin embargo, estas transformaciones aportan variabilidad a la base de datos, por lo que pueden ayudar a que el modelo resultante generalice mejor.

En segundo lugar, se realizó una transferencia de aprendizaje. Esta técnica consiste en utilizar una red preentrenada con una gran base de datos muy general y especializarla para llevar a cabo un cierto objetivo deseado. En un entrenamiento desde cero se inicializarían los pesos de las neuronas con valores aleatorios. Sin embargo, al aplicar transferencia de aprendizaje, se utilizan como punto de partida valores previamente obtenidos por un entrenamiento general y que están disponibles de forma pública. En este proyecto se han empleado los valores de Imagenet, una red preentrenada con más de 14 millones de imágenes etiquetadas manualmente, disponible en la librería segmentation_models (Yakubovskiy, 2019). La red finalmente fue entrenada con la base de datos propia para que se especializara en la tarea deseada. Esta opción solo ha sido posible gracias a la utilización de una red neuronal con una estructura predefinida como lo es Unet, ya que de otra manera sería imposible disponer de los valores necesarios para inicializar los pesos de las neuronas.

En tercer lugar, se optó por utilizar un muestreo de baldosas. Esta técnica permitió generar miles de imágenes a partir del pequeño conjunto de datos disponible.

Muestreo de baldosas

Uno de los principales inconvenientes que nos encontramos en el desarrollo del sistema es que cada imagen del conjunto de datos tiene una relación de aspecto (ratio entre ancho y altura) y resolución diferentes. Esto supone un problema debido a que, al definir una ANN, se especifica la resolución de las imágenes que se introducirán. Además, las imágenes siempre deben ser cuadradas. Si todas las imágenes tuvieran una relación de aspecto equivalente, se podría optar por deformar las imágenes para transformarlas en cuadradas; ya que los cambios en la morfología celular y tisular serian equivalentes en todas las muestras. Sin embargo, en el caso de estudio esta opción no es válida puesto que, al tener cada imagen una relación de

aspecto diferente, en algunos casos la imagen se comprimiría verticalmente, mientras que en otros casos se comprimiría horizontalmente.

Como solución a estos problemas se optó por el muestreo de baldosas. Esta técnica consiste en preprocesar el conjunto de datos para dividir cada imagen en pequeñas baldosas de un tamaño preestablecido y con la superposición deseada; tal y como se aprecia en la *Figura 10*. Gracias a esta técnica se obtienen múltiples beneficios. Por una parte, se producen imágenes cuadradas a partir de cualquier muestra sin necesidad de aplicarle ninguna transformación, independientemente de la relación de aspecto que esta tuviera originalmente. Por otra parte, el tamaño de las imágenes y, en consecuencia, el tamaño de la red neuronal, disminuyen considerablemente sin necesidad de mermar la calidad de imagen y, por lo tanto, sin que se produzca una pérdida de detalle o definición. Finalmente, esta técnica permite generar un gran conjunto de datos a partir de uno mucho menor. Por ejemplo, si se parte de una imagen de 1.000 x 1.000 píxeles a la que se le hace un muestreo con baldosas de 100 x 100 píxeles y una superposición de baldosas del 50%. Como resultado, se obtienen varios cientos de imágenes. Esto ha sido de gran utilidad en el presente caso, ya que se dispone de un conjunto de datos muy limitado (alrededor de 150 imágenes). Muchas de las baldosas resultantes únicamente incluían fondo, por lo que se descartaron las imágenes que no contuvieran un mínimo del 5% clasificado como tejido.

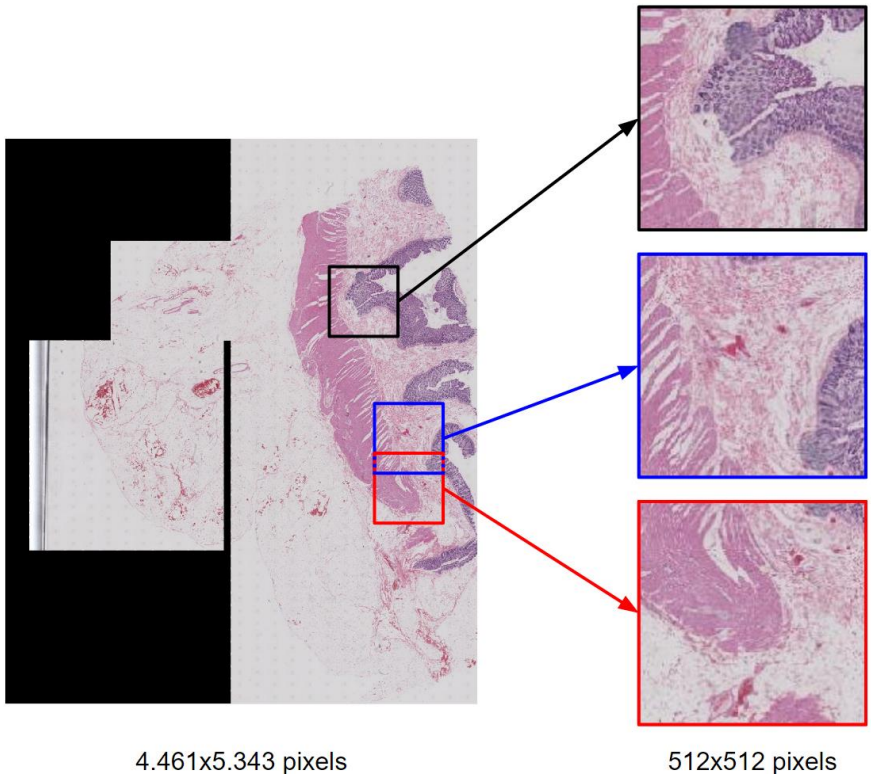


Figura 10: Muestreo de baldosas.

Sin embargo, al aplicar el muestreo de baldosas se modifica considerablemente la tarea que lleva a cabo la red neuronal, la cual aprende a segmentar estas pequeñas baldosas en lugar de las muestras completas que introducirán los oncólogos. Para solucionar este problema se utilizó una librería que permite dividir la imagen original en baldosas, hacer las predicciones por separado y, finalmente, volver a unir las baldosas utilizando las partes superpuestas para calcular medias con las que suavizar las uniones (Bhattiprolu, 2021).

Interfaz gráfica

La finalidad del proyecto es obtener un sistema robusto de inteligencia artificial que facilite el trabajo a los médicos. Estos profesionales, en principio, no tienen por qué contar con los conocimientos informáticos necesarios para trabajar de forma ágil a través de un terminal informático. Por lo tanto, se propuso desarrollar una interfaz gráfica simple que facilite la tarea de cargar, procesar y guardar las imágenes a toda clase de usuarios. Debido a la escasa formación recibida sobre el desarrollo de

interfaces gráficas y a que el software no persigue ningún objetivo comercial, se pretende desarrollar una interfaz sencilla, tanto en su diseño como en su utilización, y sin grandes pretensiones estéticas.

De hecho, debido a que el desarrollo de la interfaz gráfica no aporta ningún valor científico o académico al proyecto, se ha tratado como un objetivo secundario al que se le han dedicado los periodos de tiempo disponibles mientras no se ha podido avanzar con las tareas prioritarias del proyecto.

Diagrama de Gantt

En la *Figura 11* se muestra un gráfico de Gantt que representa la distribución temporal del periodo de desarrollo del proyecto durante mis prácticas curriculares en Biodonostia.

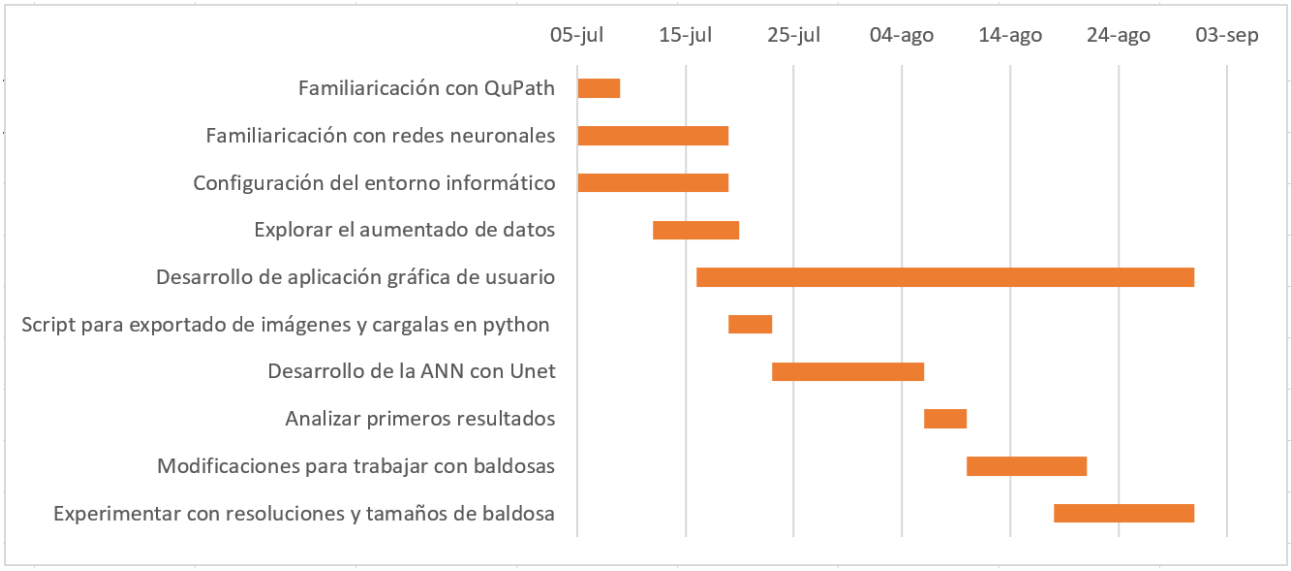


Figura 11: Diagrama de Gantt ilustrando la distribución temporal del desarrollo

RESULTADOS

En primer lugar, se ha desarrollado un sistema de aprendizaje automático totalmente capaz de clasificar los tejidos presentes en imágenes histológicas de colon. Además de la valoración visual, para evaluar los resultados se ha utilizado la fórmula de intersección sobre unión, que consiste en calcular la intersección del conjunto de píxeles del resultado esperado y del resultado obtenido y dividirlo por la unión de estos mismos conjuntos. En base a este cálculo, se ha obtenido una tasa de hasta el 0.85 de píxeles acertados. Aunque esta no parezca una tasa extremadamente alta, visualmente se puede comprobar que el resultado es muy similar al etiquetado por los oncólogos. (véase *Figura 12*).

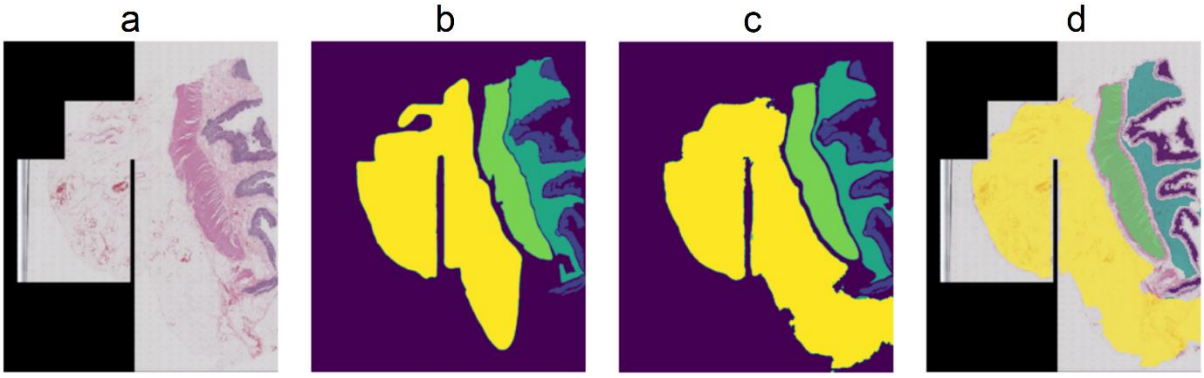


Figura 12: Resultado de la predicción de HTP en una muestra sana. a: imagen original de tejido sano; b: imagen etiquetada por los oncólogos; c: predicción realizada por la ANN; d: predicción superpuesta a la imagen original.

Código cromático: Amarillo: subserosa; Verde: muscular; Azul claro: submucosa; Azul oscuro: mucosa; Morado: fondo.

El sistema también ha sido capaz de clasificar regiones que los oncologos no habían etiquetado debido a que no estaban completamente seguros de a qué tipo de tejido correspondían, tal y como se observa en la parte inferior derecha de la *Figura 12*. Tras haber estudiado las imágenes con más detenimiento, los médicos han corroborado que estas nuevas regiones han sido clasificadas correctamente por la ANN. Aunque la imagen representada suponga un ejemplo sencillo, este resultado es un indicativo de que, potencialmente, los sistema de redes neuronales podrían tener la capacidad de superar el rendimiento de un especialista, al menos en un análisis preliminar. En cualquier caso, se ha de tener presente que la fiabilidad del resultado

no es equiparable al criterio de un oncólogo que ha examinado la muestra detenidamente.

Por otra parte, el ensayo sistemático con distintas resoluciones y tamaños de baldosa ha permitido encontrar un punto medio en el que la red neuronal es capaz de reconocer las regiones correspondientes a los linfocitos sin sacrificar por ello la calidad de la predicción de las regiones más grandes. En la *Figura 13* se puede observar la misma imagen mostrada en la *Figura 12*, pero con mayor aumento para apreciar como dos de las tres regiones de linfocitos etiquetadas por los médicos han sido clasificadas correntamente por la red neuronal.

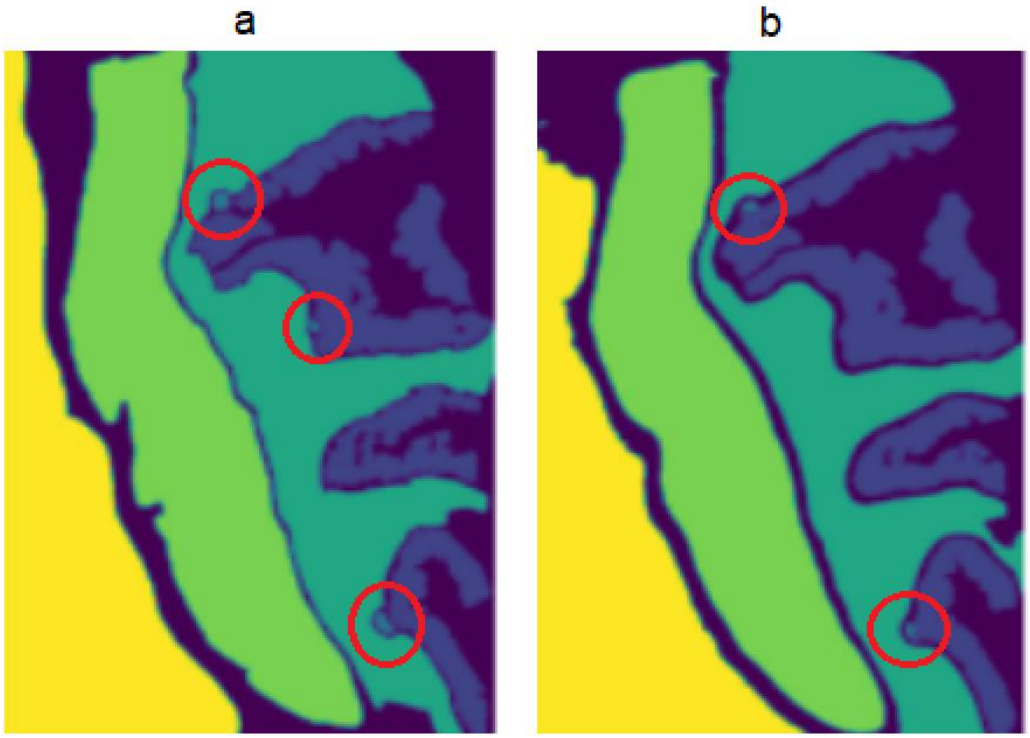


Figura 13: Resultado de la predicción de HTP en una muestra sana. a: imagen etiquetada por los oncólogos; b: predicción realizada por la ANN.

Código cromático: Amarillo: subserosa; Verde: muscular; Azul claro: submucosa; Azul oscuro: mucosa; Morado: fondo; círculos rojos: linfocitos.

En cuanto a la segunda fase, cuyo objetivo es desarrollar una segunda red neuronal que identificara las regiones afectadas por el cáncer colorrectal, no ha sido posible obtener resultados todavía. Debido a que durante el corto periodo de tiempo de prácticas, a los médicos sólo les fue posible etiquetar cinco muestras con tumor; cantidad con la que resulta inviable llevar a cabo el entrenamiento. Por esta razón,

mientras los patólogos recolectaban más imágenes de CRC y las etiquetaban, se decidió continuar con la siguiente fase: preparación del sistema para superponer el resultado de ambas redes neuronales y, de esta forma, poder identificar los tejidos afectados por el tumor.

Se utilizaron las cinco imágenes con tumor disponibles para llevar a cabo esta tarea. En primer lugar, se obtuvo la predicción de la red HTP de dichas imágenes, tal y como se muestra en la *Figura 14 b*. A continuación, y debido a que la red CTP todavía no era funcional, se utilizó el etiquetado realizado a mano por los médicos como si fuera la predicción de la red neuronal (*Figura 14 c)*. Finalmente, se superpusieron las dos máscaras de segmentación a la imagen original con cierto grado de transparencia (*Figura 14 d)*.

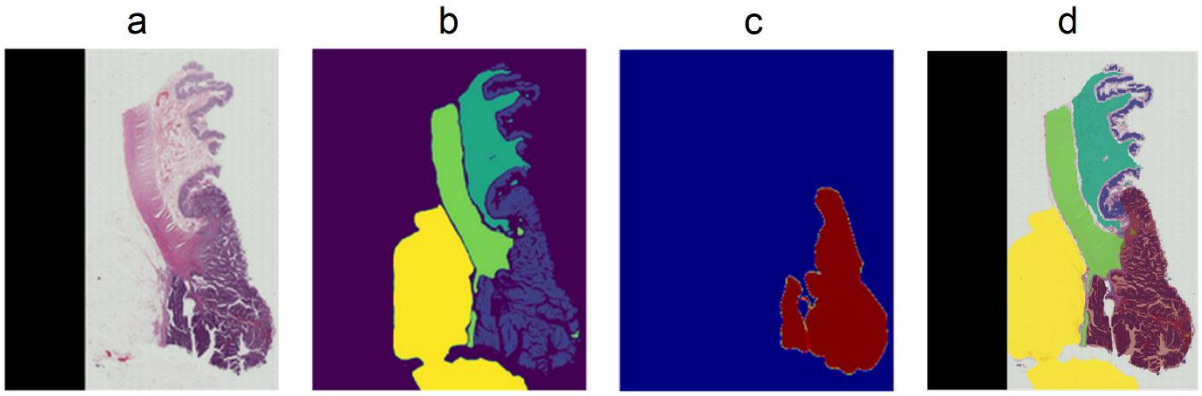


Figura 14: Resultado de la predicción de HTP en una muestra con tumor. a: imagen original con tumor; b: predicción de tejidos realizada por la ANN; c: tumor etiquetado por los oncólogos; d: predicción de tejidos y etiquetado del tumor superpuestos a la imagen original.

Como se observa en la *Figura 14*, la clasificación de los tejidos sigue siendo muy correcta en la mayoría de la imagen. Sin embargo, todo el tejido afectado por el tumor ha sido clasificado como mucosa (azul oscuro). Tras analizar detenidamente estos resultados se llegó a dos posibles razones. La primera hipótesis es que el color resultante de la tinción H&E es muy similar en el tumor y en el tejido perteneciente a la mucosa. La segunda hipótesis es que, tal y como esperábamos, los tejidos afectados por el tumor sufren grandes cambios morfológicos. Sin embargo, esperábamos que las modificaciones sufridas no fueran tan notorias como para provocar la pérdida total de características visuales de los tejidos. En otras palabras, la red neuronal extrae toda la información a partir de las características visuales de la

imagen y se basa en estas características extraídas para clasificar cada píxel como una clase concreta. Por lo tanto, es posible que el motivo del problema se deba a una conjunción de ambas razones expuestas anteriormente, ya que los dos casos afectan a las características visuales de los tejidos y contribuyen simultáneamente a dificultar la detección del tejido.

Hay que subrayar que la clasificación de tejidos ha funcionado correctamente en las zonas limítrofes del tumor. Esto permite distinguir a simple vista cuáles son los tejidos que pueden haber sido afectados por el CRC. Por lo tanto, y aunque por el momento el funcionamiento del sistema no es tan preciso como se esperaba en un principio, si podemos afirmar que cumple con la función de ayudar a distinguir de forma extremadamente sencilla qué tejidos están en contacto con la región afectada. Además, sigue siendo viable el objetivo buscado; esto es, que el sistema señale de forma automática cuáles son los tejidos afectados. Inicialmente se propuso comparar las máscaras de segmentación resultantes para comprobar a qué tejido pertenecen los píxeles clasificados como tumor y, así, presentar una lista de los tejidos afectados. Esta tarea ha funcionado correctamente, ya que en las zonas más limítrofes sí se han hallado algunos píxeles de tumor superpuestos a los píxeles del tejido correcto. Aun así, con el fin de optimizar el funcionamiento del sistema, será necesario elaborar un algoritmo más complejo, de manera que, en lugar de confiar en que se produzca esa pequeña superposición de píxeles, tenga, per se, la capacidad de realizar la clasificación teniendo en cuenta la composición de las regiones vecinas.

DISCUSIÓN

Existen diversos estudios previos que han abordado el diagnóstico de CRC mediante el enfoque de las redes neuronales. De todos ellos, el trabajo realizado por Kather (Kather et al., 2019) resulta de especial interés dado que, en dicho estudio, utilizaron una red neuronal para la prognosis de CRC. Su planteamiento también se basó en clasificar los tejidos presentes en las muestras y la región afectada por el tumor, pero distinguieron un mayor número de tejidos y trataron el tumor como un tejido en sí mismo, sin pretender identificar a qué clase pertenecía este tejido antes de ser afectado por el carcinoma. Asimismo, también aplicaron transferencia de aprendizaje y muestreo de baldosas, consiguiendo alcanzar una tasa de acierto del 99% (calculado de forma diferente a la intersección sobre la unión utilizada en este trabajo). Esta elevada tasa de acierto refuerza la idea de que estas técnicas que hemos utilizado resultan una adecuada solución a la escasa disponibilidad de datos, además de una forma conveniente de generar imágenes con relación de aspecto cuadrada. En cuanto a la disponibilidad de datos, en el estudio de Kather únicamente disponían de 86 imágenes etiquetadas, frente a las más de 150 que se nos han proporcionado para realizar nuestro estudio. Por lo que la alta disponibilidad de datos de gran calidad resulta un aporte especialmente destacable de nuestro trabajo.

Por otro lado, en el estudio de Ben Hamida, A. (ben Hamida et al., 2021) se proponen varias soluciones a la detección de tumores mediante DL. Por un lado, utilizan la técnica de muestreo de baldosas. Con estas baldosas toman dos caminos, uno consistente en llevar a cabo una segmentación semántica píxel a píxel; y otra, consistente en clasificar la baldosa de forma global en una clase de tejido. Concluyen que la mejor alternativa es la clasificación píxel a píxel, ya que la segunda no funciona correctamente cuando hay más de un tipo de tejido presente en una sola imagen. Por otro lado, efectúan pruebas con Unet y Segnet, otra estructura de CNN similar a Unet. Además, en ambas estructuras prueban a entrenar el modelo desde cero y a utilizar un modelo preentrenado. Los resultados muestran que Segnet no preentrenado es la opción con mayor rendimiento, especialmente respecto a los falsos positivos. En cuanto al etiquetado de las imágenes, en este artículo también tratan el tumor como un tejido más. Además, hacen hincapié en la dificultad y el coste temporal del etiquetado manual de imágenes y la poca disponibilidad de estas.

En lo que se refiere a las perspectivas de futuro, y basándonos en los estudios previamente mencionados, habría que considerar los siguientes puntos:

1. La distribución de clases. Ambos estudios coinciden en tratar al tumor como un tejido más, en lugar de pretender identificar el tejido que ha sido afectado por el carcinoma. Esto posiblemente sea debido a que, tal y como nos hemos encontrado, los cambios morfológicos que el cáncer provoca en las células y tejidos son tales, que imposibilita la adecuada clasificación de estos.
2. Aunque elegimos utilizar Unet por ser una red neuronal ampliamente establecida para tratamiento de imágenes biomédicas, habría que considerar otras estructuras neuronales como Segnet y Resnet.
3. Los parámetros clave en el desarrollo han sido la resolución y el aumento de las baldosas utilizadas. Optimizar estos valores para obtener baldosas que permitan visualizar las características clave de cada tipo de tejido podría favorecer su identificación.
4. Como ya se ha mencionado repetidas veces, el mayor factor limitante al tratar con imágenes biomédicas suele ser la escasez de muestras disponibles para el entrenamiento. Por esta razón, y aunque ya contamos con una mayor base de datos que la mayoría de los estudios al respecto, sería vital seguir aumentando el conjunto de imágenes de entrenamiento para optimizar el funcionamiento de la red neuronal. A tal fin, los patólogos con los que se ha colaborado durante este proyecto continúan trabajando para aportar más imágenes etiquetadas que permitan seguir entrenando el modelo de CNN y maximizar su rendimiento.

CONCLUSIONES

El cáncer colorrectal es una de las clases de cáncer más diagnosticadas, además de tratarse de una de las que más muertes causa anualmente. El mayor factor de riesgo asociado a esta patología es la dieta con excesivo consumo de carnes rojas, habitual tanto en los países desarrollados como en aquellos en vías de desarrollo. Además, la gran mayoría de los casos no presentan ningún historial familiar asociado a este tipo de cáncer.

El diagnóstico temprano es determinante en la esperanza de vida de los pacientes que sufren CRC. Por lo tanto, es necesario contar con herramientas sencillas de utilizar y poco invasivas que agilicen el análisis de muestras y, así, poder realizar screening frecuentes a la población. En este aspecto, tomar una muestra de tejido de colon es un procedimiento considerablemente invasivo. Sin embargo, es necesario llevar a cabo una confirmación mediante este tipo de análisis cuando los métodos de cribado menos invasivos indican un posible caso de cáncer. Por lo tanto, resulta ciertamente conveniente proporcionar una herramienta que agilice el trabajo de los oncólogos.

En este proyecto se ha comenzado el desarrollo de una herramienta informática que efectúa el análisis de muestras histológicas de adenocarcinoma de color. En primer lugar, identificando el cáncer y delimitando el área afectada por éste y, en segundo lugar, segmentando los tejidos presentes en la muestra, lo que permite informar de forma automatizada sobre los tejidos afectados por el CRC. Para el desarrollo de dicha herramienta se han combinado múltiples técnicas de tratamiento de imágenes y métodos para optimizar redes neuronales a partir de una base de datos limitada, obteniendo resultados satisfactorios. Adicionalmente, este proyecto cuenta con una de las mayores bases de datos de imágenes de cáncer etiquetadas por especialistas.

Como conclusiones derivadas del objetivo principal del proyecto podemos señalar que:

- Se ha logrado desarrollar una red neuronal que segmente los tejidos presentes en muestras de tejido de colon teñidas con H&E con un grado de acierto satisfactorio.

- Se ha programado una red neuronal que identifica las regiones que están afectadas por adenocarcinoma de colon en muestras de tejido de colon teñidas con H&E. Sin embargo, debido a la falta de imágenes de adenocarcinoma etiquetadas para entrenar esta red, los resultados todavía no son suficientemente precisos.

- Se ha logrado superponer ambos resultados de forma que el sistema muestra una sola imagen en la que se expone toda la información obtenida de ambas redes neuronales, a la vez que se informa al usuario automáticamente de los tejidos que han sido afectados por el carcinoma.

- En cuanto al desarrollo de la interfaz gráfica, se ha invertido únicamente el tiempo en el que no se ha podido avanzar con el resto de tareas. Esto es debido a que no aporta ninguna funcionalidad al sistema más allá de facilitar su utilización, por lo que no será necesaria hasta el final del proyecto. Durante el presente trabajo no se ha profundizado en este aspecto del proyecto, por considerar que carece de cualquier interés biotecnológico.

AUTOEVALUACIÓN

En conclusión, mi contribución en el Instituto Biodonostia ha sido una experiencia muy gratificante y satisfactoria, además de haber sido un excelente primer contacto con el mundo laboral. Asimismo, he adquirido unos conocimientos y experiencias extremadamente enriquecedoras y que sin ninguna duda me resultaran realmente útiles en el futuro. Durante este proyecto he podido poner en práctica los conocimientos obtenidos durante los cinco años de formación en Biotecnología e Ingeniería Informática. Ha sido especialmente gratificante el haber tenido la oportunidad de combinar habilidades de ambas disciplinas para plantear posibles soluciones a los problemas que hemos encontrado durante el proyecto.

Creo que el resultado obtenido ha sido altamente satisfactorio teniendo en cuenta mi falta de experiencia en este campo. Sin embargo, como se ha mencionado en la discusión, el proyecto tiene potencial de mejora ya que existen ciertos aspectos en los que se debería profundizar para intentar optimizar el rendimiento del sistema.

En cuanto a lo aprendido, he tenido la oportunidad de adquirir nuevos conocimientos y de consolidar otras destrezas relacionadas con el campo de la informática que se habían abordado someramente durante los estudios de Ingeniería Informática; como son el tratamiento y preprocesado de imágenes, familiarizándome con las diferentes maneras existentes de codificarlas; la computación masiva accediendo de forma remota a servidores con gran capacidad computacional; o el desarrollo de redes neuronales, las cuales son vagamente estudiadas durante la carrera, pero sin la posibilidad de llevar a cabo aplicaciones prácticas. Por supuesto, también he podido asimilar nuevos y enriquecedores conocimientos sobre las características del cáncer de colon y el análisis de muestras histológicas para su detección precoz.

REFERENCIAS

American Cancer Society. (2021). *Colorectal Cancer Early Detection, Diagnosis, and Staging*. Retrieved from www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html

American Cancer Society. (2021). *Understanding Your Pathology Report: Invasive Adenocarcinoma of the Colon*. Retrieved from https://www.cancer.org/treatment/understanding-your-diagnosis/tests/understanding-your-pathology-report/colon-pathology/invasive-adenocarcinoma-of-the-colon.html

American Cancer Society. (2022). *Colorectal Cancer Facts & Figures 2020-2022*. Retrieved from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2020-2022.pdf

An overview of semantic image segmentation. (2022). Retrieved from https://www.jeremyjordan.me/semantic-segmentation/

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1). doi: 10.1038/S41598-017-17204-5

ben Hamida, A., Devanne, M., Weber, J., Truntzer, C., Derangère, V., Ghiringhelli, F., Forestier, G., & Wemmert, C. (2021). Deep learning for colon cancer histopathological images analysis. *Computers in Biology and Medicine*, *136*, 104730. doi: 10.1016/J.COMPBIOMED.2021.104730

Bhattiprolu, S. (2021). *python_for_microscopists/smooth_tiled_predictions.py at master · bnsreenu/python_for_microscopists*. Retrieved from https://github.com/bnsreenu/python_for_microscopists/blob/master/229_smooth_predictions_by_blending_patches/smooth_tiled_predictions.py

Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *The Lancet*, *383*(9927), 1490–1502. doi: 10.1016/S0140-6736(13)61649-9

Cancer Today. (2021). Retrieved from https://gco.iarc.fr/today/home

Chollet, F. (2017). *DEEP LEARNING with Python*.

Collis, J. (2022). *Glossary of Deep Learning: Bias*. Retrieved from https://medium.com/deeper-learning/glossary-of-deep-learning-bias-cf49d9c895e2

Edge, S. B., & Compton, C. C. (2010). The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology*, *17*(6), 1471–1474. doi: 10.1245/S10434-010-0985-4

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *3*, 4. doi: 10.3389/FRAI.2020.00004/BIBTEX

Enyinna Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). *Activation Functions: Comparison of Trends in Practice and Research for Deep Learning*.

Fischer, E. G. (2020). Nuclear Morphology and the Biology of Cancer Cells Keywords Nuclear membrane irregularity · Cancer · Nuclear envelope · Signal transduction · Chromatin. *Acta Cytologica*, *64*, 511–519. doi: 10.1159/000508780

Gurney, K., & York, N. (1997). *An introduction to neural networks*.

Ibtehaz, N., & Rahman, M. S. (2019). MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. *Neural Networks*, *121*, 74–87. doi: 10.1016/j.neunet.2019.08.025

Jabalameli, A., Ettehadi, N., & Behal, A. (2018). *Edge-Based Recognition of Novel Objects for Robotic Grasping*. doi: 10.3390/robotics8030063

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, *16*(1). doi: 10.1371/JOURNAL.PMED.1002730

Keum, N. N., & Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology and Hepatology*, *16*(12), 713–732. doi: 10.1038/s41575-019-0189-8

Labianca, R., Beretta, G. D., Kildani, B., Milesi, L., Merlin, F., Mosconi, S., Pessi, M. A., Prochilo, T., Quadri, A., Gatta, G., de Braud, F., & Wils, J. (2010). Colon cancer. *Critical Reviews in Oncology/Hematology*, *74*(2), 106–133. doi: 10.1016/J.CRITREVONC.2010.01.010

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2020). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2021.3059968

Munro, M. J., Wickremesekera, S. K., Peng, L., Tan, S. T., & Itinteang, T. (2018). Cancer stem cells in colorectal cancer: a review. *Journal of Clinical Pathology*, *71*(2), 110–116. doi: 10.1136/JCLINPATH-2017-204739

National Institutes of Health. (2021). *Exámenes para detectar el cáncer colorrectal y los pólipos*. Retrieved from https://www.cancer.gov/espanol/tipos/colorrectal/hoja-informativa-deteccion

National Institutes of Health. (2021). *MedlinePlus: Pólipos colorrectales*. Retrieved from https://medlineplus.gov/spanish/ency/article/000266.htm

Neo, J. H., Ager, E. I., Angus, P. W., Zhu, J., Herath, C. B., & Christophi, C. (2010). Changes in the renin angiotensin system during the development of colorectal cancer liver metastases. *BMC Cancer*, *10*. doi: 10.1186/1471-2407-10-134

Nowakowski, G., Dorogyy, Y., & Doroga-Ivaniuk, O. (2018). Neural network structure optimization algorithm. *Journal of Automation, Mobile Robotics and Intelligent Systems*, *12*(1), 5–13. doi: 10.14313/JAMRIS_1-2018/1

OMS Cancer. (2021). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cancer

Rawla, P., Sunkara, T., & Barsouk, A. (2019). Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad Gastroenterologiczny*, *14*(2), 89–103. doi: 10.5114/PG.2018.81072

Sagaert, X., Vanstapel, A., & Verbeek, S. (2018). Tumor Heterogeneity in Colorectal Cancer: What Do We Know So Far? *Pathobiology : Journal of Immunopathology, Molecular and Cellular Biology*, *85*(1–2), 72–84. doi: 10.1159/000486721

Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*. doi: 10.1109/ACCESS.2021.3086020

Simon, K. (2016). Colorectal cancer development and advances in screening. *Clinical Interventions in Aging*, *11*, 967–976. doi: 10.2147/CIA.S109285

Song, M., Garrett, W. S., & Chan, A. T. (2015). Nutrients, foods, and colorectal cancer prevention. *Gastroenterology*, *148*(6), 1244-1260.e16. doi: 10.1053/J.GASTRO.2014.12.035

Wen, B., Zeng, W. F., Liao, Y., Shi, Z., Savage, S. R., Jiang, W., & Zhang, B. (2020). Deep Learning in Proteomics. *Proteomics*, *20*(21–22). doi: 10.1002/PMIC.201900335

Yakubovskiy, P. (2019). qubvel/segmentation_models. *GitHub*. Retrieved from https://github.com/qubvel/segmentation_models

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., & Langlotz, C. P. (2018). Deep Learning in Neuroradiology. *AJNR. American Journal of Neuroradiology*, *39*(10), 1776–1784. doi: 10.3174/AJNR.A5543