UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

CLASIFICACIÓN DE IMÁGENES DE CÁNCER GÁSTRICO APLICANDO APRENDIZAJE PROFUNDO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

LUIS FELIPE ESCARES GARAY

PROFESOR GUÍA:
MAURICIO CERDA VILLABLANCA

MIEMBROS DE LA COMISIÓN:
FRANCISCO RIVERA SERRANO
JUAN BARRIOS NÚÑEZ

Este trabajo ha sido parcialmente financiado por FONDECYT 1221696.
Esta investigación fue apoyada por el supercomputador Patagón
de la Universidad Austral de Chile (FONDEQUIP EQM180042).

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE MAGÍSTER EN CIENCIAS
DE LA INGENIERÍA
POR: LUIS FELIPE ESCARES GARAY
FECHA: 2022
PROF. GUÍA: MAURICIO DAVID CERDA VILLABLANCA

CLASIFICACIÓN DE IMÁGENES DE CÁNCER GÁSTRICO APLICANDO APRENDIZAJE PROFUNDO

En Chile, el cáncer gástrico es el segundo tipo de cáncer con mayor mortalidad. Su alta mortalidad se debe a que no presenta sintomatología temprana por lo que frecuentemente es detectado cuando el cáncer está en etapa avanzada. Para detectarlo, un método es extraer una biopsia y someterla a un análisis inmunohistoquímico (IHC) lo cual genera una tensión particular en las células donde la proteína HER2 está sobreexpresada. La expresión de la proteína HER2 es un indicador de un subtipo de cáncer gástrico, el cuál tiene un tratamiento específico más efectivo. Las muestras son escaneadas generando imágenes de gran tamaño del orden de los 40 GB. Los médicos patólogos observan las imágenes y analizando la forma y color de las células determinan la clase de expresión HER2 de la muestra.

En esta memoria se entrena y evalúa una arquitectura de *deep learning* llamada InceptionV3 para realizar la clasificación de expresión HER2, con la idea de buscar mejoras en el rendimiento. Esta memoria propone: primero, replicar un método existente y, luego, realizar una propuesta sobre cómo realizar un entrenamiento con un subconjunto de imágenes sobre el cual se tiene un mayor grado de certeza de que esté correctamente etiquetado. Para realizar la replicación del método existente, en la etapa de entrenamiento, se utilizó la base de datos del estudio PRECISO que son imágenes de resecciones con anotaciones de patólogos (NCT01633203, N= 34 pacientes). Además, se ejecutan los entrenamientos en el supercomputador de la universidad Austral de Chile llamado Patagón.

Los principales resultados de este trabajo de memoria, son la implementación de la arquitectura InceptionV3 para los problemas de: clasificación tumor / no-tumor (2 clases) y reactividad HER2 (5 clases). En el primer caso se obtuvo un *accuracy* de 0.86. En el problema de 5 clases se obtuvo un *accuracy* de 0.5. Por otra parte, la implementación utilizó GPU de última generación que permitieron bajar los tiempos de cálculo de varios días a 8 horas. Por otra parte, se detalla e implementa un algoritmo para filtrar los parches de imágenes.

A mi madre Marianela
y a mi padre Luis.

Agradecimientos

Le agradezco a mi mamá Marianela por todo su amor, afecto, comprensión y apoyo en este largo proceso que ha sido mi etapa de formación. Le agradezco a mi papá Luis por todo su apoyo y adaptación en este proceso.

Le agradezco mi profesor guía Mauricio Cerda por haber sido el mejor profesor guía que un estudiante pueda haber tenido. Le agradezco por toda su paciencia y apoyo en los momentos difíciles vividos. Le agradezco por haber abierto las puertas del laboratorio SCIAN, que me permitió desarrollar esta memoria en un ambiente de trabajo enriquecedor y de crecimiento. Le agradezco a todas las personas del laboratorio por el recibimiento y por haberme contagiado la visión y espíritu de trabajar por un sistema de salud mejor.

Le agradezco a mi psicóloga Helena Silva por ser una excelente psicóloga y ayudarme mucho en este proceso. Le agradezco a mi psiquiatra Sergio Barroilhet por todos sus consejos y su ayuda médica.

Le agradezco a mis amigues Jorge, Chelo, Alexis, Sergio, Claudio, Valeska, Kevin, Martin y a todas las personas que he conocido en mi tiempo en la universidad y que han aportado a mi formación.

Le agradezco a los pacientes anónimos del estudio PRECISO cuyas muestras sirvieron para realizar este trabajo. A la Dra. Bettina Müller, por facilitar el acceso acceso a la base de datos para este trabajo. También agradecemos a los doctores que realizaron el proceso de etiquetado manual.

Le agradezco al Dr. Cristóbal Navarro líder del equipo del supercomputador Patagón de la Universidad Austral de Chile por haberme facilitado el acceso al cluster de computo y el soporte técnico.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Este capítulo se inicia exponiendo sobre el cáncer gástrico. Luego, se describe brevemente el estado del arte de los modelos de *deep learning* aplicados para detectar cáncer gástrico. Finalmente, se entrega una descripción general del trabajo realizado en esta memoria.

1.1. Introducción al cáncer gástrico

El cáncer gástrico es el 5º cáncer con mayor incidencia del mundo y es el 4º con mayor tasa de mortalidad [1]. Presenta mayor tasa de incidencia en Asia Oriental (22.9[1]), Europa Oriental (12.4) y Sudamérica (9), comparado al promedio mundial (7.1) [1]. En Chile, es el 4º cáncer de mayor incidencia en ambos sexos y el 4º de mayor mortalidad [2]. En suma, el cáncer gástrico constituye un importante problema de salud pública.

La principal causa de mortalidad del cáncer gástrico es el diagnóstico tardío debido a que cuando se presentan síntomas, son leves e inespecíficos. Incluso, es más, se estima que dos tercios de los pacientes consultan cuando la enfermedad se encuentra avanzada [3] [4]. Otro problema del proceso de diagnostico es que, al ser visual, la asignación de puntuación sobre estado del cáncer, consecuentemente, el proceso se vuelve propenso a errores dados por la variabilidad interobservador[5]

Existen diferentes tipos de cáncer gástrico. En particular, un tipo de cáncer gástrico es el asociado a la sobreexpresión de la proteína HER2. Se estima que entre el 9 % y el 38 % de los casos de cáncer gástrico son por tumor HER2 positivo[6]. La proteína HER2 actúa en la membrana de las células de la pared estomacal [7]. La proteína HER2 es parte del sistema que regula la proliferación y el crecimiento celular. Las células con proliferación y crecimiento excesivo, se asocian con tumores cancerosos.

[1] Tasa de incidencia por cada 100.000 habitantes.

El método usado para detectar el cáncer gástrico asociado a la sobreexpresión de HER2 es someter al paciente a una endoscopia; de la cual se obtiene una muestra de la pared gástrica. Este tejido se deposita en placas de vidrio. Luego, se somete a una tinción inmunohistoquímica (*immunohistochemistry* o IHC), lo que genera una coloración marrón en la membrana de las células donde la proteína HER2 está sobreexpresada. El tejido con tinción IHC se estudia con un microscopio generando imágenes en el espectro de la luz visible. Para diagnosticar las imágenes de IHC, los patólogos analizan la forma, el color, y los patrones de distribución de las células gástrica siguiendo guías clínicas establecidas [7]. El análisis de los patólogos puede ser negativo o positivo para cáncer. De los casos positivos, estos pueden ser HER2 positivos o no; siendo graduados en 0+ (negativo), 1+(negativo), 2+(equívoco), 3+ (positivo).

1.2. Breve introducción al estado del arte en clasificación automática de cáncer

En la literatura se encuentran métodos automáticos para realizar la clasificación de imágenes IHC, en particular para medición de HER2 en cáncer de mama y gástrico. Los estudios enfocados en la detección de cáncer de mama con técnicas de *deep learning* han mostrado resultados que pueden transformar el campo de la histopatología [8].

Los métodos basados en aprendizaje automático requieren de imágenes etiquetadas las cuales se generan al definir regiones de interés (*Region Of Interest* o ROI). Una ROI es una región de la imagen de biopsia. En particular, las ROI son circulares en la base de datos con que se trabaja en esta memoria. Cuando un patólogo etiqueta una ROI significa que, según la experiencia del experto, la mayoría de las células de esa región presentan una clase de expresión de HER2. Las ROI son divididas en parches. Un parche es un recorte de la imagen original que contiene una gran cantidad de células.

En particular, para el cáncer de mama existen diversas bases de datos como *BreaKHis* (8 mil imágenes-parches) [9] y IDC (53 mil imágenes-parches) [10], sobre las cuales, se ha aplicado el enfoque de *deep learning* clasificando en tejido canceroso y no-canceroso (o también, *normal* y *abnormal*). Los clasificadores entrenados en dichas investigaciones logran una exactitud entre 80 % y 85 %.

Lamentablemente, si se quisiera aplicar las mismas arquitecturas para el caso del cáncer gástrico, se presentan dificultades entre las cuales se pueden mencionar: (1) Que se dispone de bases de datos con cantidad de imágenes mucho menor que para el cáncer de mama. (2) Que las etiquetas son muy generales, en el sentido que solamente se clasifica en las clases normal y anormal. (3) La mayor complejidad del patrón de expresión en el cáncer gástrico en comparación al cáncer de mama. Sin embargo, el trabajo de tesis de Alegría [11] busca

diagnosticar sobreexpresión de la proteína HER2 (4 clases graduadas de 0+ a 3+) basado en imágenes con tinción HDAB (Hematoxilina-Diaminobencidina) de biopsias de cáncer gástrico. Esto es novedoso porque las bases de datos encontradas existentes están conformadas por imágenes con tinción HE.

En el trabajo de Alegría [11] se clasifican parches de las imágenes de biopsia. Esto significa que se asume el supuesto de que todas las células de un parche pertenecen a la misma clase de sobreexpresión de la proteína HER2. Cuando en realidad, eso sólo es cierto en la mayoría de los casos. Además, presenta la simplificación de que la información se considera a nivel de parche, es decir, a nivel de conjunto de células. Cuando, según la guía clínica, el análisis debería realizarse célula a célula.

Asumiendo el supuesto de que todas las células de una misma ROI son de la misma clase y procesando la información a nivel de parche, el trabajo de Alegría [11] se compone de tres etapas que presentan los siguientes resultados. La primera etapa es un clasificador de parches de clases Tumor y NoTumor, con un 88 % de exactitud. La segunda etapa es un clasificador de los parches clasificados como Tumor, en 5 clases dependiendo de su reactividad HER2, logrando una exactitud de 70 %. La tercera es la aplicación de la guía clínica [7] con la información de los parches clasificados, es un algoritmo que no debe ser entrenado, solamente es aplicado.

Entonces, globalmente se distinguen entre parches que pueden ser de 6 clases (NoTumor y las 5 clases de reactividad). Estas clases se mapean a 3 clases, simplificando el problema usando el algoritmo de la guía clínica[7]. Con lo que, se constituye un sistema de diagnóstico que logra una exactitud del 96 % con las clases: Negativo/ Equivoco/ Positivo.

El trabajo de Alegría presenta una exactitud alta porque se reducen los errores al mapear las clases (del problema de 6 clases) a un problema de 3 clases. En este mismo trabajo se declara que todos los modelos entrenados confunden la clase 0 con la clase 1+.

Por lo que, se identifica la oportunidad de construir un sistema que filtre los parches mal etiquetados. Para luego, reentrenar el modelo esperando mejorar el desempeño del sistema al clasificar las subclases.

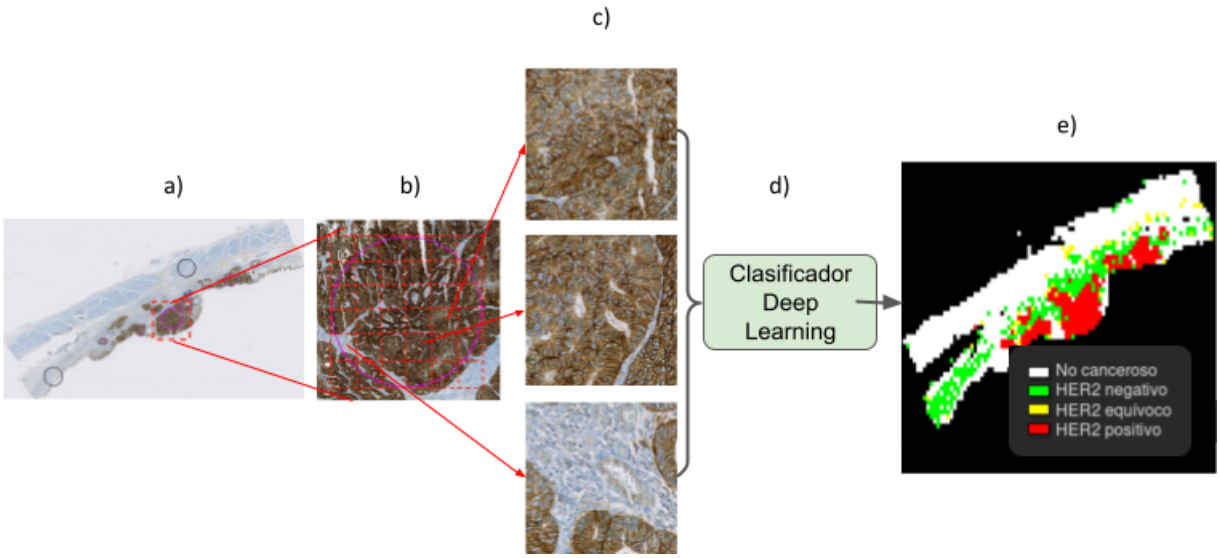


Figura 1.1: a) Imagen biopsia con círculos donde el color indica la anotación del patólogo (magenta es clase reactividad lineal fuerte), b) Anotación patólogo ampliada, c) Parches, d) Clasificador, e) Clasificación global del tejido usando la guía clínica [7].

1.3. Metodología

El trabajo realizado por Alegría [11] es un modelo base que logra un sistema de diagnóstico. La problemática que se observa es que dicho modelo es entrenado con datos de mala calidad. Los datos con que se entrena el modelo son parches. Un parche es una pequeña imagen cuadrada extraída de una ROI. Una ROI es una región de interés circular marcada por el patólogo en que la mayoría de las células presentan un tipo de reactividad HER2.

Los datos son de mala calidad en el sentido de que el etiquetado es muy grueso. Esto quiere decir que existen parches dentro de las ROI que no presentan la misma reactividad que la asignada por el patólogo a la ROI, sobre todo en los bordes.

El problema de fondo de trabajar con este dataset, o más bien, en este campo, es que a simple vista el programador **no** puede determinar si un parche está mal clasificado. No así, en otros dataset como Imagnet[12] en que es directo observar una imagen mal etiquetada. El programador no puede determinar si un parche está mal etiquetado porque no es un médico patólogo. Exclusivamente el médico patólogo puede clasificar un parches.

El *conocimiento del problema* que se tienen de que existen áreas de las ROI mal etiquetadas, se obtuvo de la descripción del dataset. Lamentablemente, durante la realización de este trabajo de tesis no fue posible ponerse en contacto con el doctor Pablo Zoroquiain, el cual realizó este etiquetado.

Se desconoce el porcentaje de parches mal etiquetados, o una métrica de qué tan ruidoso es el etiquetado. Por otro lado, de lo **sí** se tiene noción, es la cantidad de parches que están en los bordes de las ROI. También, de cuándo un parche posee sólo una esquina o pequeña área con células con coloración muy diferente, lo cual es un indicio que el parche está mal etiquetado. En el capítulo de origen de los datos y línea base se profundiza y se presentan ejemplos de parches mal etiquetados.

Respondiendo a la pregunta, ¿por qué se realizó de forma tan gruesa este etiquetado?. La respuesta es que el patólogo Dr. Zoroquiain realizó el etiquetado de forma voluntaria sin mediar pago alguno. Tal como se describe en este mismo capítulo, en la sección de introducción al cáncer gástrico, examinar visualmente una biopsia es un proceso que consume mucho tiempo de trabajo de una persona altamente calificada, lo cual tendría un alto costo monetario. En adición, fue necesario etiquetar 34 imágenes. En suma y dadas las condiciones expuestas, el etiquetado grueso realizado fue lo mejor que se pudo hacer con los recursos disponibles.

Esta memoria surge para responder a la pregunta, ¿podría mejorar el desempeño del modelo de Alegría [11] al reentrenarlo con un subconjunto de ejemplos con etiquetado de mayor calidad?. Luego, dicho subconjunto se obtiene con el algoritmo propuesto por esta memoria que es un sistema de etiquetado en base a características extraídas de los parches.

El *software* utilizado para programar el procesamiento de datos e imágenes es *Python 3.10*, se utiliza *KERAS 2.8* como *framework* de *deep learning*. El *hardware* que se utiliza es el supercomputador de la Universidad Austral, llamado *Patagón* [13], por su disponibilidad de aceleración del entrenamiento con *GPU NVIDIA A100-40GB*, usando *CUDA 11.5*.

1.4. Objetivo General

Replicar, reentrenar y evaluar la arquitectura de *deep learning* de Alegría [11] con un sub-conjunto de datos de mejor calidad.

1.5. Objetivos Específicos

1. Generación de base de datos de mejor calidad de etiquetado.
2. Entrenamiento y evaluación de una arquitectura de un algoritmo base presentado en la memoria de Alegría [11] y del algoritmo propuesto.
3. Aplicación de algoritmo de la guía clínica [7].

1.6. Alcances

Para mejorar el rendimiento de un modelo de *machine learning* el flujo de trabajo estándar es ajustar sus hiperparámetros (*hyperparameters tuning*). Esto implícitamente asume que se esta entrenando con un *dataset* de calidad. Entonces, si bien es cierto, en esta memoria se trabaja en mejorar un modelo, este no fue entrenado con data de la mejor calidad. Es por esto que, no se trabaja en ajustar sus hiperparametros. Sino que, se trabaja en mejorar la calidad del conjunto de entrenamiento. Teniendo como hipótesis que esto mejorará el rendimiento del modelo.

Esta fuera de los alcances de esta memoria el desarrollo de una interfaz gráfica de usuario. Tampoco esta considerado el desarrollo de un sistema que permita procesar las imágenes usando la interfaz de linea de comandos *Terminal*. Los códigos son ejecutables en *Jupyters Notebooks* y se agregan comentarios para facilitar la replicación de resultados.

Capítulo 2

Marco Teórico

En este capítulo se presenta la información necesaria para entender el problema abordado en esta memoria. Este capítulo se organiza en tres secciones que son: (1) Cáncer gástrico, en la cual se define en general el cáncer y en particular el cáncer gástrico. Se expone sobre las secciones del estómago, su epidemiología en el mundo y en Chile, sus causas, tratamiento y su diagnóstico mediante análisis de las biopsias. (2) *Machine learning*, donde se realiza un descenso explicativo desde lo general a lo particular; partiendo por explicar el aprendizaje de máquinas, desglosando en aprendizaje *no supervisado* y *supervisado*, pasando por el aprendizaje *profundo* hasta las redes neuronales convolucionales, dando pie a explicar la técnica de *transfer learning*. Se describen métodos que se utilizan para mejorar el rendimiento del modelo como: *Scaling*, *Data Agmentation* y *Balanceo de Pesos*. Se finaliza con las métricas de evaluación. Por último, la sección de (3) estado del arte donde se presentan y describen brevemente las investigaciones recientes de *Machine Learning* aplicado a cáncer gástrico y patologías relacionadas, como el cáncer de mama.

2.1. Cáncer gástrico

El cáncer es una enfermedad en la cual las células de un órgano proliferan demasiado o no se mueren cuando deberían, sobrepasando las células necesarias para que el órgano cumpla sus funciones [14]. Este exceso de células genera mal funcionamiento en el órgano. Además, las células cancerígenas se pueden propagar a distintas zonas del cuerpo y generar problemas en el funcionamiento de los órganos de otras zonas.

El estómago es un órgano que forma parte del sistema digestivo. En el estómago ocurre la digestión química y física de los alimentos. Las secciones del estómago son: Cardias, Fundus, Cuerpo, Antro y Píloro. El estómago se compone por capas que desde su interior a su exterior son: Mucosa (produce enzimas digestivas), Submucosa, Muscular propia, Subserosa y Serosa. Las copas del estómago se observan en la figura 2.1. El cáncer gástrico comienza

desarrollándose como lesiones en la capa mucosa.

En el caso del cáncer gástrico, las células de la pared estomacal proliferan y no se mueren cuando deberían hacerlo [14]. Entonces, ocupan los recursos de las células que funcionan correctamente. Lo que genera dificultades gástricas al paciente. Peor aún, cuando el cáncer gástrico se ramifica, comúnmente se aloja en la médula ósea de los huesos grandes como el fémur. Lugar donde perjudica las células que renuevan el hueso. Generando fuertes dolores al paciente y susceptibilidad a fracturas.

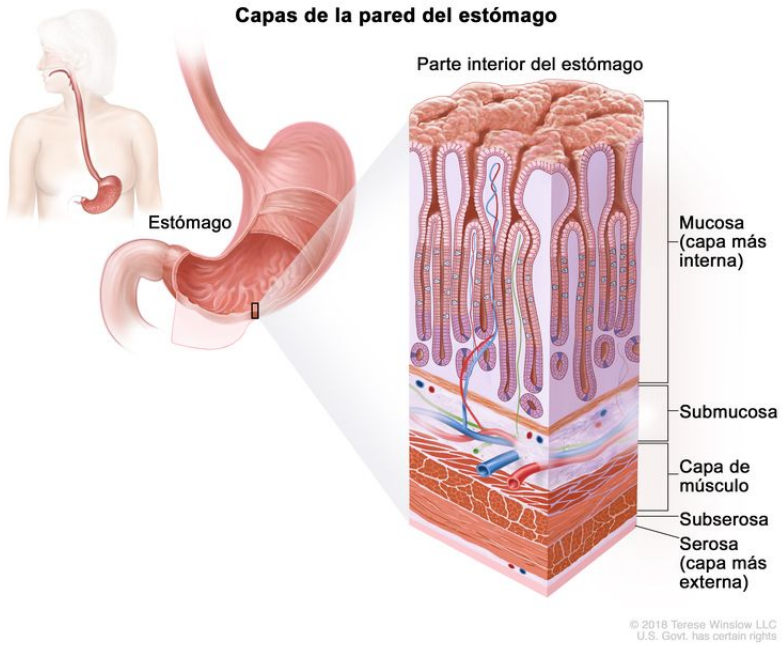


Figura 2.1: Ubicación y capas del estómago.

2.1.1. Epidemiología del cáncer gástrico en el mundo

¿Cómo afecta esta enfermedad a la población mundial?. Para responder esta pregunta, la disciplina científica de la epidemiología ha desarrollado métodos para caracterizar a la población en base a su edad, sexo, nivel económico y ubicación geográfica. Así como también, índices para describir el estado de afectación de una enfermedad en una población. Tales índices son la incidencia y la mortalidad. La incidencia es la cantidad de casos ocurridos en un periodo y lugar específicos; se mide en números absolutos o con respecto a 100.000 personas por año. La mortalidad es la cantidad de fallecimientos ocurridos en una región específica por cada 100.000 personas en un periodo de tiempo de un año. La incidencia y la mortalidad varían fuertemente con la edad, por lo que es inconcluyente comparar los índices de dos poblaciones con diferente distribución de edad. Luego, para poder realizar comparaciones se recalculan los índices ponderados por edad, con lo que se logran índices expresados en tasa

estandarizadas por edad. Los índices de incidencia y mortalidad ponderados por edad son los que se comparan para generar las observaciones que se exponen en esta sección.

La fuente principal de caracterización epidemiológica de los diferentes tipos de cáncer es el estudio GLOBOCAN, el cual analiza información sobre 36 tipos de cáncer en 185 países. La información es continuamente recopilada por el *Global cáncer Observatory* de la Organización Mundial de la Salud. Los resultados se publican en un artículo de investigación cada dos años.

Según el estudio GLOBOCAN del año 2020, al observar la distribución de los casos de cáncer gástrico en la población mundial, graficados en la figura 2.2, la ubicación geográfica de las regiones con mayor incidencia del cáncer gástrico son los países de Asia del Este, Europa del Este y Sudamérica.

En base a la clase socioeconómica de los pacientes, existen diferencias en cuanto a la sección del estómago en que se desarrolla el cáncer gástrico. El cáncer gástrico de tipo intestinal se puede ubicar en las secciones del Fundus, Cuerpos, Antro y Piloro. Además, usualmente es denominado como cáncer gástrico de ubicación *no-cardial*. Prevalece en países en desarrollo en las personas de nivel socioeconómico bajo. El cáncer gástrico de tipo difuso se ubica en la sección del estómago llamada Cardias; predomina en países desarrollados, principalmente en hombres de nivel socioeconómico alto [15].

Mundialmente, la incidencia del cáncer gástrico en el sexo masculino es 15.8 casos por cada 100.000 habitantes y en el sexo femenino es de 7.0 casos por cada 100.000 habitantes. La mortalidad del cáncer gástrico en el sexo masculino es 11.0 por cada 100.000 habitantes y en el sexo femenino es de 4.9. En cada una de las regiones del mundo la incidencia y mortalidad del cáncer gástrico es mayor en el sexo masculino que en el femenino [1].

Con respecto a cómo influye la variable edad en la epidemiología del cáncer gástrico, generalmente, el cáncer gástrico se diagnostica cuando está en etapa avanzada, porque presenta síntomas notorios sólo en su última etapa. La edad promedio de diagnóstico del cáncer gástrico es de 65 años. La tasa del mortalidad del cáncer gástrico aumenta al aumentar la edad del paciente tanto en el sexo masculino como en el sexo femenino [16].

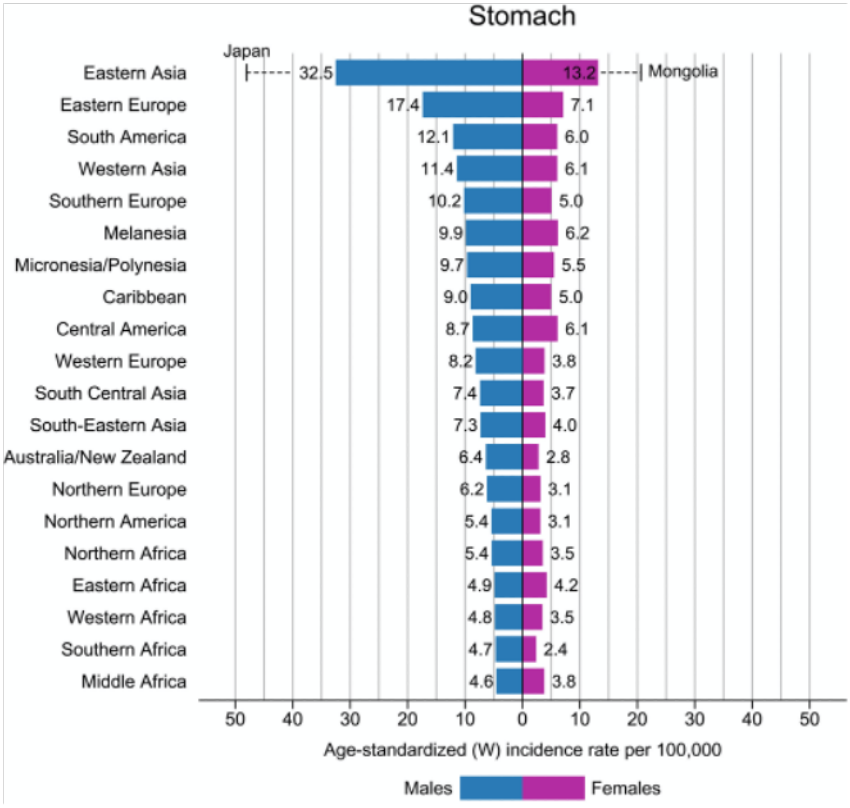


Figura 2.2: Incidencia del cáncer gástrico (estandarizada por edad) desglosada por sexo en las regiones del mundo en el año 2020 [1].

Con respecto al pronóstico del cáncer gástrico, se describe principalmente con el concepto de tasa de supervivencia. La tasa de supervivencia del cáncer es la cantidad de personas que siguen con vida una vez han sido diagnosticadas con cáncer en un periodo de tiempo de 5 años. En Estados Unidos, el cáncer gástrico tiene una tasa de supervivencia de 32 %. Lo que significa que por cada 100 personas que padecen cáncer gástrico, 32 siguen con vida después de 5 años de haber sido diagnosticadas [17].

Con lo que en suma, en la población mundial, el cáncer gástrico es una enfermedad que puede afectar a todas las personas de diversas edades y todas las regiones del mundo. Aunque más específicamente, tiene mayor mortalidad en hombres adultos mayores de Sudamérica, Asia del este y Europa de este.

2.1.2. El cáncer gástrico en Chile

En Chile, las regiones con mayor riesgo de mortalidad con respecto al país son las regiones de Maule, Bio-Bío y de los Ríos. Destacando, la comuna de Molina con la mayor mortalidad de cáncer gástrico en Chile. Situación que motivó el estudio MAUCO para indagar la relación con el aumento de contaminantes químicos en la comida y el aire de la comuna [4]. Por el

contrario, las regiones con menor mortalidad son las de Arica y Parinacota, y Antofagasta. La asimetría es tal, que resulta destacable el hecho de que los habitantes de la región del Maule tienen 2,3 veces más riesgo de mortalidad por cáncer gástrico, en comparación a las personas que viven en la región de Antofagasta [18].

Con respecto al nivel socioeconómico y contextualizando, en Chile existen dos sistemas de salud: uno público y otro privado. Las personas deciden en cuál sistema inscribirse en base a sus ingresos. Por otra parte, el estado de Chile garantiza el acceso gratuito a la detección y tratamiento de cáncer gástrico desde el año 2006; al hacerlo parte del plan gubernamental llamado AUGE-GES. Sin embargo, la estadística histórica muestra que las personas con menor escolaridad e inscritos en el sistema público, presentan mayor incidencia y mortalidad por cáncer gástrico [18]. Los chilenos que pertenecen a un nivel socioeconómico alto pueden acceder a mejores tratamientos, disminuyendo la tasa de mortalidad en ese sector de la población.

En relación al sexo biológico, los datos del estudio GLOBOCAN 2020 [1] muestran que la incidencia del cáncer gástrico en Chile, en hombres, es de 19.6 por 100.000 habitantes con un total de 2.783 casos. La incidencia del cáncer gástrico en Chile, en mujeres, es de 7.7 casos por cada 100.000 habitantes con 1.425 casos. Además, la mortalidad en hombres es de 15.3 con 2.211 fallecidos y de 5.7 en el sexo femenino con 1.106 decesos [19]. Los registros indican un consenso en que los hombres presentan mayor tasa de incidencia y mortalidad.

El pronóstico del cáncer gástrico en Chile tiene una situación similar al resto del mundo. Lamentablemente, los hombres chilenos presentan un riesgo elevado de mortalidad. Más especialmente, los habitantes de la región del Bio-bío, la cual está dentro de las 10 regiones con mayor mortalidad en el mundo desglosando por regiones.

A partir del año 2006 se incluye el cáncer gástrico en el plan AUGE-GES lo que garantiza su diagnóstico y tratamiento oportuno en plazos máximos definidos por ley. Sin embargo, el vencimiento de los plazos de detección y tratamiento si es una situación frecuente, teniendo diferencia de porcentaje de cumplimiento en las diferentes regiones de Chile. Sí bien es cierto, aunque se han hecho enormes avances en este problema de salud pública, se está llegando tarde a ofrecer un tratamiento potencialmente curativo; en el sentido de disminuir los casos cuyo mejor tratamiento es la extirpación total o parcial del estómago.

Los desafíos de Chile en el problema de salud del cáncer gástrico es aumentar la cantidad de pacientes diagnosticados con cáncer gástrico incipiente, ya que, es la única manera de mejorar el pronóstico [3].

2.1.3. Factores de riesgo y prevención

Los factores de riesgo que inducen el desarrollo de cáncer gástrico son principalmente alimenticios. Específicamente, una alimentación alta en contenido de sal, comidas preservadas y embutidos. Además, están factores clínicos, ya que, los casos de cáncer gástrico se han relacionado fuertemente con infecciones por la bacteria Helicobacter pylori (en el estómago) y reflujo gastroesofágico crónico.

Con respecto a factores de riesgo por estilos de vida, existe evidencia inequívoca que el hábito de fumar está asociado con el cáncer gástrico. No existe evidencia epidemiológica de la relación entre el consumo de alcohol y el cáncer gástrico [20].

La prevención se enfoca en la alimentación y en la detección temprana. El aumento de la ingesta de frutas es un factor de protección. En cuanto a la detección temprana, a pacientes con reflujo se recomienda realizar un control endoscópico cada 3 años. Además, se ha observado una relación entre esta enfermedad y las personas pertenecientes a etnias indígena como Inuit, Maorí y en de caso local de Chile como la etnia Mapuche [21]. La mayor parte de los pacientes con cáncer gástrico no poseen un antecedente hereditario de esta enfermedad [3].

2.1.4. Diagnóstico del cáncer gástrico

En Chile, las etapas del proceso para el diagnóstico del cáncer gástrico están escritas en la guía clínica del ministerio de salud (MINSAL) [22]. El proceso parte con una visita del paciente al médico general. Luego, se le deriva a un médico especialista. Cuándo hay sospecha de cáncer gástrico al paciente se le realiza una endoscopia. En la endoscopia se extrae una muestra de tejido gástrico. El paciente puede presentar lesiones de la pared estomacal asociada a cáncer gástrico. Un tipo de cáncer gástrico es el asociado a la sobreexpresión de la proteína HER2. Luego, específicamente para detectar el cáncer gástrico asociado a la sobreexpresión de la proteína HER2, la biopsia se somete a un análisis IHC por parte de un médico patólogo. El médico patólogo analiza la biopsia. Los resultados de la biopsia se presentan a un comité oncológico el cual determina la fase del cáncer. Finalmente, se procede al tratamiento y seguimiento del paciente.

El tiempo de demora el proceso de diagnóstico e identificación de la fase de un caso de cáncer gástrico es muy importante. Un caso de cáncer se puede diferenciar en dos tipos: cáncer ramificado y sin ramificar. Previamente, se expuso que la tasa de sobrevida del cáncer gástrico es de 32 %, lo cual incluye casos de pacientes con cáncer ramificado y sin ramificar. La sobrevida de un paciente con cáncer gástrico sin ramificar es de aproximadamente un 70 % [17]. Es por esto que los tiempos que demoren los exámenes de diagnóstico son muy

importantes porque pueden significar la diferencia entre tratar un cáncer ramificado y uno sin ramificar, el cual tiene mejor pronóstico.

En este sentido, en los casos difíciles de diagnosticar generalmente se pide una segunda opinión de otro patólogo o también se puede requerir rehacer la biopsia. Todo este tiempo es muy valioso para el paciente. Por lo que el sistema planteado en esta memoria tiene mucho sentido en este contexto. El poder generar automáticamente una segunda opinión es muy valioso en casos difíciles de diagnosticar, cuando las muestras de tejido tengan que transportarse grandes distancias para ser analizadas en caso de tener que realizar una segunda biopsia.

2.1.5. Tipos de cáncer gástrico

Existen diversas clasificaciones del cáncer gástrico. Según la ubicación existe el cáncer gástrico de tipo intestinal y el de tipo difuso. El de tipo intestinal se ubica en la zona no cardial, predominantemente en hombres de mayor edad. El tipo de difuso se ubica en el área del cardias dónde existen alteraciones de la mucosa gástrica asociado a enfermedad por reflujo gastroesofágico y con la obesidad.

También, existe la clasificación en base al estado clínico. La clasificación en base del estado clínico es usada para entregar información sobre el tratamiento y pronóstico de la enfermedad. Está clasificación entrega información sobre la extensión tamaño y estado de expansión a otros órganos. El sistema de estadificación estándar es el sistema **TNM**. En el sistema **TNM** el estado de un cáncer se expresa por las letras T, N y M; seguidas de un número que indica la clasificación. **T** que se refiere al tamaño y extensión del tumor primario. **N** se refiere a la cantidad de ganglios linfáticos a los cuales se ha diseminado o no el cáncer. Por último, **M** se refiere a la metastatización del tumor, es decir, si las células se han diseminado a otros órganos. Por ejemplo, la médula ósea de los huesos grandes, los pulmones y el hígado en el caso del cáncer gástrico.

2.1.6. Tratamiento

El tratamiento para el cáncer gástrico de tipo HER2 positivo, consta principalmente de quimioterapia y gastrectomía. Con el avance de las investigaciones, han surgido nuevos tratamientos provenientes de otros tipos de cáncer. En el año 2001, el estudio *"Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2"* de Slamon *et al* [23], mostró que el tratamiento con el medicamento llamado **Trastuzumab** está asociado a diversas mejoras en el pronóstico del cáncer de mama. El **Trastuzumab** es un anticuerpo monoclonal que se une a la proteína HER2 inhibiendo su cadena de funcionamiento con lo que se aminoran varios mecanismos de desarrollo del cáncer

de mama. Las mejoras en el tratamiento son: (1) aumento del tiempo hasta la progresión del cáncer de mama. Es decir, el tiempo desde el diagnóstico e inicio del tratamiento hasta que el cáncer comienza a empeorar. (2) Mayor tiempo de sobrevida y (3) reducción de la mortalidad.

Por otro lado, se estima que entre el 9 % y 38 % de los casos de cáncer gástrico son HER2 positivo, es decir, presentan tumores con sobreexpresión de proteína HER2. Los tumores con sobreexpresión HER2 con mayor frecuencia se ubican en la unión del estómago con el esófago.

En el 2010, el estudio clínico *"Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cáncer (ToGA)"* de [24], demostró que la quimioterapia acompañada con el tratamiento con **Trastuzumab** prolonga de forma estadísticamente significativa la sobrevida de los pacientes con cáncer gástrico. En el año 2016 la *American Society for Clinical Pathology (ASCP)* publica una guía clínica para la evaluación de la sobreexpresión de la proteína HER2 en tejidos de pacientes con sospecha de cáncer gástrico [7]. La publicación de la guía clínica logra estandarizar un sistema de evaluación para el análisis de muestras de tejido posiblemente canceroso. Este sistema de evaluación es utilizado por los médicos patólogos en todo el mundo.

La evaluación de muestras la realiza un médico patólogo analizando tejidos cancerosos. El tejido canceroso se extrae mediante dos tipos de procedimientos diferentes que son: la biopsia endoscópica y la resección. En la biopsia endoscópica se extrae una muestra de tejido durante una endoscopia. Mientras que, en la resección se extrae quirúrgicamente el tejido que posteriormente será analizado.

Finalmente, de ser detectada la sobreexpresión de HER2, el tratamiento de quimioterapia es acompañado con el medicamento **Trastuzumab**. Esto aumenta el tiempo transcurrido hasta que el cáncer comienza a empeorar. Con lo que mejora la tasa de respuesta a tratamientos, aumentando el tiempo de sobrevida y reduciendo la probabilidad de fallecimiento del paciente.

2.2. Interpretación tinsión inmunohistoquímica

2.2.1. Proteína HER2

HER2 es una proteína que participa en el mecanismo de crecimiento y proliferación celular. La proteína HER2 es codificada por el gen llamado **HER2**. El gen HER2 es un protooncogén. Un protooncogén es un gen que cuando es activado por mutaciones adquiere funciones oncogénicas, es decir, se involucra en el desarrollo del cáncer. Por esto, se reconoce a la proteína

HER2 como un marcador tumoral.

La proteína HER2 es un receptor de la hormona *tiroxina quinasa*. Esta proteína se encuentra en toda la membrana citoplasmática de las células. La membrana citoplasmática (o membrana celular) es una bicapa de moléculas lipídicas que constituyen una barrera entre el medio intracelular y el medio extracelular. Luego, las proteínas HER2, al ser receptores de una hormona, se ubican en la membrana celular entremedio de sus moléculas lipídicas.

La sobreexpresión de la proteína HER2 significa que la célula ha sintetizado muchas más proteínas HER2 de las que debería sintetizar. Entonces, en la membrana celular van a existir muchas más proteínas de las que deberían existir. La cantidad adecuada de proteínas HER2 sintetizada por la célula es la que asegura el correcto funcionamiento de los mecanismos de crecimiento y proliferación de la célula. Las proteínas HER2 se ubican a lo largo y ancho de toda la membrana celular. La sobreexpresión no significa que existan zonas de la membrana celular conformadas enteramente por proteínas HER2. Sino que, significa que hay muchas más proteínas HER2 insertas en la membrana celular en su conjunto. Es por esto que, cuando se observan las imágenes con sobreexpresión de proteína HER2, se observa una membrana celular con coloración intensa a lo largo y ancho de toda la membrana.

Cuando los patólogos observan un tejido de biopsia en el microscopio, les es difícil diferenciar células, debido a que las imágenes tienen muy poco contraste. Además, las proteínas HER2 de las membranas celulares no presentan una coloración visible al ser observadas en el microscopio. Entonces, para constatar la expresión de proteínas HER2, es decir, ver las moléculas HER2 que la célula sintetiza, es necesario marcar de alguna manera las proteínas HER2. Esto se realiza con la tinción IHC. Una vez marcadas visualmente las proteínas HER2 es que los patólogos pueden realizar una evaluación de si existe o no sobreexpresión.

Con respecto a la detección de la proteína HER2 es pertinente aclarar que siempre se encuentra proteína HER2, es decir, las membranas de las células siempre tendrán leve coloración marrón. Existe un nivel normal de coloración marrón. La tinción IHC permite detectar cantidades anormalmente grandes de proteínas HER2. Entonces, la interpretación de si la intensidad del color obtenido corresponde a tal nivel de sobreexpresión está sujeto a la visión del patólogo en particular[2]. Por eso existen diagnósticos *equívocos*, que resultan en repetir la biopsia.

[2] Lo que parecería ser relativo, sin embargo, no lo es tanto debido a que los patólogos reciben formación muy similar.

2.2.2. Tinción inmunohistoquímica IHC

La tinción inmunohistoquímica (IHC) es un procedimiento bioquímico que sirve para detectar, amplificar y hacer visible un marcador proteico. En este caso el marcador proteico es la proteína HER2 ubicada en la membrana citoplasmática de las células. La tinción IHC permite teñir las células y su membrana. Gracias a esto, el análisis de los patólogos aporta información para el diagnóstico y estadificación de un paciente con cáncer.

Las moléculas de la célula en general no son visibles bajo la luz del microscopio. Se requiere un procedimiento poder constatar la existencia de moléculas particulares particulares. Este procedimiento se llama tinción, a lo largo de la historia se han desarrollado varias técnicas de tinción. En particular, la aplicada en los tejidos de la base de datos con que se trabaja en esta memoria se llama tinción IHC. En la tinción IHC lo que se visualiza con el microscopio y en las imágenes no es el marcador proteico en sí. Sino que, lo que se visualiza de un color particular es un complejo molecular compuesto por: un sustrato, un anticuerpo y una enzima. La figura 2.3 es un ejemplo de las imágenes obtenidas con tinción IHC.

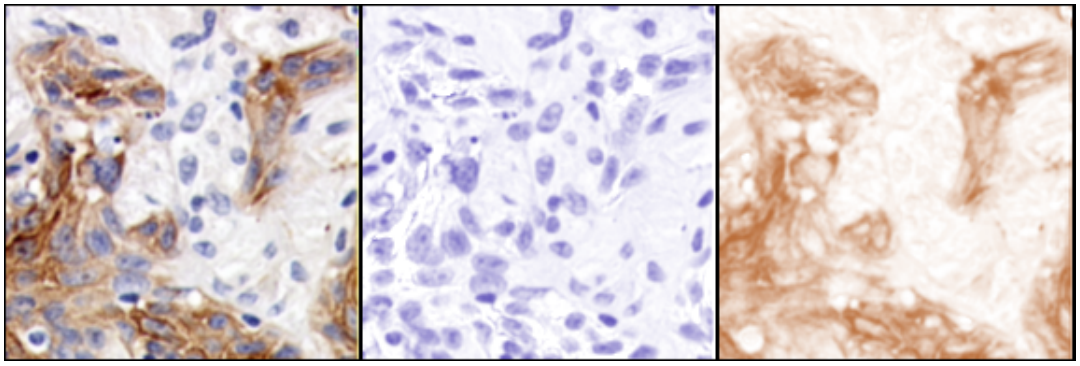


Figura 2.3: A la izquierda imagen con tinción IHC, al centro imagen post-procesada de canal H, a la derecha imagen post-procesada de canal DAB.

Explicando componente a componente el complejo que se visualiza en las imágenes 2.3: el sustrato es la proteína HER2; es la proteína que originalmente se quiere hacer visible. Al sustrato también se le conoce como antígeno. Un antígeno es una molécula que tiene la capacidad de ser reconocida por el sistema inmunitario; esto es importante porque para poder realizar la tinción tiene que existir un mecanismo de manera tal que las sustancias que se agregan a la biopsia reconozcan específicamente las proteínas que se necesita teñir. En este caso, la proteína HER2 es el antígeno, es decir, el sistema inmunitario genera moléculas específicas llamadas anticuerpos que cuando se agregan al tejido logran 'detectar o reconocer' las proteínas HER2.

El anticuerpo es una proteína sintetizada por las células blancas del sistema inmunitario de los animales en respuesta a un antígeno. Su estructura está hecha para reaccionar solamente

con el antígeno. Entonces, en el contexto práctico de la tinción IHC en el laboratorio. Lo que sucede es que las empresas desarrollan anticuerpos para detectar antígenos específicos. Por ejemplo, en el Kit de tinción IHC llamado *Herceptest*, desarrollado por la empresa *Agilent-Dakos*, el anticuerpo se llama *Rabbit Anti-Human HER2 protein* (es una molécula generada por el sistema inmunitario del conejo enlazada con la molécula Hemocianina proveniente del molusco llamado Lapa Californiana). Más allá de entender la profundidad química, el punto es que las empresas diseñan y sintetizan estas proteínas anticuerpo, que van a reaccionar específicamente con la proteína HER2 de las células humanas.

El siguiente problema que aparece es que ambas moléculas antígeno y anticuerpo son invisibles al microscopio. Entonces, lo que hacen las empresas que diseñan los Kit de tinción IHC es 'adherir' (mediante un enlace químico) al anticuerpo una molécula llamada enzima, o también llamado agente cromógeno. Entonces, el kit de tinción IHC consiste principalmente en el anticuerpo más la enzima.

En general, las enzimas son proteínas que aceleran las reacciones bioquímicas del cuerpo. Además, las enzimas son moléculas que poseen especificidad. La especificidad significa que la molécula solamente reacciona con otras moléculas específicas. Entonces, en el caso de la tinción IHC, la enzima reacciona con el antígeno y el anticuerpo, produciendo un color específico.

El triunfo de la tecnología de tinción IHC es haber desarrollado una enzima que reacciona específicamente con una molécula, generando una coloración específica. **La enzima que se utiliza en la tinción IHC de las proteínas HER2 es la DAB (Diaminobencidina)**. La enzima DAB genera una coloración **marrón** en la posición donde se encuentra la molécula de proteína HER2.

En suma, para realizar la tinción IHC, la muestra de tejido es la que posee el sustrato o antígeno que es la proteína HER2. A la muestra se le agrega una solución con el anticuerpo y la enzima, en todas las áreas del tejido por igual. En las áreas del tejido donde exista proteína HER2, sucederá la reacción bioquímica con el complejo: Antígeno (proteína HER2) - Anticuerpo - Enzima (DAB), y se generará una coloración marrón/rojiza.

Comúnmente, se conoce como tinción IHC a todo el procedimiento en general de tinción. Sin embargo, en la práctica son dos ciclos de tinción, es decir, dos ciclos en que suceden reacciones químicas que generan coloración en la posición de una molécula en particular. En las imágenes utilizadas en este trabajo, el primer ciclo de tinción genera la coloración marrón/rojizo con el cromógeno DAB. El segundo ciclo de tinción genera coloración azul usando como agente cromógeno a la hematoxilina (H). **La hematoxilina (H) genera coloración azul en el núcleo de la célula al reaccionar con los ácidos nucleicos**. En otras palabras, la hematoxilina marca los núcleos celulares. En adición, la tinción con hematoxilina (H)

también sirve para aumentar el contraste de la imagen para distinguir mejor las membranas celulares.

La técnica de tinción IHC posee ventajas y desventajas con respecto a otras técnicas de tinción. Las ventajas de la tinción IHC son que permite retener la microestructura del tejido, que la señal de color puede ser obtenida varias veces y que la tinción de la muestra se conserva por años. Su desventaja es que la coloración obtenida es una variable **cualitativa** (no cuantitativa); esto significa que no se puede observar la cantidad de moléculas teñidas, sino que el patólogo solo puede expresar cualitativamente, es decir, con palabras categóricas un descripción de la señal de color que observa. En la subsiguiente sección de guía clínica se explican las categorías de coloración que utilizan los patólogos.

2.2.3. Obtención de imágenes

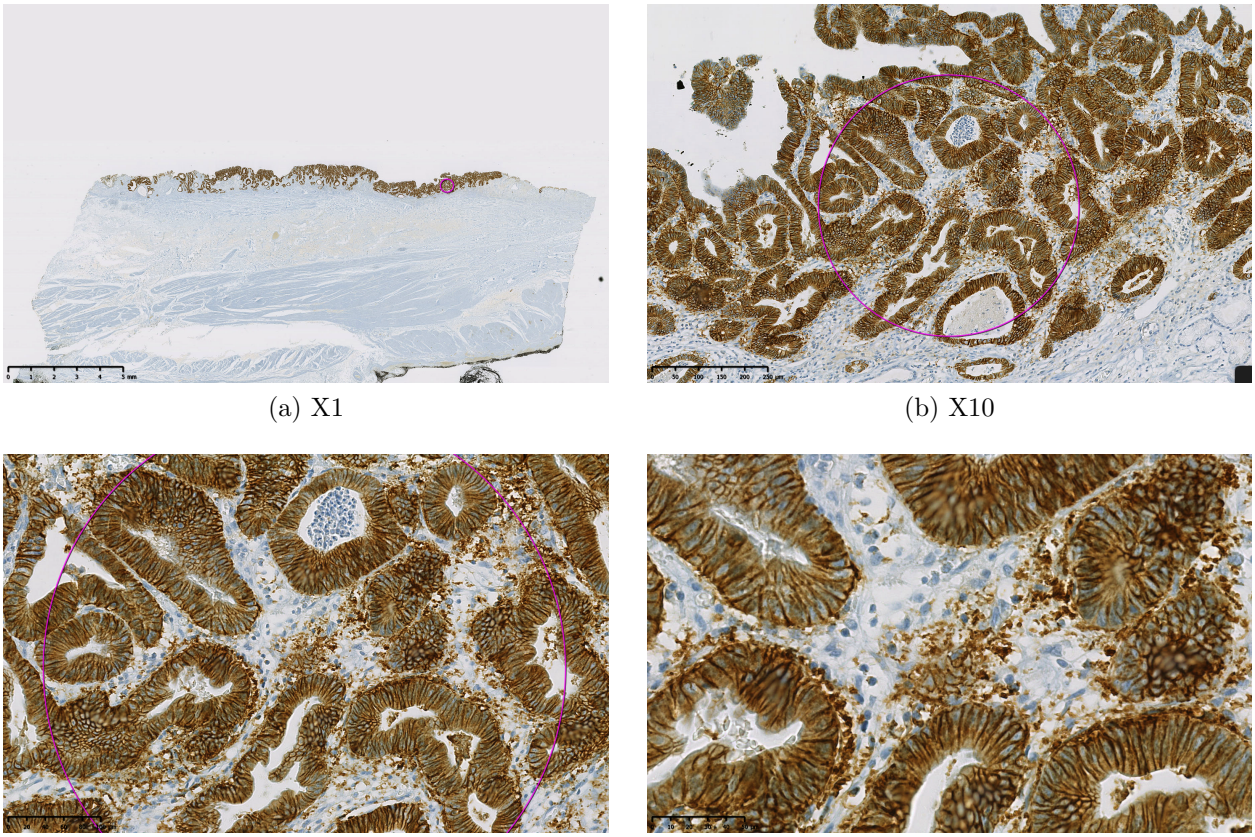
Las muestras de tejido con tinción IHC se depositan en un portaobjetos. Un portaobjetos es una lámina de vidrio de forma rectangular usada para almacenar muestras de tejido. Las muestras depositadas en el portaobjetos se escanean en un escáner de placas histológicas. Un escáner de placas histológicas es un dispositivo que dirige un haz de luz hacia la muestra, los cuantos de luz interactúan con las moléculas del tejido y el resultado es capturado por un sensor transductor generando una imagen digital en el computador.

Las imágenes generadas por microscopio digital son de alta resolución. Aproximadamente, de 12 mil millones de pixeles a resolución completa y sin usar compresión. Estas imágenes pesan del orden de 30 Gb por lo cual no pueden ser manipuladas por la memoria RAM de un computador. La solución es almacenar las imágenes en formato piramidal. Cuando se almacena una imagen en formato piramidal lo que se guarda es una serie de magnificaciones distintas, cada magnificación es representada un piso de un pirámide. Luego, cuando el patólogo analiza la imagen utiliza un software especial que carga en memoria solo la magnificación que se requiera; todo esto, a través de una interfaz gráfica que le permite hacer zoom a nivel de celular.

Algunas consideraciones sobre el proceso de tinción y captura de imágenes, son que: **todas las tinciones de las imagenes del dataset *PRECISO* son realizadas con el mismo tipo de kit de tinción y usando la misma metodología**. Además, todas las imágenes son obtenidas con el mismo microscopio digital. Esto se menciona porque la intensidad de los colores depende levemente del kit de tinción y de qué tan rigurosa en tiempos fue la metodología de tinción.

En la figura 2.4 se observan cuatro imágenes de distintas magnificaciones para una imagen de cáncer gástrico HER2 positivo. La figura (a) es una vista panorámica de la muestra

de tejido (biopsia). La figura (b) es un acercamiento a magnificación x10 y muestra un círculo violeta, que es ejemplo de las anotaciones que realizan los patólogos. La figura (c) es un acercamiento a x20 de dicha anotación. Finalmente, la figura (d) es un acercamiento a magnificación x40, que es la máxima magnificación posible, en ella los patólogos pueden observar a nivel de célula.



(a) X1

(b) X10

(c) X20

(d) X40

Figura 2.4: Ejemplo de múltiples magnificaciones

2.2.4. Guía clínica

La guía clínica es un artículo de investigación titulado *"HER2 Testing and Clinical Decision Making in Gastroesophageal Adenocarcinoma: Guideline From the College of American Pathologists, American Society for Clinical Pathology, and the American Society of Clinical Oncology"* publicado el año 2017 por la ASCO, la CAP y la Sociedad Americana de Patología Clínica (ASCP) en la revista *Journal of Clinical Oncology*. Su función es guiar a los médicos patólogos en el análisis de una muestra de tejido para el diagnóstico de cáncer gástrico HER2 positivo. Su importancia es que especifica un sistema de diagnóstico para las biopsias y otro para las resecciones. Ambos sistemas de clasificación se condensan en los datos que se presentan en la tabla 2.1.

El proceso de análisis de una muestra consiste en las siguientes etapas. Los patólogos reciben una imagen del tejido generada por el microscopio, es decir, un archivo digital. Las células del tejido presentan diferentes reactividades producidas en el proceso de tinción inmunohistoquímica (IHC). Las imágenes con tinción IHC presentan regiones de células con cierto tipo de reactividad predominante. El patólogo identifica las regiones y les asigna una puntuación. Una muestra puede tener varias regiones con distintas puntuaciones. Luego, considerando las regiones identificadas y usando la regla expresada en la tabla 2.1 se genera una puntuación global. La puntuación global se traduce directamente en el diagnóstico usando la tabla 2.1.

La tabla 2.1 de la guía clínica establece un sistema cuando la muestra es una resección y cuando es una biopsia. La principal diferencia es que cuando la muestra es una resección la puntuación depende del número porcentual de células que presenten la reactividad con respecto al total de células de la muestra. Se establece un umbral del 10 % de las células de la muestra para darle cierta puntuación. En cambio, cuando la muestra es una biopsia y se identifica un cluster de células tumorales (agrupación de 5 o más células) no importa el porcentaje de células para asignar la puntuación, sólo se considera la reactividad del cluster encontrado. En esta memoria se trabaja con emular la evaluación de resecciones, por lo que se debe tener presente la regla del umbral del 10 %.

Tabla 2.1: Pauta de puntuación para interpretación de la tinsión inmunohistoquímica de HER2 en carcinoma gástrico. Traducción de Bartley *et al* [7].

Patrón de teñido en espécimen quirúrgico por resección	Patrón de teñido en espécimen obtenido por biopsia	Puntuación	Clasificación HER2
Sin reactividad alguna o reactividad membranosa en <10 % de las células	Sin reactividad alguna o sin reactividad membranosa en ninguna célula tumoral	0	Negativo
Reactividad débil/apenas perceptible en ≥10 % de las células tumorales; las células son reactivas sólo en parte de su membrana.	Cluster* de células tumorales con reactividad membranosa débil/apenas perceptible, sin importar el porcentaje de células tumorales teñidas	1+	Negativo
Reactividad membranosa débil a moderada en ≥10 % de las células tumorales; la reactividad es completa, basolateral o lateral	Cluster* de células tumorales con reactividad membranosa débil a moderada, sin importar el porcentaje de células tumorales teñidas; la reactividad es completa, basolateralo lateral	2+	Equívoco
Reactividad membranosa fuerte en ≥10 % de las células tumorales; la reactividad es completa, basolateral o lateral	Cluster* de células tumorales con reactividad membranosa fuerte, sin importar el porcentaje de células tumorales teñidas; la reactividad es completa, basolateral o lateral	3+	Positivo

Por último, en la figura 2.5 se presentan ejemplos de imágenes representativas de cada tipo de puntuación. Las puntuaciones y sus correspondientes imágenes son: En la imagen A) la puntuación es 0 y representa un diagnóstico negativo, en la imagen las membranas de las células no presentan coloración marrón. En B) la puntuación es 1+ también representa un diagnóstico negativo, en la imagen se observan algunas pocas células con tinción en sus membranas débil o apenas perceptible. En C) la puntuación es 2+ significa diagnóstico equívoco, en la imagen se aprecian células con sus membranas con coloración marrón débil, el siguiente paso que debe seguir el paciente es repetir la biopsia. Finalmente, en la imagen D) se observa gran cantidad de células tumorales con sus membranas con tinción completa de un color marrón intenso; esta imagen representa la puntuación de 3+, lo que significa un diagnóstico positivo.

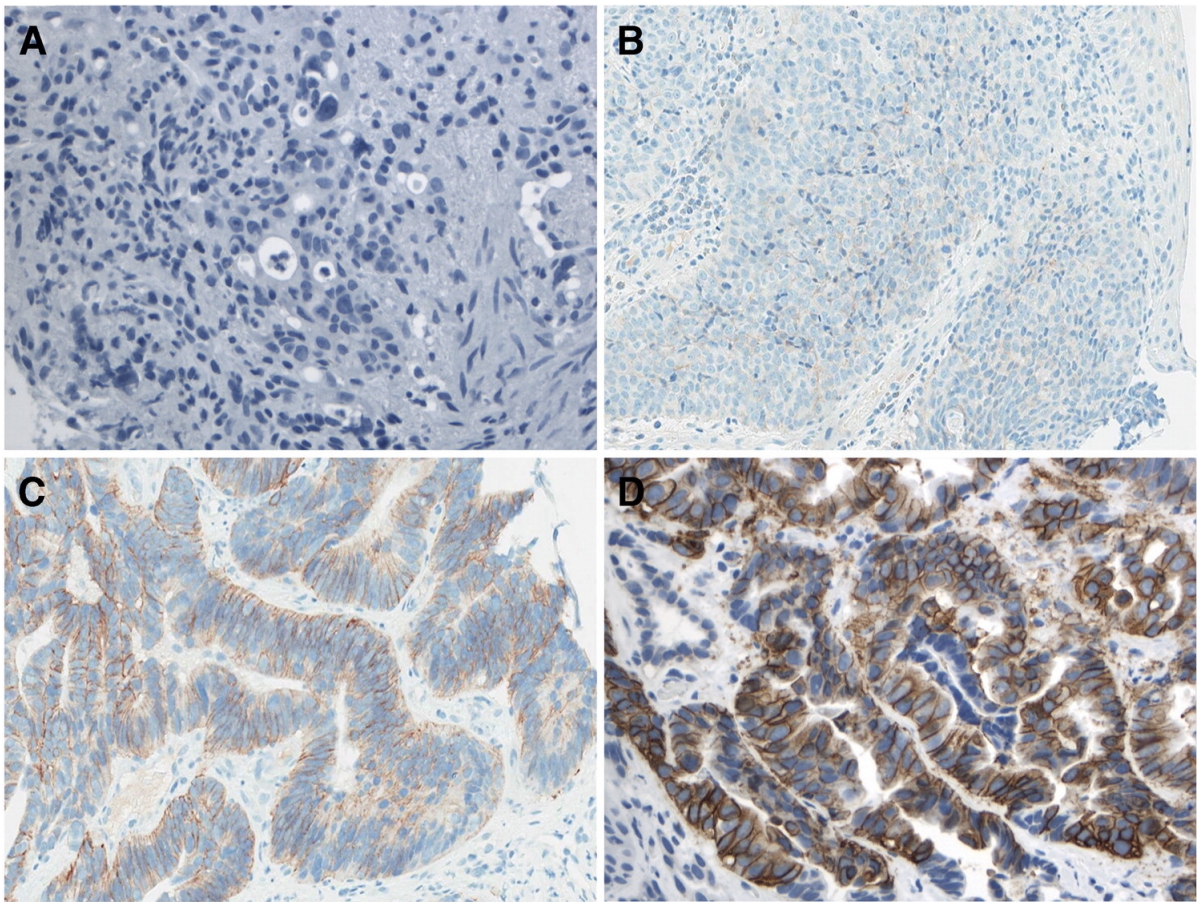


Figura 2.5: Análisis inmunohistoquímico en muestras representativas de expresión de HER2 en cáncer gástrico. A) 0, negativo. B) 1+, negativo. C) 2+, equívoco. D) 3+, positivo. Fuente: Bartley *et al* [7]

2.3. Machine Learning

El *machine learning* es uno de los enfoques con que se aborda el desarrollo de la inteligencia artificial. La inteligencia artificial es una rama de las ciencias de la computación que tiene como objetivo desarrollar algoritmos que posean la capacidad de la inteligencia. Según el Ph.D Andriy Burkov autor del libro *Machine Learning Engineering*, el *machine learning* es *"un área de las ciencias de la computación dedicada a construir algoritmos que, para ser útiles, se basan en una colección de ejemplos de algún fenómeno"* [25].

Por ejemplo, en el caso de esta memoria se busca desarrollar un algoritmo que detecte sobreexpresión de HER2. Esto significa que el algoritmo tiene que recibir una imagen y luego, generar una respuesta que es el tipo de sobreexpresión que presenta la imagen. Además, el algoritmo aprende, es decir, actualiza sus parámetros internos minimizando el error de clasificación.

El campo del *Machine learning* se puede dividir en: aprendizaje *no supervisado*, *aprendizaje supervisado*, *aprendizaje semi supervisado*, *aprendizaje reforzado*. En este documento se explicarán los dos primeros debido a que son los que se utilizan en este trabajo.

2.3.1. Aprendizaje no supervisado

El objetivo de un algoritmo de aprendizaje no supervisado es crear un modelo que tome un vector de características como entrada y lo transforme en otro vector o en un valor que pueda usarse para resolver un problema práctico [25]. El aprendizaje no supervisado se trata de construir algoritmos basándose en ejemplos de los cuales no se poseen etiquetas.

Algunos ejemplos de problemas donde se aplica el aprendizaje no supervisado son: la *reducción de dimensionalidad*, la *detección de outliers* y el *clustering*. En la *reducción de dimensionalidad* se ingresa un vector de características y el modelo de aprendizaje no supervisado retorna un vector con menos dimensiones que el de entrada. Esto se utiliza para hacer visualizaciones en 2 o 3 dimensiones de datos que tengan más de 3 dimensiones. En la *detección de outliers*, el modelo de aprendizaje no supervisado analiza los datos. Luego, al ingresar un ejemplo se retorna un valor que indica si el ejemplo es un valor atípico con respecto al comportamiento observado en los datos.

En el *clustering*, cuando se entrena un modelo se están buscando patrones en el conjunto de ejemplos, con los que se generan *clusters*. Un *cluster* es un conjunto de datos que tienen patrones en común. El resultado de un modelo de clustering es un vector de dimensión uno, es decir, es un identificador que indica el cluster al cual pertenece un ejemplo.

El funcionamiento de un modelo de *clustering* consiste en ingresar un conjunto de datos y la cantidad de *clusters* que se conjetura que existen, luego, se procesan los datos resultando los *clusters*. Los parámetros con que se caracteriza un *cluster* dependen del algoritmo de *clustering* elegido. Por ejemplo, el algoritmo de *clutering* de *k-means*, define un *cluster* mediante el concepto de *centroide*. La calidad de los *centroides* se mide usando métricas de *clustering*. Una vez obtenidos los *centroides*, es posible clasificar un nuevo ejemplo (o vector de características) en base a la mínima distancia entre el vector de características y los centroides que representan a los clusters. Un modelo de este tipo de funcionamiento es el que se propone para mejorar el dataset con que se entrena el modelo de Alegría [11]. Este modelo de *clustering* se describe en detalle en el capítulo 5 de algoritmo propuesto.

2.3.2. Aprendizaje supervisado

En el aprendizaje supervisado la tarea que se optimiza es clasificar el tipo de un fenómeno. Por ejemplo, en esta memoria se trabaja con imágenes que presentan el fenómeno de mostrar una clase de sobreexpresion de proteína HER2, los tipos o clases de este fenómeno son las distintas clases de sobreexpresión presentadas anteriormente. Cada manifestación del fenómeno se llama ejemplo o muestra. En este caso, un ejemplo es una imagen más la etiqueta que le asigna él patólogo.

En el aprendizaje supervisado se parte con un modelo matemático con parámetros variables. El modelo matemático tiene dos modos de funcionamiento: predicción (también llamado clasificación) y entrenamiento. En el modo de predicción, el modelo recibe una muestra y retorna la clase de las muestra. La clase de la muestra es un número, 0 o 1, en el caso de un modelo de clasificación binaria. En el modo de entrenamiento, el modelo mejora su rendimiento de clasificación, esto lo hace recibiendo la muestra y su clase real. La clase real es la clase que realmente debería obtenerse para la muestra. La clase real proviene de un supervisor humano, el supervisor humano clasifica la muestra usando sus conocimientos. El modelo mejora su rendimiento de clasificación modificando sus parámetros. Sus parámetros se llaman pesos. Los pesos se modifican usando un algoritmo llamado *backpropagation*.

División de base de datos

Al entrenar un modelo de aprendizaje supervisado está presente la pregunta: ¿como se sabe si el modelo entrenado logra generalizar?. La capacidad de generalizar tiene que ver con que la red logra clasificar correctamente un nuevo *feature vector* ligeramente diferente a los que la red estaba entrenada para clasificar. Entonces, para evaluar correctamente la capacidad de generalización de la red entrenada lo que se hace es separar la base de datos en tres conjuntos llamados *training*, *validation* y *evaluation*. Comúnmente, a la división de la base de datos se le llama *dataset split* y se usan las proporciones de *60 % training (train) , 20 % validation (val)*

y 20 % evaluation (eval). Durante la etapa de desarrollo, la red se entrena con los datos del conjunto de *train*, y se evalúa con los datos del conjunto de *validation*. En base a las métricas obtenidas con el conjunto de *validation* es que se varían los hiperparametros de entrenamiento de la red. Con esto se logra que la red sea evaluada en datos ligeramente diferentes a los que fue entrenada, es decir, se logra una medida de la capacidad de generalización de la red.

Ahora bien, ¿donde se ocupa el conjunto de *evaluation*?. Lo que sucede es que, durante la etapa de desarrollo del modelo, existen varias ocasiones en que se toma la decisión de reentrenar la red con nuevos hiperparametros. Esa decisión se toma con información proveniente de datos del conjunto de *validation*. Es decir, la información del conjunto de *validation* sirve para modificar la red. Luego, no es correcto que las métricas que describen el rendimiento de la red sean obtenidas con datos que sirven para modificar la red. Aquí es donde entra el conjunto de *evaluation*. El conjunto de *evaluation* es una porción de los datos que **no** se usa para ninguna decisión de desarrollo. El conjunto de *evaluation* solo se utiliza al final de la etapa de desarrollo de la red para obtener las métricas finales, que se reportan como resultado del estudio.

Early stopping

Una vez introducida la idea de división de la base de datos y continuando con las preguntas relacionadas al entrenamiento, surge la pregunta: ¿durante cuanto tiempo se deben entrenar la red?. Para responder esta pregunta es necesario introducir el concepto de época, una época es un ciclo de procesamiento en que la red ve el conjunto de datos de entrenamiento por completo.

El método de *early stopping* responde a la pregunta de qué cantidad cantidad de épocas la red debe entrenarse de modo que conserve su capacidad de generalizar.

Entonces, y en retrospectiva, se entrena el modelo con el conjunto de *train*. Luego se calcula un *accuracy* usando el conjunto de *train* (acc_{train}) y otro *accuracy* usando el conjunto de *validation* (acc_{val}). Esto se repite iterativamente durante cierta cantidad de épocas obteniendo un gráfico como el de la figura 2.6. En el se observa que el *accuracy* de *validation* comienza a descender a partir de cierta época, a diferencia del *accuracy* de *train* que crece asintóticamente. Lo que significa que a partir de esa época el modelo pierde su capacidad de generalizar, es decir, entra en *overfitting*. El método de *early stopping* indica que se debe parar de entrenar en el punto en que el *accuracy de validation* comienza a descender. Finalmente, se toma el modelo entrenado hasta dicha *época* y se calcula el *accuracy* en el conjunto de *evaluation*. El *accuracy de evaluation* es el resultado final con que se valora el modelo.

Figura 2.6: Gráfico de *accuracy vs ápocas* donde se visualiza la *epoch* de *early stopping*.

LOOCV

LOOCV es una estrategia de entrenamiento y evaluación de modelo predictivos. La estrategia *LOOCV* permite reducir la variabilidad que se origina al dividir las observaciones de una forma específica. Con la estrategia *LOOCV* se dividen las observaciones en todas las maneras posibles, es decir, se prueban todos los *fold* posibles. Si bien es cierto, realizar un estudio con la estrategia *LOOCV* demora mucho más tiempo de cómputo que usar una partición aleatoria de la base de datos. La ventaja es que usar la estrategia *LOOCV* permite entregar argumentos de mayor peso con respecto a si se comprueba o no la hipótesis de este estudio.

El método de *LOOCV* significa *leave one out cross validation*. En este método se extraen las imágenes provenientes de una biopsia y se entrena el modelo con el resto de las imágenes. Luego, el modelo se evalúa con las imágenes del conjunto de imágenes que se extrajo. En el caso del trabajo de esta memoria se tienen imágenes de 34 pacientes. Por lo tanto, en cada *corrida* de entrenamiento se toman las imágenes de un paciente para la evaluación del modelo y se entrena con las imágenes de los otros 33 pacientes. Entonces, se finaliza con 34 modelos entrenados y sus respectivas métricas. Los modelos se entrenan y evalúan con esta metodología porque un modelo no debe ser evaluado con las imágenes que fue entrenado. Cuando se evalúa un modelo con los mismos datos con que fue entrenado, las métricas obtenidas no entregan información utilizable para cuantificar la capacidad de generalización del modelo.

2.3.3. Métricas de clasificación

Las métricas de clasificación son fórmulas matemáticas que se aplican sobre los resultados de un modelo de clasificación. Las métricas sirven para determinar el rendimiento del clasificador. Para definir las métricas de clasificación, primero se debe introducir el esquema

fundamental de una matriz de confusión asociada al problema de clasificación binaria, la cual se observa en la figura 2.7:

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 2.7: Matriz de confusión genérica.

La observación es el resultado del examen y la predicción es el valor predicho por el clasificador. La matriz de confusión introduce conceptos de verdadero positivo (VP), falso positivo (FP), verdadero negativo (VN) y falso negativo (FN). Introducidos estos conceptos, en el problema de clasificación las métricas son:

Accuracy

El *accuracy* es una métrica de qué tan bien funciona el clasificador enfocándose en las predicciones tanto positivas como negativas. Lo importante de un clasificador es que cuando predice correctamente una clase, también debe predecir correctamente la otra. Se puede interpretar como una medida de qué tan verdaderas son las predicciones de un modelo. Matemáticamente, el *accuracy* se calcula como:

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

El **accuracy** es un número entre 0 y 1. El mejor caso es cuando es 1, en el cual, la predicción del clasificador siempre es verdadera, es decir, es el clasificador funciona de forma correcta. Un *accuracy* sobre 0.5 significa que el clasificador aprende a clasificar muestras y puede mejorar. Los algoritmos de *machine learning* en operaciones, y que han revolucionado diversas industrias, funcionan sobre el 0.9, es decir, tienen arriba de un 90 % de probabilidades de que su predicción sea verdadera. Este es un punto de referencia para identificar un buen modelo.

La interpretación del *accuracy* de un modelo depende de los datos con que fue entrenado. En este aspecto el modelo puede ser entrenado con datos balanceados y con datos no balanceados. Cuando se entrena con datos balanceados, un *accuracy* bajo 0.5 significa que nuestro clasificador funciona peor que lanzar una moneda al aire, es decir, el algoritmo clasifica aleatoriamente las muestras y, por lo tanto, se debe reentrenar y/o rediseñar su arquitectura.

Por otro lado, cuando se entrena con datos *no balanceados*, el *accuracy* falla al describir que tan bien está trabajando el modelo. Lo que sucede es que el clasificador tiende al *sesgo* de clasificar mejor la clase de la cual tiene más ejemplos. Para solucionar ese problema se utiliza la métrica del *F1-score*. Es por esto que es muy importante balancear los datos. Además de, realizar *class weighting* es decir balancear los pesos con los que el modelo aprende, asignando más peso a los datos con menor frecuencia de aparición en el dataset de entrenamiento.

Precision

La métrica de Precisión indica qué tan bien lo hace el modelo detectando casos de clase negativo. Se calcula como:

$$Precision = \frac{VP}{VP + FP} \tag{2.2}$$

Por lo que, igual que *accuracy*, se mueve en valores de entre 0 y 1. Siendo 1 el mejor desempeño para detectar correctamente las muestras de clase negativa.

Recall

Por último, la métrica recall se enfoca en los positivos. Indica qué tan bien el modelo lo hace para detectar las muestras que originalmente eran positivas. Al igual que las métricas anteriores es un número entre 0 y 1. Se calcula con:

$$Recall = \frac{VP}{VP + FN} \tag{2.3}$$

En el capítulo de discusión se realizará una extinción de la interpretación del recall y falso negativos (FN) para aplicaciones médicas.

2.3.4. Multi layer perceptron (MLP)

Una arquitectura básica del *aprendizaje supervisado* se llama MLP que significa *multi layer perceptron*. Una MLP puede ser usada para clasificaciones y regresiones. Cuando una MLP es usada para clasificaciones su modo de funcionamiento es que entran variables numéricas que definen una clase y sale un valor que representa la clase a la cual pertenece un ejemplo.

La figura 2.8 representa una referencia del orden en que se hacen los cálculos para lograr este objetivo. La primera etapa de cálculos es la capa de entrada donde ingresan las *features*,

que son variables independientes numéricas. La siguiente es la capa oculta que posee pesos que se ajustan para aproximar la función objetivo. Finalmente, la capa de salida que posee tantos nodos como clases se quieran clasificar. La matemática implícita, que se realiza por dentro, es que en realidad existe una función objetivo que toma por entrada las *features* y devuelve como salida las clases. Entonces, lo que hace una MLP es aproximar dicha función, ajustando sus parámetros usando el algoritmo de *backpropagation*.

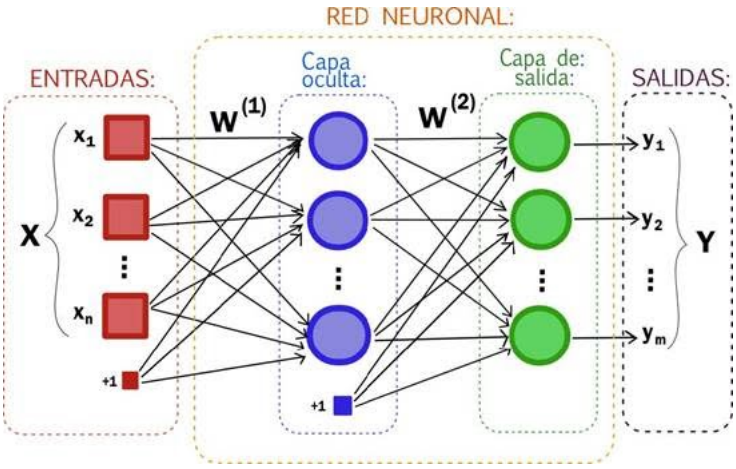


Figura 2.8: Esquema red *multi layer perceptron.*

En general, la arquitectura de *MLP* se puede usar para clasificar cualquier objeto del mundo real siempre y cuando se dispongan de características numéricas que describen ese objeto. Por ejemplo, si se quiere clasificar una imagen de un lunar, esta puede ser clasificada como un carcinoma (cáncer maligno) o una pigmentación de la piel inocua. Para clasificar la imagen del lunar se debería investigar qué características numéricas se deben extraer de la imagen para clasificar el lunar, por ejemplo, diámetro, forma o color. Estas características son denominadas *hand crafted*. Para luego, en base a esas características, entrenar la red.

Función de pérdida

En lo medular, **entrenar una red neuronal es resolver un problema de minimización con un método de optimización**. La función que se debe minimizar es la función de pérdida o *Loss function*. Básicamente, la función de pérdida es un función que expresa qué tan diferentes son los valores que predice la red con los valores reales que debería entregar, para un subconjunto de ejemplos (llamado *batch de entrenamiento*).

Una red neuronal funciona bien cuando clasifica correctamente la mayoría de los datos, lo que se observa cuando el valor de la función de pérdida es cercano a 0. Por otro lado, mientras más alto sea el valor de la función de pérdida peor estará clasificando la red. La ecuación 2.4 expresa la función de pérdida llamada *Binary Cross Entropy* ($Loss_{CE}$) que se

utiliza para el problema de clasificación binaria.

$$Loss_{CrossEntropy} = -\frac{1}{N}\sum_{i=1}^{N} y_i^{real}\log\Big(p(y_i^{real})\Big) + (1-y_i^{real})\log\Big(1-p(y_i^{real})\Big) \tag{2.4}$$

La entropía cruzada binaria se calcula como la entropía cruzada promedio en todos los ejemplos de datos en el *batch de entrenamiento*. En cada *batch de entrenamiento* se tienen N ejemplos de datos.

Una función de pérdida necesita la información que representa al valor predicho por la red y el valor real del ejemplo. Para cada ejemplo, la información que representa al valor predicho por la red y^{pred} está representada por $p(y_i^{pred})$, que es la probabilidad *softmax* para el i-ésimo ejemplo. La información que representa el valor real del ejemplo está expresada dentro de la ecuación 2.4 con el valor de y_i^{real}, este valor puede puede ser de 0 o 1, en el caso de clasificación binaria.

Por ejemplo, en el trabajo realizado en esta memoria. Uno de los clasificadores implementados es un clasificador binario entrenado con *batch de entrenamiento* de 32 parches. Entonces, cuando se entrena el clasificador, en la primera iteración de entrenamiento del algoritmo de *backpropagation*, de la primera época, se calcula el $Loss_{CE}$ usando la ecuación 2.5

$$Loss_{CrossEntropy} = -\frac{1}{32}\sum_{i=1}^{32} y_i^{real}\log_2(p(y_i^{real})) + (1-y_i^{real})\log_2(1-p(y_i^{real})) \tag{2.5}$$

Entonces, para el primer parche del *batch* suponiendo que sea de clase tumor, que se se representa como (1, 0), y que de la neurona *softmax* salgan las probabilidades (0.87, 0.13), se tiene que el *cross entropy* los para ese parche es:

$$Loss_{CE,i=1} = 1 * \log_2(0.87) + 0 * \log_2(1-0.87) = -0.2009 \tag{2.6}$$

Ya que los resultados de *cross entropy* para los ejemplos dan un números negativos, cuando se realiza el promedio sobre los 32 ejemplos, el promedio se multiplica por -1. De esta forma, la función de *loss* resulta en un valor positivo, esto permite que tenga sentido que cuando disminuya el *loss* mejore el rendimiento de la red.

2.3.5. Convolutional neural network (CNN)

El problema de la metodología planteada en la sección anterior es el alto costo de investigar cuáles *hand crafted features* sirven para cada tarea de clasificación específica. Este es el problema que resuelven las redes llamadas *convolutional neural networks*, o también, *CNN* por su sigla en inglés. Estas implementan un tipo especial de operación matemática llamada convolución.

Lo importante de una *CNN* es que, tal como se aprecia en la figura 2.9, a ella se ingresa directamente la imagen. Luego, en su proceso de entrenamiento, aprende a extraer las mejores features de la imagen para maximizar su rendimiento en la tarea de clasificación.

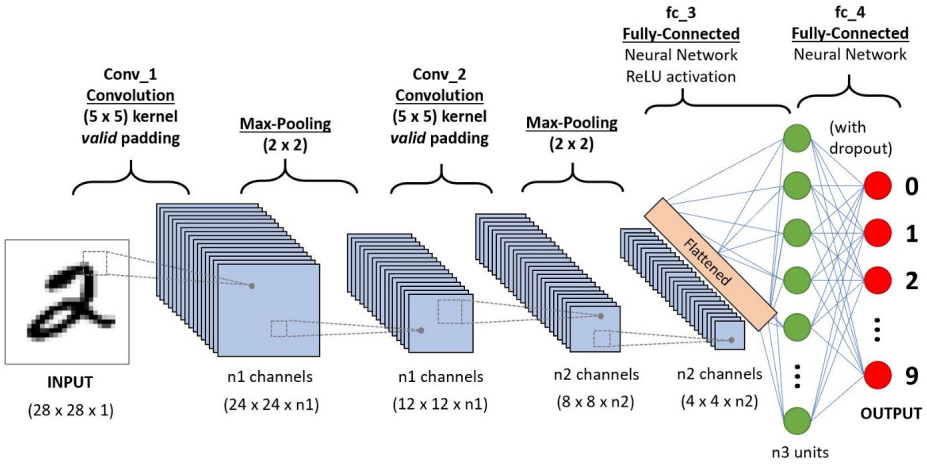


Figura 2.9: Esquema de una*convolutional neural network*

En la imagen 2.9 se observan distintas capas ocultas de procesamiento. Las capas que se observan son:

(1) Las **capas convolucionales** cuya función es implementar la operación de convolución, que en el fondo es un filtro, que se compone de neuronas que se activan con ciertos patrones de forma y color que son aprendidos a través del procesos de entrenamiento.

(2) Las **capas max pooling** tienen como función comprimir la información proveniente de la capa convolucional, con lo que se reduce la cantidad de parámetros que la red debe aprender en capas posteriores. Estas capas conforman la etapa de *feature extraction*.

(3) Las **capas fully conected** tienen la función de procesar la información, que ha sido extraída a través de las capas convolucionales y de *max pooling*, para generar la predicción de la clase de la imagen.

Capítulo 3

Estado del arte

Esta sección es producto de un análisis del campo de investigación en *machine learning* aplicado a la detección de cáncer gástrico y de mama usando imágenes médicas. Las primeras investigaciones se centran en detectar el cáncer de mama. Pasados los años, se avanza hacia la detección del cáncer gástrico. Es por eso que, para situarse en el estado del arte, es pertinente exponer sobre modelos de detección de cáncer de mama. A continuación se presentan descripciones de las investigaciones.

3.0.1. Detección de células de Vandenberghe

En este estudio titulado *"Relevance of deep learning to facilitate the diagnosis of HER2 status in brest cáncer"* [26] se describe el contexto del trabajo de los médicos patólogos, explicando que la detección de cáncer HER2+ es propensa a errores dependiendo del criterio de cada patólogo. Se expone sobre los inconvenientes en tiempo y dinero que genera realizar una nueva biopsia para una segunda opinión. Se utiliza una CNN para clasificar parches de biopsias de cáncer de mama. Se clasifica en medio extracelular, 0, 1+,2+ y 3+ (recordando sección guía clínica). Se inicia utilizando umbralización, se divide la imagen de la biopsia en parches, a los cuales aplica deconvolución para separarlos en 3 imágenes. La primera de ellas representa la tinción de hematoxilina que contiene la información para identificar los núcleos celulares. La segunda imagen, de tinción marrón, posee la información que caracteriza a las membranas celulares HER2 positivas. La tercera imagen es una combinación lineal de estas dos últimas imágenes para obtener una sola imagen, que es resultado meramente matemático que no captura información de la realidad, por lo que la tercera imagen obtenida de la deconvolución no se utiliza. Para comenzar a entrenar, se etiquetan manualmente los parches que poseen mayoritariamente una sola clase de sobreexpresión. Se extraen *handcrafted features* biológicamente relevantes que describen el color, textura y forma del núcleo; y para la membrana celular se extrae la intensidad de la tensión de HER2 y la proporción de la membrana que es HER2 positiva. Con estas features se infieren modelos *support vector machine (SVM)* y *Random Forest*. Con los parches etiquetados manualmente se entrena una *CNN* de 3 capas

convoluciones. Esta última es la que logra los mejores resultados.

El aporte de esta investigación es exponer la forma en que este tipo de sistemas se integra en un rol de apoyo a la labor de los patólogos y no en reemplazo. Esto es importante, porque en la mayoría de otras investigaciones se limitan a indicar que las biopsias son costosas, susceptibles a errores y que tardan mucho tiempo. También, es un aporte en el sentido de que explícita *handcrafted features* que van a servir para generar una base de datos de máscaras de células HER2+ de cáncer estomacal.

Este estudio no soluciona el problema de esta memoria porque se trabaja sobre células de cáncer de mama y no es evidente que las células de cáncer gástrico tengan la misma distribución; en el sentido de los clusters que forman. Ni que la tinción observada sea de las mismas intensidades. También, en este trabajo se realiza etiquetado manual lo cual es inviable por el tiempo que tomaría debido a la alta cantidad de parches.

3.0.2. U-NET

Las redes de la familia U-NET [27] son arquitecturas tipo encoder-decoder sirven para segmentación semántica de imágenes biomédicas. Fueron desarrolladas por Ronneberger, Fischer y Brox [27] en el año 2015. Las arquitecturas de estas redes tienen codificadas como conocimiento las formas y texturas que presentan a las células en general en imágenes biomédicas. El avance que implementan es que la información codificada de cada nivel, de la etapa del encoder, se usa para entrenar ese mismo nivel en la etapa del decoder.

La tarea de segmentación semántica consiste en determinar a qué clase pertenece cada píxel de una imagen. La arquitectura que realiza esta tarea se llama *Autoencoder*, o también, *encoder-decoder*. El esquema general de este tipo de red se aprecia en la figura 3.1. La entrada de esta red es la imagen y la salida es otra imagen en que cada píxel es asignado a un color que representa la clase al cual pertenece. La primera parte de la red se llama *encoder*, la cual codifica los *features* aprendidos de la imagen a un *espacio latente* de menor dimensionalidad.

La segunda parte de esta red implementa un tipo de operación llamado *up-sampling* que sirve para 'descomprimir' (aumentar de tamaño) la información almacenada en los features que aprende la parte convolucional de la red (el *encoder*). A través del entrenamiento se ajustan los parámetros de las capas de *up-sampling*, con el objetivo de generar una imagen de las mismas dimensiones que la original pero coloreando específicamente cada píxel con la clase a la que pertenece según los ejemplos con que se entrenó.

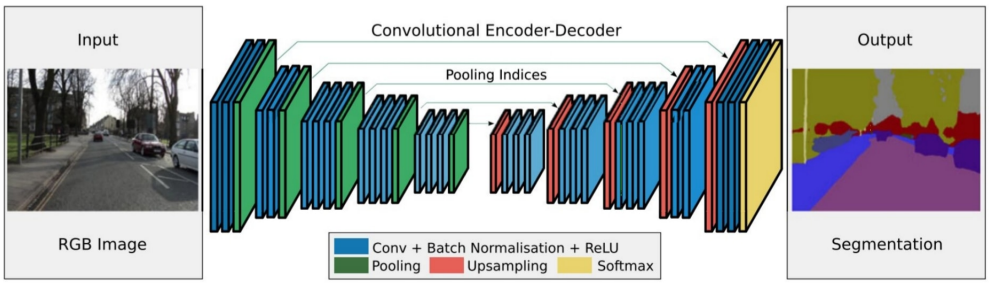


Figura 3.1: Esquema de una red *encoder-decoder*

3.0.3. HER2net

La red HER2net [28] por Saha y Chakraborty en el año 2018. También implementa una estructura encoder-decoder. La innovación que se implementa son los módulos trapezoidal LSTM al final de la etapa convolutiva (encoder) y al principio de la etapa deconvolutiva (decoder). Estos módulos agregan una suerte de memoria al modelo.

Esta arquitectura es mucho más sofisticada que U-NET. HER2net es entrenada con imágenes con tinción IHC de tejidos de biopsias de cáncer de mama. Esto es un ejemplo que le da validez la utilización de deep learning en imágenes IHC.

3.0.4. Clasificación parches cáncer mamario HER2+ de Pitkäaho

En el articulo *"Classifying HER2 breast cancer cell samples using deep learning"* por Pitkäaho *et al* [29] del año 2016. Se utiliza una *convolutional neural network (CNN)* para clasificar parches de biopsias de cáncer de mama. Las imágenes provienen del dataset *HER2 Scoring Contest* de la University of Warwick [5]. Este es un *benchmark* lanzado en el año 2016 con el fin de acelerar el desarrollo de algoritmos para la automatización de la clasificación de sobreexpresión de HER2 en imágenes biomédicas.

La metodología utilizada en Pitkäaho *et al* [29] se basa en la experiencia de los investigadores para seleccionar manualmente secciones de las imágenes que mejor representen la clase de la imagen. Estas se dividen en parches de 128x128 píxeles. Los parches poseen una o más células. Cada parche se clasifica en 5 clases: fondo blanco sin textura, 0, 1+, 2+, 3+. Se usan 319.032 parches en la etapa de entrenamiento. Se alcanza un 97.7 % de *accuracy*. Se concluye indicando que el modelo desarrollado podría ser usado para clasificar una biopsia completa, basándose en la guía clínica.

3.0.5. Clasificación de biopsias con Inception V3 de Alegría

El presente trabajo de memoria es la continuación del trabajo de tesis de magíster de Alegría [11]. En dicho trabajo se desarrolla un sistema de diagnóstico de cáncer gástrico HER2 usando el enfoque de clasificación entrenando una arquitectura InceptionV3. Se utiliza la misma base de datos del estudio PRECISO y se desarrolla un sistema de diagnóstico basado en la clasificación de los parches en que fue dividida la imagen original de la biopsia.

Este trabajo constituye una línea base desde la cual comparar los resultados que se obtendrán en esta memoria. Aporta con el marco general de identificación y justificación de la problemática del cáncer gástrico. Hace un análisis de la concordancia de los resultados de diagnóstico de dos patólogos comprobando la existencia de disimilitudes para el diagnóstico de las clases más difíciles que son 1+ y 2+.

Si bien es cierto, este trabajo soluciona el problema planteado, asume el supuesto de que todas las células de una anotación son de la misma clase. Esta es una solución gruesa del problema. Lo que sucede es que cuando se divide la imagen de la biopsia en parches cuadrados, cada parche pertenece a una clase. Luego, se hace el supuesto de que todas las células que este contiene, presentan la misma sobreexpresión de HER2. Este supuesto no siempre es cierto.

En suma, se está aplicando tecnología de un problema ampliamente abordado que es el cáncer de mama a un problema con pocas investigaciones como el cáncer gástrico.

Capítulo 4

Origen de los datos y Línea base

4.1. Estudio PRECISO

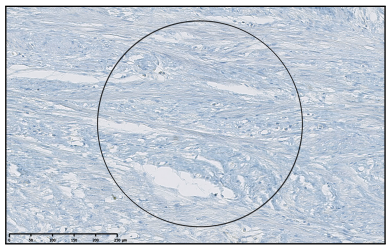
El origen de los datos es el estudio PRECISO (NCT01633203) [30], que significa *Prospective Observational Study of Patients With Locally Advanced Gastric Cancer Treated With Perioperative Chemotherapy and Surgery*. El objetivo de este estudio es evaluar la eficacia y toxicidad de la quimioterapia perioperatoria con Epirubicin + Cisplatin + Capecitabina (ECX) en la práctica clínica de rutina en una red de hospitales públicos en Santiago de Chile. El estudio se realizó con 61 pacientes, entre los años 2012 y 2020. De los cuales sólo 48 autorizaron la determinación de la sobreexpresión de la proteína HER2 en sus muestras. Todas las imágenes generadas son anonimizadas.

Las imágenes se obtienen utilizando el escáner de placas histológicas *Nanozoomer XR* de la empresa japonesa *Hamamatsu Photonics*. Las imágenes son almacenadas en formato especial de esta empresa llamado *ndpi*. Son imágenes del tamaño aproximado de 40 GB guardadas a una magnificación máxima de 40x.

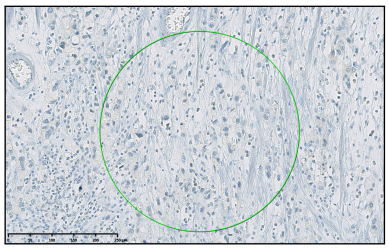
El estudio cuenta con imágenes con tinción H&E (usadas para detectar la presencia de cáncer) y con tinción IHC. En total 3 patólogos realizan anotaciones sobre las imágenes del estudio. Sin embargo, para esta memoria las anotaciones que se utilizan son las del tercer patólogo. El patólogo 3 realiza anotaciones de las regiones de interés en cada muestra y además realiza el etiquetado de la muestra global utilizando la guía clínica. El patólogo 3 solamente realiza anotaciones y clasifica 34 pacientes. Se cuenta con imágenes de 34 pacientes correspondientes a sus biopsias por endoscopía y de sus biopsias por resección. Finalmente, en esta memoria se trabajó con las 34 biopsias por resección; las cuales poseen una etiqueta global a nivel de tejido y anotaciones de sobreexpresiones HER2 particulares en regiones de interés.

Ahondando en la identificación de regiones de interés, los patólogos observan la imagen en

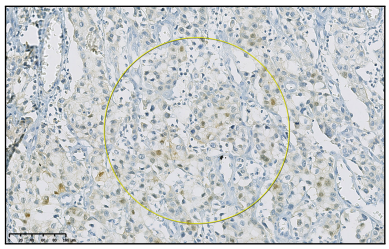
busca de regiones de células con reactividad local particular. La determinación de la región celular con reactividad particular es en base al criterio experto. Una vez identificada la región, el patólogo marca manualmente la región circular que contiene esa región celular y la etiqueta con la reactividad representativa de sus células. La etiqueta es asignada en base al criterio del médico patólogo. El criterio del médico patólogo es producto de su formación profesional, su experiencia y en base a la información que continuamente adquieren de investigaciones científicas. Los patólogos buscan grupos de células que posean la reactividad. Los patólogos no trabajan observando célula a célula. El mínimo objeto sobre el que él patólogo decide no es una célula, sino que es una región de células. La información que caracteriza el grupo de células se observa a nivel de grupo no a nivel individual de células. En la siguiente figura 4.1 se observan ejemplos de las distintas sobreexpresiones que identifica el patólogo en zonas de interés circulares.



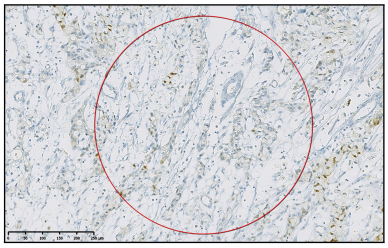
(a) No tumor - nt



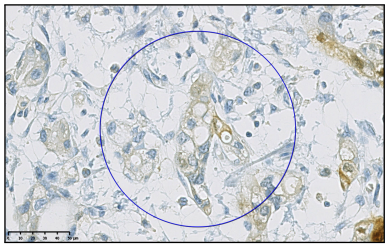
(b) Sin reactividad - sr



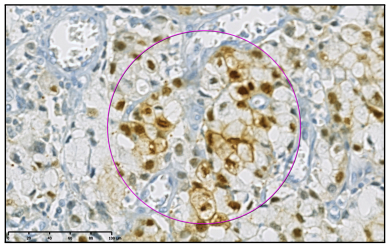
(c) Reactividad no lineal - rnl



(d) Reactividad lineal casi imperceptible - rlci



(e) Reactividad lineal debil - rld



(f) Reactividad lineal fuerte - rlf

Figura 4.1: Diferentes etiquetas para sobreexpresión de HER2

Recordando lo expuesto en la sección pasada, el color rojo/marrón representa las proteínas HER2. La intuición simple es que una coloración marrón intensa está asociada a la sobreexpresión de proteína HER2 en esa vecindad celular. La coloración azul está asociada a los núcleos celulares. Finalmente, las zonas blancas son el medio extracelular.

En síntesis, para cada muestra de tejido, el patólogo la analiza detectando regiones de interés y etiquetando su reactividad. Una vez identificadas las regiones de interés se procede a hacer el etiquetado de la muestra global, en el que la etiqueta de reactividad de cada región de interés se traduce a una puntuación y clasificación HER2. Esta traducción se realiza utilizando la tabla 4.1. Luego considerando las puntuaciones obtenidas se clasifica globalmente la muestra usando la guía clínica. Por ejemplo, para una muestra en que solo se detecten

regiones de interés de “sin reactividad”, luego todas estas son traducidas a puntuación negativa (0). Luego, considerando todas las regiones de interés según la guía clínica se clasifica globalmente la muestra como negativa y el diagnóstico es negativo.

A modo de ejemplo, se explica cómo se aplica la guía clínica para un caso positivo. Considérese el caso de una muestra en la cual se hayan detectado regiones de interés con todas las posibles clases de reactividad, entonces, según la guía clínica, si se tienen más del 10 % de las células con clase reactividad lineal fuerte, la muestra pasa a ser clase positiva (3+), sin importar las reactividades de las otras regiones de interés detectadas. Esto se conoce como la regla del umbral del 10 %.

Tabla 4.1: Pauta de puntuación para reactividades de regiones de interés

Etiqueta HER2	Clasificación HER2	Nro ROI
No Tumor - nt	No aplica	52
Sin reactividad - sr	0 Negativo	50
Reatividad no lineal - rnl	0 Negativo	37
Reactividad lineal casi imperceptible - rlci	1+ Negativo	33
Reactividad lineal debil - rld	2+ Equivoco	33
Reactividad lineal fuerte - rlf	3+ Positivo	20

En la tabla 4.1, en la columna Nro ROI se observa la cantidad de anotaciones que se tienen para cada etiqueta. Por otro lado, para la etiqueta NO tumor, se informa que **no aplica** una puntuación ni una clasificación HER2. Esto significa que el patólogo determina que las regiones de interés etiquetadas como No Tumor, no poseen células tumorales. Se aclara que, finalmente, en la práctica se puntúan como negativo. Como se aprecia en la tabla 4.1 se trata de un problema desbalanceado (3+ tiene 20 ROIs, y 0 Negativo tiene 87 ROIs), pero no en extremo.

Cabe destacar que un paciente puede padecer cáncer gástrico pero no ser de HER2 positivo. El cáncer HER2 positivo es un subtipo de cáncer, una persona puede ser HER2 negativa y pese a eso tener cáncer gástrico de otro tipo. Por ejemplo, los parches con la etiqueta sin reactividad son parches en los cuales los patólogos determinan que existen células tumorales cancerígenas, aunque no asociadas a la sobreexpresión de HER2.

4.2. Algoritmo de Línea base

Como algoritmo de comparación se selecciona el algoritmo de Alegría [11] el cual se encuentra bien detallado en su implementación. Sin embargo, al ser un algoritmo extenso, en esta sección se detallan tareas o elementos que no son evidentes desde el documento de tesis

donde se presenta dicho método.

Formato de imágenes utilizado

Las imágenes de las biopsias son archivos grandes del orden de los 40 GB (al descomprimir) por lo que leerlos todos en la memoria RAM es inviable. Por esto, las imágenes se guardan en un tipo de archivo llamado *ndpi*. Esto sirve para cargar en memoria solo una parte de la imagen con un nivel específico de acercamiento mediante una estructura *piramidal multi-tile*. Los patólogos acceden a estos archivos utilizando el *software NDP.view2*. En él, visualizan las imágenes con acercamiento desde 1x hasta 40x y realizan anotaciones identificando visualmente las células con sobreexpresión de HER2.

A la fecha no existen librerías que permitan realizar procesamiento de imágenes programando en *Python* sobre las imágenes en formato *ndpi*. Por eso, se utiliza el *software ndpisplit*, el cual se ejecuta usando la terminal de *Linux*, para transformar las imágenes a formato *TIFF*. Usando *ndpisplit* se generan 7 imágenes, una para cada escala de magnificación con que se dispone, que en este caso son: 40x, 20x, 10x, 5x, 2.5x, 1.25x y 0.625x. La imagen con magnificación 40x es la imagen original. Las imágenes con magnificación 2.5x, 1.25x y 0.625x son las copias de cada vez más baja resolución (y con ello menos peso). Son útiles para visualizar a escala los resultados que se van obteniendo en el proceso.

Formato de etiquetas

Las etiquetas realizadas por el patólogo son las clases de las regiones de interés. Cuando el patólogo las dibuja en la biopsia, el *software NDP.view2* genera un archivo de tipo *.ndpa*. Este es un tipo de archivo específico desarrollado por la compañía japonesa *Hamamatsu Photonics*. Los archivos *ndpa* implementan la estructura de datos de árboles en formato *XML*. El detalle del formato *.ndpa* se presenta en el Anexo A.

Transformación sistema coordenadas ndpa a pixeles

Desde los archivos *ndpa*, se obtienen las coordenadas del centro y el radio de las ROI. Las cuales están en el sistema de coordenadas de *ndpi*. Se necesita pasarlo al sistema de coordenadas de *OpenCV*, en el cual, el origen está en la esquina superior izquierda de la imagen. Para acceder a un píxel, la primera componente es la fila (que sería el eje x) y la segunda, la columna del píxel que sería el eje y.

Se necesita una función que transforme el sistema de coordenadas de *ndpi* al de *OpenCV*. Primero hay que partir conociendo el sistema *ndpi*. Se comienza buscando manuales de fun-

cionamiento, pero lo que se encuentra son manuales de usuario enfocados en uso del *software* a nivel de usuario. No se logra encontrar documentación del código del *software* o explicaciones más detalladas de cómo se guardan los datos. Dado que no existe documentación se procede a usar puntos de referencia en la imagen, como se muestra en la figura 4.2 y de esa manera calcular la transformación de coordenadas.

Figura 4.2: Pins colocados para entender sistema de coordenadas

La transformación de un sistema de coordenadas a otro se define entonces como,

$$x_{pixeles} = \lfloor \left(\frac{x_{ndpi} - x_{offset}}{1000 * mpp_x} + \frac{widthl_0}{2} \right) * \frac{10}{40} \rfloor, \tag{4.1}$$

$$x_{pixeles} = \lfloor \left(\frac{y_{ndpi} - y_{offset}}{1000 * mpp_y} + \frac{lengthl_0}{2} \right) * \frac{10}{40} \rfloor, \tag{4.2}$$

$$radio_{pixeles} = \lfloor \frac{radio_{ndpi}}{227} \rfloor. \tag{4.3}$$

El detalle del proceso de obtención de la conversión anterior se muestra en el Anexo B. Además, se detallan los significados de las variables y las librerías con que se obtienen esos datos desde los archivos *ndpi*.

Manejo de imágenes de gran tamaño

Como se mencionó, las imágenes utilizadas en su conjunto no se pueden cargar en memoria principal por su tamaño. Tampoco las librerías de *Python* dan soporte el formato *ndpi*. Por esto se trabaja con la imagen en partes, los cuales se generan con un paquete llamado

NDPISplit. *NDPIsplit* [31] es un *software open source* desarrollado por el *Modelling Team del IMNC laboratory* en París, lanzado el año 2013. Convierte archivos *ndpi* en archivos *tiff*, además, puede cortar partes del archivo *ndpi* en imágenes pequeñas en formato *jpeg* o *tiff*.

El comando presentado en el código 3.1 muestra un ejemplo de la redacción de un comando para cortar un parche.

Código 4.1: Ejemplo extracción de parche con ndpisplit.

```
1 ndpisplit -Ex40,z0,120541,12171,299,299,parche-a0f0c0 -K ejemplo.ndpi
```

Al ejecutar este comando *NDPIsplit* se posiciona en la magnificación 40x, con z-offset de 0, del archivo *ejemplo.ndpi* recorta un parche cuya esquina superior derecha está en el punto (120541, 12171), del sistema de coordenadas *ndpi*. El parche tiene un ancho de 299 píxeles y un alto de 299 píxeles. El archivo final se llamará: *ejemplo-x40-z0-parche-a0f0c0* . Con la flag -K se imprime en la terminal los subprocesos que va ejecutando el comando. Para utilizar *NDPIspit*, se complementa con un script en *Bash* que se detalla en el Anexo C.

Uso de GPU

Un entrenamiento con todo el conjunto de datos PRECISO en un computador personal (CPU 2 GHZ 4Gb RAM) puede llegar a tomar cerca de dos semanas. Debido a esto para hacer pruebas sistemáticas es necesario utilizar una *GPU*. En este trabajo se utilizó el *cluster* de *GPU* disponibles en el cluster de cómputo *Patogón* [13]. Este *cluster* cuenta con 8 tarjetas *Nvidia A100* de 40 Gb. Usando una *GPU Nvidia A100* de 40 Gb es posible entrenar el modelo en 8 horas. Este *cluster* de cómputo utiliza una arquitectura basada en *Dockers* y una cola de prioridad *Slurm*.

Implementación de método de línea base

El modelo de Alegría [11] se basa en una red neuronal profunda conocida como *InceptionV3*. Sin embargo, no es evidente cómo se inicializan los pesos de las diferentes capas de la red entre corridas de entrenamiento del *LOOCV*. Luego de experimentar se definió que en la primera corrida los pesos se inicializan con los pesos de *InceptionV3* entrenada en *ImageNet* [12]. A continuación, se repite el proceso para cada una de las corridas en el *loocv*.

Data Augmentation y Desbalance de los datos

Inicialmente, se cuenta con 3.987 imágenes provenientes de recortar las regiones de interés en parches cuadrados. Sobre estos se aplican transformaciones aleatorias como rotación (en

rango de $\pm 10^{o}$), traslación horizontal y vertical (en rango 10 % ancho y alto de imagen), zoom de acercamiento y alejamiento, en rango de hasta 10 % [3], reflexión horizontal y vertical. Esto se realiza usando el módulo de data augmentation de la librería *KERAS* llamado *ImageDataGenerator*. Lo destacable es que *KERAS* permite hacer el data augmentation como parte del pipeline de procesamiento. Las imágenes aumentadas son generadas solamente para el contexto del entrenamiento. Luego, no es necesario guardar en el disco duro las imágenes aumentadas, sino que solamente la base de datos original.

Al ser un problema desbalanceado, se utilizó el paquete *sklearn class weight* que **pondera inversamente proporcional a la frecuencia de los ejemplos**.

[3] Para futuras continuaciones se recomienda no usar la aumentación por zoom porque dichos datos estarán enseñando al modelo a reconocer patrones en una magnificación diferente a la que el modelo usará en producción. Lo cual no es directamente negativo para el rendimiento. Sino que, solo agrega información que no es la requerida, es decir, no agrega información de mejor calidad para esta tarea en particular

Capítulo 5

Algoritmo propuesto

En este capítulo se presenta el algoritmo propuesto para mejorar el rendimiento del modelo de Alegría [11]. Fundamentalmente, el algoritmo trata de filtrar el dataset en que originalmente se entrena la red de Alegría, en base a los parches con etiquetado de mayor calidad. El proceso de filtrado se detalla a continuación.

El algoritmo propuesto se compone de dos etapas principales: (1) Creación del filtro y (2) Aplicación del filtro. La etapa de creación del filtro se explica en la sección de 4.1 llamada estimación de centroides de colores por clase. La etapa de aplicación del filtro se explica en la sección 4.2 denominada filtrado de parches por distancia a centroide.

5.1. Estimación de centroides de colores por clase

Como se ilustra en la figura 5.1 desde el dataset *PRECISO* se toman las imágenes en formato *ndpi* y las anotaciones en formato *ndpa* para generar parches. A continuación, se propone un algoritmo que generará un color característico para cada clase (sr, rnl, rlci, rld, rlf), las cuales servirán luego para filtrar los parches según su calidad de etiquetado. Al color característico de una clase se le llama centroide. Las etapas del algoritmo propuesto se explican a continuación.

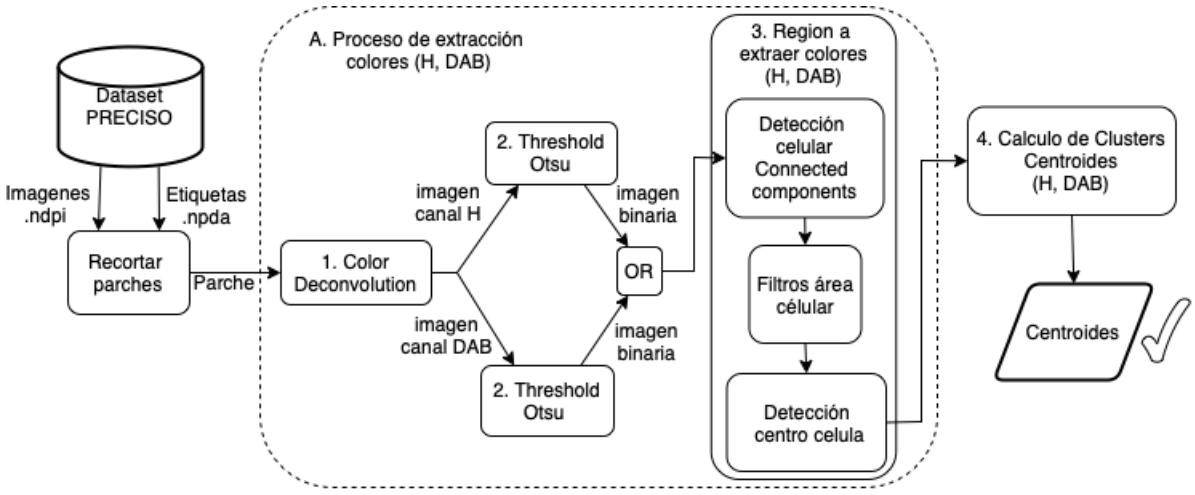


Figura 5.1: Diagrama de creación de filtro para obtener centroides de colores.

5.1.1. Deconvolución de color

Las imágenes IHC o HE son adquiridas con microscopios en 3 canales, R: Red, G: Green, B: Blue (I_R, I_G, I_B). Sin embargo, las resecciones utilizadas son tinciones de *hematoxilina* y *DAB*, donde ambos compuestos no son colores puros en RGB. Para simplificar la segmentación, se transforma el espacio de colores tal que un canal se asocie a núcleos celulares(I_H), otro a la membrana teñida reactiva HER2 (I_{DAB}) y, un tercero, a la eosina (I_E). La transformación se puede definir como una operación matricial,

$$I_{HDAB} = I_{RGB}D$$

donde la matriz D se llama matriz de deconvolución de color, que ya se ha establecido para la tinción H-DAB en Ruifrok [32]. En este trabajo solo se ocupan los colores H y DAB que dan lugar a dos imágenes de un canal cada una.

5.1.2. Segmentación basada en umbral de Otsu

En esta parte del proceso se requiere segmentar entre el fondo de la imagen y las células, que incluye el núcleo y la membrana. Esto se realiza segmentado en el canal H (núcleos) y en el canal DAB (membranas). El algoritmo de Otsu logra esta segmentación mediante un umbral único por imagen.

El algoritmo de Otsu [33] asume que existen dos cluster, uno del objeto de interés (células) y el otro es el fondo de la imagen. Mediante un método iterativo, dicho algoritmo minimiza

la varianza intra-clase definida por:

$$\begin{aligned} \sigma_w^2 &= w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t), \\ w_0(t) &= \sum_{i=0}^{t-1} p(i), \\ w_1(t) &= \sum_{i=t}^{L-1} p(i), \end{aligned}$$

donde $p(i)$ es la frecuencia de la intensidad de píxel i en la imagen, t es el umbral de intensidad buscado, $\sigma_0(t)$ es la varianza de las intensidades hasta la intensidad $t-1$, y $\sigma_1(t)$ es la varianza de las intensidades desde la intensidad t.

5.1.3. Identificación de región celular

El objetivo de esta etapa es identificar las regiones de donde se van a extraer los colores. Con este objetivo se combinan las segmentaciones de los canales H y DAB mediante un OR lógico. A continuación, se detectan los objetos mediante la función *connected components*. Como se muestra en la figura 5.1, los objetos detectados se filtran por área mínima de 250 pixeles y área máxima de 1250 pixeles, que corresponde en promedio al tamaño de una célula. Los objetos resultantes representan células, es decir, el núcleo y la membrana de una célula. Luego se detecta el centro de cada célula y se genera una ventana circular de 10 píxeles de radio en torno al centro. Esta ventana circular es la región de la cual se extraen los colores. Este proceso genera las coordenadas de las cuales se extrae el color H de la imagen de canal H y el color DAB de la imagen de canal DAB. Concretamente, se extrae πr^2 (314 puntos) de dos colores (H, DAB) por cada célula de cada parche.

5.1.4. Estimación Centroides de colores

El objetivo de esta etapa es obtener un color representativo asociado a cada clase de reactividad HER2. Para ello se agrupan todas las ventanas celulares asociadas a parches de una cierta clase. Asumiendo, que los parches tendrán en su mayoría células de la clase indicada por el patólogo, los puntos de las ventanas celulares agrupados formarán un cluster mayoritario que representará el color de la clase en el espacio H-DAB.

Entonces, el color asociado a una clase es un punto en el espacio H-DAB. Este punto es obtenido promediando todos los puntos H-DAB de una clase. Finalmente, los centroides obtenidos son 5 puntos H-DAB, llamados c_i, donde cada punto representa una clase de reactividad.

5.2. Filtrado de parches por distancia a centroides

Esta sección describe los procesos de la etapa de aplicación del filtro. Esta etapa comienza con el dataset PRECISO y termina con un dataset filtrado compuesto por parches etiquetados con mayor calidad. Los procesos descritos a continuación se aplican sobre un parche en cada iteración de esta etapa. Los procesos se describen a continuación:

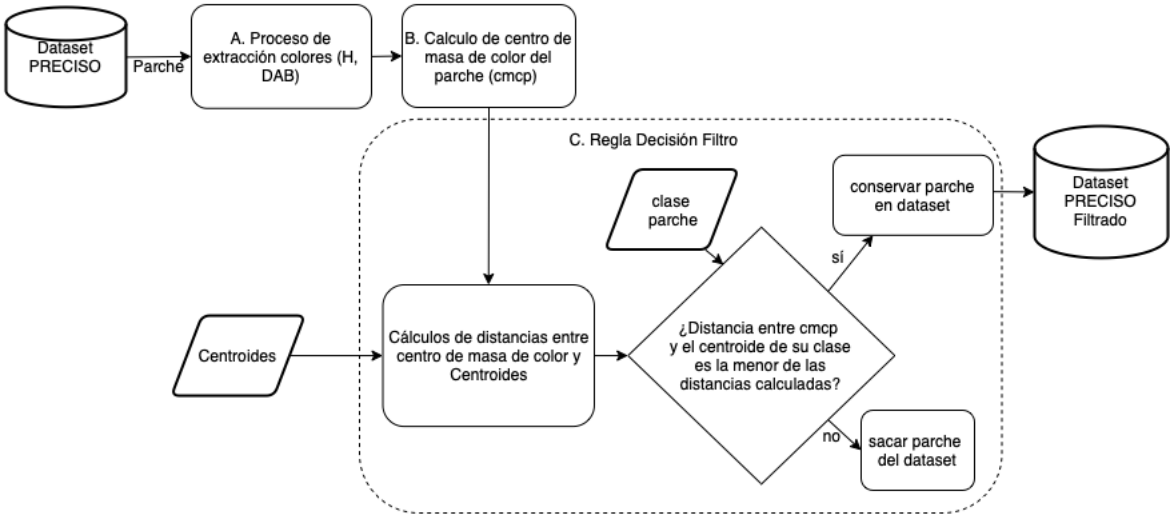


Figura 5.2: Diagrama aplicación del filtro para generar dataset de mayor calidad.

5.2.1. Calculo de color representativo de las regiones celulares del parche (B)

El objetivo de este proceso es definir un color representativo para cada parche, es decir, un punto en el espacio H-DAB que represente la muestra. Para comenzar, se utiliza el mismo procedimiento detallado en 4.1.3, donde se toman los colores H-DAB asociados a cada célula detectada, generando una nube de puntos en el espacio H-DAB. Luego, se calcula el promedio de todos los puntos que se denomina p, el cual es finalmente el punto o color representativo del parche.

5.2.2. Regla de decisión de filtro (C)

En la figura 5.2 se observa la sección D, a la cual ingresan cinco distancias calculadas $(d_1, d_2, d_3, d_4, d_5)$ entre el color representativo del parche p y el centroide estimado que repre-

senta cada clase de sobreexpresión $(c_1, c_2, c_3, c_4, c_5)$ es decir,

$$d_i \quad = \quad |p - c_i| \, .$$

Por otra parte se conoce la clase a la cual pertenece el parche clase(p) que fue indicado por el patólogo. Luego, si la clase del parche es la misma que la asociada a la distancia más corta, el parche se mantiene en el dataset, de lo contrario se elimina del dataset.

El supuesto realizado es que una parche que está correctamente clasificado es aquel en que su distancia entre su color representativo y el centroide del cluster de su clase, es la menor de todas las distancias calculadas a los centroides representantes de las clases. Es decir, los parches que se mantienen cumplen con clase$(\min d_i) = c_p$.

A modo de resumen, una vez procesados todos los parches, se finaliza con un dataset filtrado. Los parches pertenecientes del dataset filtrado tienen anotaciones con mayor calidad. Se espera que reentrenando los modelos con este dataset filtrado, estos mejoren su rendimiento.

Capítulo 6

Resultados y Discusión

En este capítulo se utilizan 34 modelos distintos bajo la partición explicada en los antecedentes en la sección de *LOOCV*. En resumen, significa que si se tienen 34 muestras de pacientes se van a entrenar y evaluar 34 modelos. El primer modelo se entrena tomando los parches provenientes de un paciente, estos se asignan como conjunto de validación. Luego, como conjunto de entrenamiento se toman los parches de las otras 33 muestras, ya sea para el problema de 2 clases (Tumor/No tumor) o de 5 clases de reactividad.

Los experimentos realizados en este capítulo se entrenan y evalúan por 60 épocas. Este proceso se denomina una *"corrida"*. Cuando se itera de una corrida a otra, se borra la red entrenada y se crea otra red en blanco (partiendo de los pesos de *InceptionV3* entrenada en *Imagenet*). Para todos estos experimentos se miden las métricas de *accuracy*, *loss*, *precision* y *recall*.

6.1. Resultados

6.1.1. Clasificación Tumor/NoTumor

El objetivo de este experimento es replicar en detalle el experimento de clasificación de parches de células tumorales y no tumorales, primero propuesto por Alegría [11]. En la figura 6.1 se observan los resultados del entrenamiento del modelo biclase, es decir, el modelo que discrimina si en una imagen hay células tumorales o no hay células tumorales.

En la figura 6.1 se observa una rápida velocidad de convergencia de la arquitectura *InceptionV3*, llegando a un *accuracy* de validación entorno al 0.8 en 5 épocas de entrenamiento. Posterior a la época 5, se observa que la velocidad de aumento de *accuracy* de validación disminuye. El *accuracy* de entrenamiento de los modelos se acerca asintóticamente a 1 lo que es lo esperado para un modelo de *machine learning*. Esto es un éxito porque significa

que la arquitectura con que se experimenta realmente está aprendiendo a realizar la tarea de clasificación. Sin embargo, dado que es el *accuracy* de entrenamiento, mientras más épocas se entrene, el modelo entra más en sobreajuste y va perdiendo su capacidad de generalización. La línea punteada verde representa el *accuracy* de validación reportado por Alegría [11] de 0.88 en la tarea de clasificación Tumor/NoTumor. Al observar el *accuracy* de validación obtenido en este trabajo se observa que a partir de la época 30 se estabiliza en torno al 0.86. Con esto se confirma que en un rango de 2 % se logra replicar una parte del trabajo de Alegría [11].

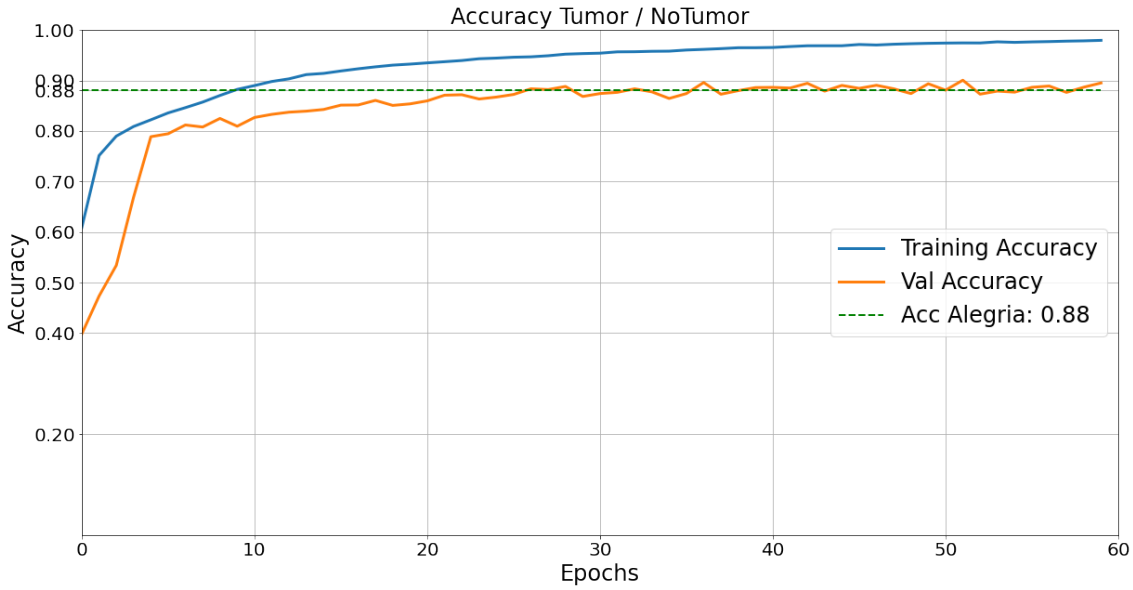


Figura 6.1: Promedio de la evolución del *Acurracy* en entrenamiento y evaluación de 34 modelos de clasificador de Tumor/No-Tumor

La figura 6.1 muestra que el *accuracy* de entrenamiento y validación se estabilizan después de las 30 épocas, aproximadamente. Esto significa que los modelos entrenados efectivamente realizan satisfactoriamente la tarea de clasificación. Sin embargo, a partir de cierta época de entrenamiento los modelos dejan de aprender patrones que les permitan mejorar su rendimiento. Este comportamiento se puede apreciar mejor en la figura 6.2 que muestra la evolución de las métricas *loss* durante las épocas.

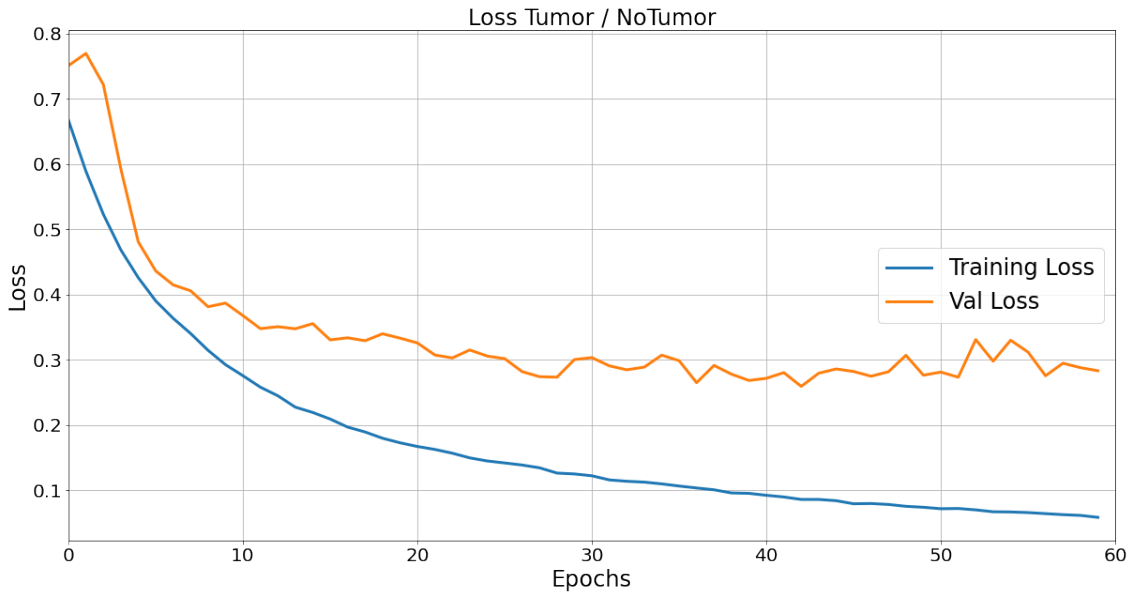


Figura 6.2: Promedio de la evolución del *loss* en entrenamiento y evaluación de 34 modelos de clasificador de Tumor/No-Tumor

La figura 6.2 también es creada calculando el promedio de *loss* para los 34 modelos durante 60 épocas. Se aprecia que el *loss* de entrenamiento disminuye exponencialmente lo cual es esperable. La curva de *loss* de validación disminuye hasta las épocas 30-40, luego comienza a subir. Lo que significa que entre 30 y 40 épocas en adelante el modelo se sobreajusta.

6.1.2. Clasificación HER2 (5 clases)

El objetivo de este experimento es nuevamente replicar en detalle el experimento de clasificación de reactividad HER2 (5 clases) propuesto por Alegría [11]. En la figura 6.3 se observan los resultados del entrenamiento del modelo de 5 clases, es decir, el modelo que discrimina los diferentes grados de reactividad HER2.

Las figuras 6.3 y 6.4 y son calculadas bajo el concepto de *mean per epoch*, es decir, teniendo los resultados de *accuracy* y *loss* para cada época de cada uno de los modelos entrenados. Lo que se realiza para crear las figuras es que, para cada época, se toman los *accuracy* de los modelos en esa época, y luego, se promedian. Es por eso que cuando se menciona *accuracy* o *loss* realmente se está refiriendo a *accuracy* promedio y *loss* promedio.

En la figura 6.3 se observa la evolución del *accuracy* promedio en los modelos entrenados. En estas curvas se puede apreciar que ya a las 10 épocas el *accuracy* promedio en el conjunto de validación se estabiliza en 0.5 sin decaer posteriormente. En la misma figura 6.3 se indica el rendimiento reportado en Alegría [11] el que es superior en un 11 %.

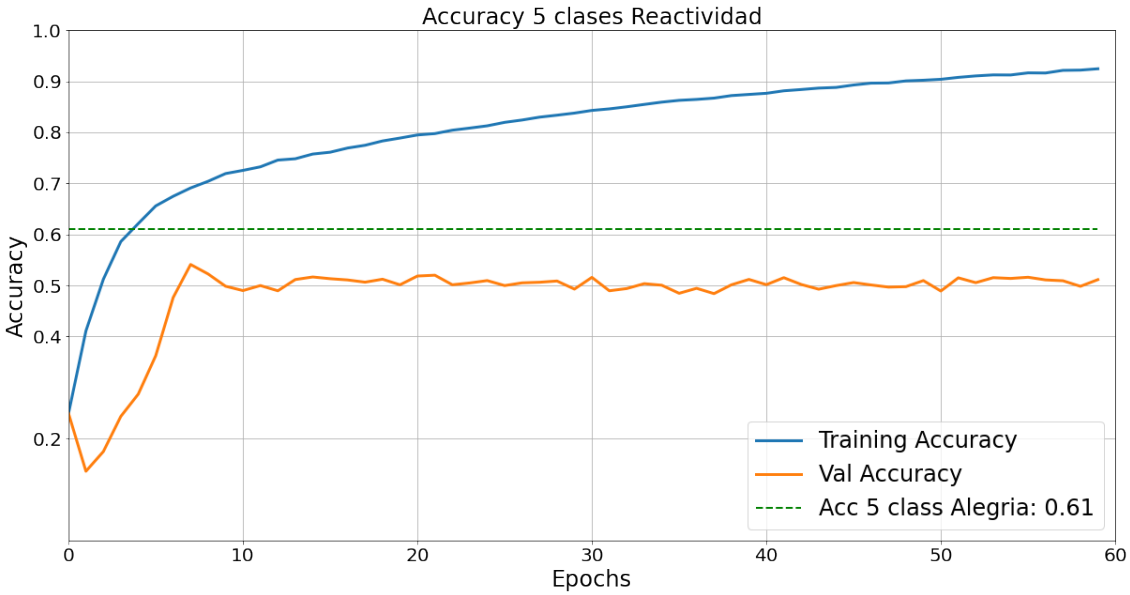


Figura 6.3: Promedio de la evolución del *accuracy* en entrenamiento y evaluación de 34 modelos de clasificador de 5 clases de reactividad.

Cabe mencionar que en esta segunda etapa se entrenan 33 modelos porque existe un paciente sin etiquetas para un tipo de reactividad, lo que genera indefiniciones matemáticas al calcular sus métricas.

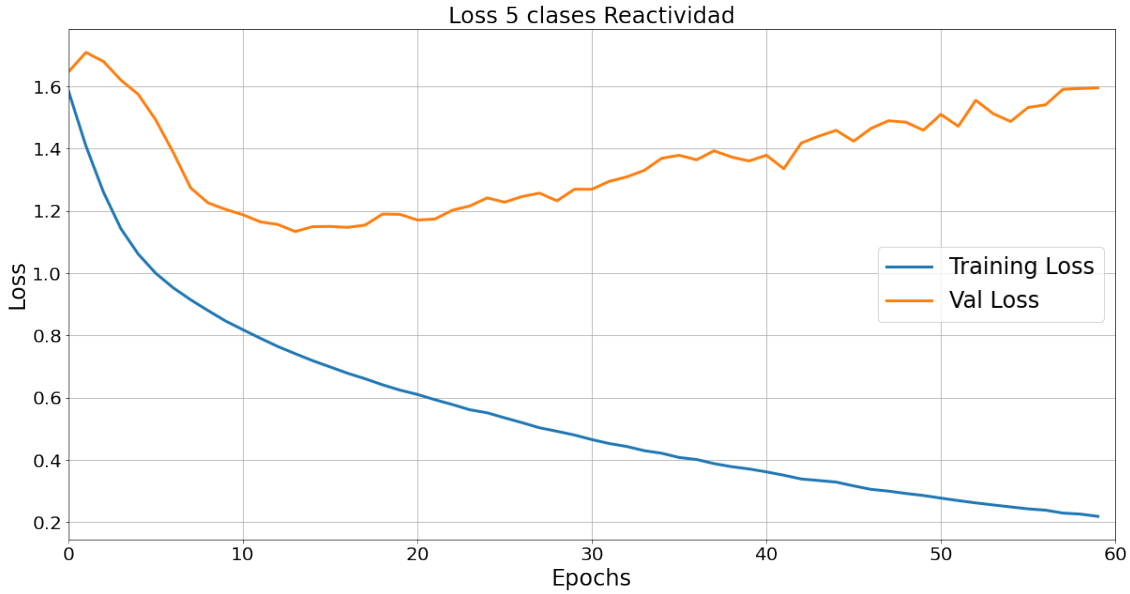


Figura 6.4: Promedio de la evolución del *loss* en entrenamiento y evaluación de 33 modelos de clasificador de 5 clases de reactividad.

La figura 6.4 también es creada calculando el promedio de *loss* para los 33 modelos durante 60 épocas. Se puede observar que el promedio de la función de *loss* cae rápidamente,

disminuyendo hasta alrededor de 12 épocas.

A modo de resumen, se presenta la tabla 6.1, donde se reporta el promedio de las métricas a partir de la época 15. De estas métricas obtenidas para el caso de la clasificación Tumor/-NoTumor, al tener similares valores de *precision* y *recall*, quiere decir que es un clasificador que no está orientado a minimizar los errores de tipo I ni tipo II, específicamente. En el caso de la clasificación de 5 clases, afirmativamente, la *precision* es mayor al *recall*.

Tabla 6.1: Métricas resultados clasificador 5 clases.

	Precision	Recall	F1-score	Accuracy
Tumor/No tumor	0.79	0.78	0.79	0.86
5 Clases Reactividad	0.46	0.34	0.39	0.504

6.2. Discusión

Los resultados de *accuracy* y *loss* para cada época son el promedio de 34 observaciones de los experimentos de clasificación Tumor/NoTumor y del problema de clasificación de 5 clases de reactividad HER2. Entonces, esta metodología promedia 34 observaciones, y se acerca a un uso más real, al nunca usar parches de una muestra a evaluar en el conjunto de entrenamiento. En su conjunto, los experimentos realizados validan la aplicación y replican el trabajo propuesto por Alegría [11].

Es relevante que, en el problema de 5 clases, el rendimiento es inferior al reportado por Alegría [11]. En el problema de 5 clases, Alegría [11] calcula los *accuracy* de una manera diferente a la presentada en este trabajo (forma *mean-per-epoch*). Lo que realiza es concatenar todos los resultados de clase real y clase predicha al predecir con los modelos las clases de los parches de los conjuntos de validación respectivos a cada corrida. Para luego, calcular las métricas como si la concatenación de resultados fuera el resultado de un solo ensamblaje de modelos. Esta metodología se explica en profundidad en Forman [34], donde también se explica que estas formas de calcular las métricas entregan resultados similares.

Un punto a discutir importante con respecto a los modelos en salud es que los falsos negativos son particularmente importantes. Si el modelo predice muchos falsos negativos habrían muchos pacientes que poseen la enfermedad pero que el clasificador predijo que no la poseían, lo cual es un costo muy alto y un error muy grave; en comparación al costo de repetir el examen que es bajo en comparación a la salud de una persona.

Lo que importa es minimizar la tasa de falso negativo. Esto se realiza variando el umbral de decisión sobre las probabilidades entregadas por las neuronas softmax del final de la red. El umbral escogido deberá ser el que garantice un valor mínimo de métrica de sensibilidad.

Entonces, se realiza la variación del parámetro del umbral, resultando en que la sensibilidad del modelo Tumor/NoTumor varía entre 0.68 y 0.71, lo cual no es una diferencia significativa para diversos umbrales. En perspectiva, no es mucho lo que se puede optimizar la red, así que se el umbral escogido es de 0.3 que es el que entrega la sensibilidad máxima de 0.71.

A diferencia de los realizado anteriormente, se utilizó el *cluster Patagón* [13] (Nvidia A100 40 Gb), lo que implicó reducciones de tiempo importantes, pasando de 2 días de cálculo a 8 horas (aprox 16 min por época). Lo que acerca la posibilidad de realizar pruebas extensivas de exploración de parámetros, o de filtrado de imágenes como el propuesto.

En este trabajo se descompuso el problema en un clasificador tumor/NoTumor y, para los parches de clase tumor, se sub-clasificaron con un segundo clasificador de 5 reactividades. Luego, estas 6 clases obtenidas (NoTumor más 5 reactividades) se mapean al problema de 5 clases que resuelven los métodos del estado del arte. Este mapeo de problemas de clasificación se ordena en la tabla 4.1.

Se compara el algoritmo propuesto por Alegría [11], y replicado en esta memoria, con métodos del estado del arte en clasificación automática de imágenes de cáncer de mama como Vandenberghe [26] y Pitkäaho [29]. Se observa que los métodos del estado del arte resuelven el problema de 5 clases (*medio-externo, 0, 1+, 2+ y 3+*), obteniendo un *accuracy* entre $0,78$ en Vandenberghe [26] y $0,98$ en Pitkäaho [29]; ambos usando una *CNN*. Las diferencias se podría deber a: un mayor número de ejemplos (71 *WSI* en [26] y 86 en [29]), mejores etiquetas (etiquetas a nivel de parche en [26] y a que se utilizan diferente tipos de imágenes (cáncer gástrico vs cáncer de mama).

Lamentablemente no se alcanza, por temas de tiempo, a realizar el mapeo que permita comparar el trabajo realizado con el estado del arte en los mismos problemas de clasificación. Sin embargo, revisando de 0.86 *accuracy* Tumor/NoTumor y 0.5 *accuracy* en reactividades, se proyecta que se obtienen **peores resultados** que el estado del arte. Esto se puede explicar porque se entrenó con mucha menor cantidad de imágenes. Sin embargo, la gran ventaja es que en este trabajo se ha logrado realizar la clasificación sin la necesidad de que un patólogo etiqueta parche por parche, lo cual representa un ahorro de tiempo considerable.

Una vez presentadas discusiones sobre los resultados de la línea base, se presentan discusiones sobre el proceso de filtrado de la base de datos. El punto a mencionar es que se toma la decisión de no realizar un filtro en la etapa Tumor/NoTumor. Esta decisión se toma porque los parches no se pueden diferenciar bien de su forma y color porque no funcionaria bien la detección de células.

Para futuros experimentos con esta arquitectura, se recomienda elegir una cantidad de épocas de entrenamiento en una vecindad en torno a las 15 épocas. A partir de la época 15 el

loss de validación comienza a subir en ambos clasificadores y, recordando la sección de *early stopping*, los clasificadores comienzan a sobreajustarse.

Capítulo 7

Conclusión y trabajo futuro

7.1. Conclusión

En este trabajo se replicó un algoritmo existente, alcanzando en una de sus modalidades (2 clases) un rendimiento equivalente, y en otra modalidad (5 clases) un rendimiento inferior. Por otra parte se propuso un método que podría mejorar el rendimiento, mediante el uso exclusivo de parches con mayor calidad de etiquetado, lo que resta a ser evaluado.

Los experimentos realizados muestran que se puede usar un clasificador de *machine learning*, para primero, automatizar la labor de diferenciar parches con tejido tumoral de no tumoral. Segundo, valida que la arquitectura *Inception V3* puede resolver el problema.

Por otra parte, si bien se cambio el *hardware* de cómputo, los resultados de la aproximación se mantuvieron, bajando considerablemente los tiempos de cálculo de varios días a 8 horas, lo que permitirá explorar sistemáticamente otros métodos y espacios de parámetros en trabajos futuros.

Respecto al cumplimiento de los objetivos específicos planteados en este trabajo, estos se lograron de manera parcial. Más detalladamente: (1) Con respecto a la generación de base de datos de mejor calidad de etiquetado. Se logra entender y trabajar con imágenes en formato especial llamado *ndpi*. Se propone un algoritmo para filtrar parches (que se puede utilizar para generar una base de datos de mayor certeza). Posterior al plazo de entrega del borrador de esta memoria, se ejecuta el algoritmo y se logra una base de datos de mayor calidad. También, como parte de la etapa exploratoria de los datos, se programa un script que extrae solamente los parches desde el interior de las ROI, lo que constituye una base de datos de etiquetado de aun mejor calidad, que la utilizada en esta memoria para validar el trabajo de Alegía [11].

Por otro lado, en lo que respecta al objetivo específico (2) entrenamiento y evaluación

de una arquitectura de un algoritmo base presentado en la memoria de Alegría [11] y del algoritmo propuesto. Se logra entrenar y evaluar sólo la arquitectura ya propuesta por Alegría [11]. Obteniendo un accuracy de 86 % en el clasificador biclase Tumor-NoTumor y un accuracy de 50.4 % en el clasificador de cinco clases de reactividades HER2 específicas. No se logra entrenar y evaluar la arquitectura *InceptionV3* en la base de datos filtrada con el algoritmo propuesto.

Por último, en lo que concierne al objetivo de la (3) aplicación del algoritmo de la guía clínica [7], no se logró programar ni aplicar el algoritmo descrito en la guía clínica, el cual consiste principalmente en aplicar la regla del 10 %. Con lo que no se tienen resultados para comparar el sistema propuesto con el desempeño de médicos patólogos.

Finalmente y en suma, dentro de lo logrado, a pesar que no se cumplieron todos los objetivos propuestos, el trabajo realizado constituye una confirmación de que se pueden entrenar modelos de *machine learning* para clasificar imágenes de tejido tumoral de uno noTumoral. Esto es un salto tecnológico en las aplicaciones de *computer vision* en la medicina, en particular en la especialidad de histopatología.

7.2. Trabajo futuro

El principal trabajo futuro es efectivamente evaluar el algoritmo propuesto en el Capítulo 4. En esta evaluación se identifica como el mayor desafío, que los parámetros usados para realizar las operaciones sobre las imágenes sean efectivos en las diferentes muestras. En particular, dado que la intensidad de los colores depende levemente del kit de tinción y de qué tan rigurosa en tiempos fue la metodología de tinción. Por lo que, es un aspecto de posible mejora el explorar métodos que mejoren la robustez del modelo, ya sea, entrenando con otro conjunto de imágenes o desarrollando técnicas de data augmentation que modifiquen levemente la coloración de las muestras, es decir, los valores de la coloración del canal H y del canal DAB.

Otro trabajo a futuro es experimentar con utilizar el dataset que solo posee parches extraídos desde el interior de la circunferencia de la ROI. Es decir, el dataset que excluye los parches que se ubican entre el cuadrado de encierra la circunferencia de la ROI y el círculo de la ROI. Estos parches tienen etiquetado de mejor calidad, con lo que se espera que la arquitectura *InceptionV3* mejore levemente su rendimiento.

Por otro lado, en el presente año 2022, el modelo llamado *Visual Transformer(ViT)* ha sido aplicado en el campo de la histopatología en [8]. El modelo de *Visual transformer* es presentado el año 2021 en el paper titulado *"An image is worth 16x16 words"* de Kolesnikov *et al* [35]. La investigación de *"From CNNs to Vision Transformers – A Comprehensive*

Evaluation of Deep Learning Models for Histopathology" ha mostrado que usando el modelo de *Visual Transformer* se obtienen *accuracy* en torno al 97 % en tres datasets de histopatología. Constituye un trabajo a futuro probar la arquitectura *Visual Transformer* en las imágenes de la base de datos *PRECISO*.

Cabe señalar también que, en el trabajo realizado, se utilizaron sólo resecciones y no biopsias. Es importante notar que las reglas clínicas son diferentes para ambos tipos de muestras y que, desarrollar y evaluar métodos que estén adaptados a ambos tipos de muestras, es un importante trabajo a realizar.

Bibliografía

[1] H. Sung y col, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

[2] H. Sung y col, "Global cancer observatory: Cancer today," *CA: A Cancer Journal for Clinicians*, 2021. https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020&mode=cancer&mode_population=countries&population=900&populations=152&key=asr&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=10&group_cancer=1&include_nmsc=1&include_nmsc_other=1&type_multiple=%257B%2522inc%2522%253Atrue%252C%2522mort%2522%253Atrue%252C%2522prev%2522%253Afalse%257D&orientation=horizontal&type_sort=1&type_nb_items=%257B%2522top%2522%253Atrue%252C%2522bottom%2522%253Afalse%257D# (visitado 21-04-2022).

[3] A. Csendes and M. Figueroa, "Situación del cáncer gástrico en el mundo y en chile," *Revista Chilena de Cirugía*, vol. 69, no. 6, pp. 502–507, 2017. https://www.sciencedirect.com/science/article/pii/S0379389316301533.

[4] C. Ferreccio, J. C. Roa, and C. Bambs, "Study protocol for the maule cohort (mauco) of chronic diseases, chile 2014–2024," *BMC Public Health*, vol. 16, p. 122, Feb 2016. https://doi.org/10.1186/s12889-015-2454-2.

[5] Q. T, M. A, and R. P. C, "Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues," *Histopathology*, vol. 72, pp. 227–238, Jan 2018. doi: 10.1111/his.13333. Epub 2017 Oct 27. PMID: 28771788.

[6] J. A. A. y col, "Nccn clinical practice guidelines in oncology (nccn guideli- nes)," 2017. https://www.nccn.org.

[7] A. N. Bartley, M. K. Washington, and C. B. Ventura, "Her2 testing and clinical decision making in gastroesophageal adenocarcinoma: Guideline from the college of american pathologists, american society for clinical pathology, and american society of clinical oncology," *Archives of Pathology & Laboratory Medicine*, vol. 140, pp. 1345–1363, 12 2016. https://doi.org/10.5858/arpa.2016-0331-CP.

[8] M. Springenberg, A. Frommholz, M. Wenzel, E. Weicken, J. Ma, and N. Strodthoff, "From cnns to vision transformers – a comprehensive evaluation of deep learning models for histopathology," 2022. https://arxiv.org/abs/2204.05044.

[9] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016. https://ieeexplore.ieee.org/document/7312934.

[10] A. M. Andrew Janowczyk, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, no. 1, p. 29, 2016. https://www.sciencedirect.com/science/article/pii/S2153353922005478.

[11] J. J. Alegría, "Clasificación automatizada de sobreexpresion de proteína her2 en biopsias digitalizadas de cáncer gástrico teñidas hinmunohistoquímicamente," Master's thesis, Departamento de ciencias de la computación, Universidad de Chile, 2019. https://repositorio.uchile.cl/handle/2250/176896.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[13] "Patagón supercomputer austral university of chile." https://patagon.uach.cl, 2021. fecha ultima visita 15 de junio 2022.

[14] N. C. Institute, "Diccionario de cáncer del nci," *National Cancer Institute*, 2022. https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/buscar/cancer/?searchMode=Begins (visitado 5-05-2022).

[15] K. D. Crew and A. I. Neugut, "Epidemiology of gastric cancer," *World J Gastroenterol*, vol. 12, pp. 354–362, jan 2006. https://pubmed.ncbi.nlm.nih.gov/16489633/.

[16] A. A. C. Society, "Cancer facts & figures 2022," 2022. https://www.cancer.net/es/tipos-de-cancer/cancer-de-estomago/estad%C3%ADsticas.

[17] A. A. C. Society, "Cancer facts & figures 2022," 2022. https://www.cancer.net/es/tipos-de-cáncer/cáncer-de-estómago/sobrevivencia.

[18] P. Brenner, S. Dathe, M. T. González, F. Hofmann, P. Jara, V. Montes, and E. Montes, "Descripción epidemiológica del cáncer gástrico en chile," *Revista Confluencia*, vol. 3, pp. 57–61, dic. 2020. https://revistas.udd.cl/index.php/confluencia/article/view/462.

[19] H. Sung y col, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," 2022. https://gco.iarc.fr/today/online-analysis-table?v=2020&mode=cancer&mode_population=continents&population=900&populations=152&key=asr&sex=2&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_

group%5B%5D=17&group_cancer=1&include_nmsc=0&include_nmsc_other=1.

[20] T. L. Ang and K. M. Fock, "Clinical epidemiology of gastric cancer," *Singapore medical journal*, vol. 55, pp. 621–628, Dec 2014. https://pubmed.ncbi.nlm.nih.gov/25630323.

[21] M. Arnold, S. P. Moore, S. Hassler, L. Ellison-Loschmann, D. Forman, and F. Bray, "The burden of stomach cancer in indigenous populations: a systematic review and global assessment," *Gut*, vol. 63, no. 1, pp. 64–71, 2014. https://gut.bmj.com/content/63/1/64.

[22] S. de Salud Pública, "Guías clínicas auge: Cáncer gástrico," 2014. https://www.minsal.cl/sites/default/files/files/GPC%20Gástrico%20(PL).pdf.

[23] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, and L. Norton, "Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2," *New England Journal of Medicine*, vol. 344, no. 11, pp. 783–792, 2001. https://doi.org/10.1056/NEJM200103153441101.

[24] Y.-J. Bang, E. Cutsem, A. Feyereislova, H. C. Chung, L. Shen, A. Sawaki, F. Lordick, A. Ohtsu, Y. Omuro, T. Satoh, G. Aprile, E. Kulikov, J. Hill, M. Lehle, J. Rüschoff, and Y.-K. Kang, "Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of her2-positive advanced gastric or gastro-oesophageal junction cancer (toga): a phase 3, open-label, randomised controlled trial," *The Lancet*, vol. 376, no. 9742, pp. 687–697, 2010. https://doi.org/10.1016/S0140-6736(10)61121-X.

[25] A. Burkov, *Machine Learning Engineering*. True Positive Incorporated, 2020.

[26] M. Vandenberghe, M. Scott, P. Scorer, M. Söderberg, D. Balcerzak, and C. Barker, "Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer," *Scientific Reports*, vol. 7, April 2017. http://doi.org/10.1038/srep45938.

[27] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351 of *LNCS*, pp. 234–241, Springer, 2015. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.

[28] M. Saha and C. Chakraborty, "Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2189–2200, 2018. https://doi.org/10.1109/TIP.2018.2795742.

[29] T. Pitkäaho, T. M. Lehtimäki, J. McDonald, T. J. Naughton, *et al.*, "Classifying her2 breast cancer cell samples using deep learning," in *Proc. Irish Mach. Vis. Image Process. Conf*, pp. 1–104, 2016.

[30] B. Muller, "Observational study of perioperative chemotherapy in gastric cancer," *ClinicalTrials.gov*, 2012-2020. NCT01633203. https://clinicaltrials.gov/ct2/show/NCT0163320.

[31] C. Deroulers, D. Ameisen, M. Badoual, C. Gerin, A. Granier, and M. Lartaud, "Analyzing huge pathology images with open source software," *Diagnostic Pathology*, vol. 8, p. 92, Jun 2013.

[32] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, 2001.

[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[34] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *SIGKDD Explor. Newsl.*, vol. 12, p. 49–57, nov 2010.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

Anexo A

Formato *ndpa* para anotaciones

Los archivos *ndpa* de anotaciones son archivos *XML*. Para acceder a la información de las etiquetas se cambia el formato de los archivos *ndpa* a formato *XML* con un script escrito en *Python*. Los archivos *XML* resultantes se leen utilizando la librería de *Python* *xml.etree.ElementTree*. Lo que genera un objeto con estructura de dato de tipo árbol. En este objeto, cada anotación es un nodo de nombre *ndpviewstate*. Usando estas herramientas, se implementa un método para acceder a cada etiqueta.

Cada nodo llamado *ndpviewstate* tiene un nodo hijo llamado *annotation*. Cada nodo *annotation* guarda las coordenadas del centro y el radio de la ROI. Esta información está guardada en el sistema de coordenadas de *ndpi*, en que una unidad significa 1 *nm*. Por ejemplo, un radio de ROI de 5.487 significa que el radio de la ROI dibujada por el patólogo es de 5.487 *nm*. En el Anexo B se explicará este sistema de coordenadas.

Además, los nodos *annotation* tienen tres parámetros. Uno de ellos se denomina *color* y su valor es un *string* que contiene el color de la etiqueta en notación hexadecimal. Por ejemplo color="ff00ff.[es] el color magenta, que es el color con que el patólogo realiza las anotaciones de las regiones de interés de reactividad lineal fuerte. En la tabla A.1 se observa la asignación de colores a las reactividades.

Tabla A.1: Pauta de asociación entre etiquetas de reactividad y color usado por patólogo

Etiqueta reactividad	Color	Hexadecimal
Reactividad lineal fuerte	Magenta	#ff00ff
Reactividad lineal debil	azul	#0000ff
Reactividad lineal casi imperceptible	rojo	#ff0000
Reactividad no lineal	amarillo	#ffff00
Sin reactividad	verde	#00ff00
No tumor	negro	#000000

Es posible agregar anotaciones manualmente editando el archivo ndpa. Agregar una anotación equivale a agregar un nodo *ndpviewstate*.

Anexo B

Conversión coordenada *ndpi* a *pixeles*

Para deducir la conversión de coordenada con un editor de texto se agregó manualmente en los los archivos ndpa anotaciones. En particular, para descubrir el origen del sistema de coordenadas ndpi se agrego un nodo con centro en $(0, 0)$ y radio 1.000.

Para obtener la conversión de coordenadas, se agregaron marcas (pines) en el archivo ndpa, en las esquinas en el sistema de coordenadas *ndpi*. Para visualizar los pines en el software del fabricante se ajustaron los parámetros de visualización (Gamma, brillo). En las siguientes imágenes se puede ver la imagen resultante.

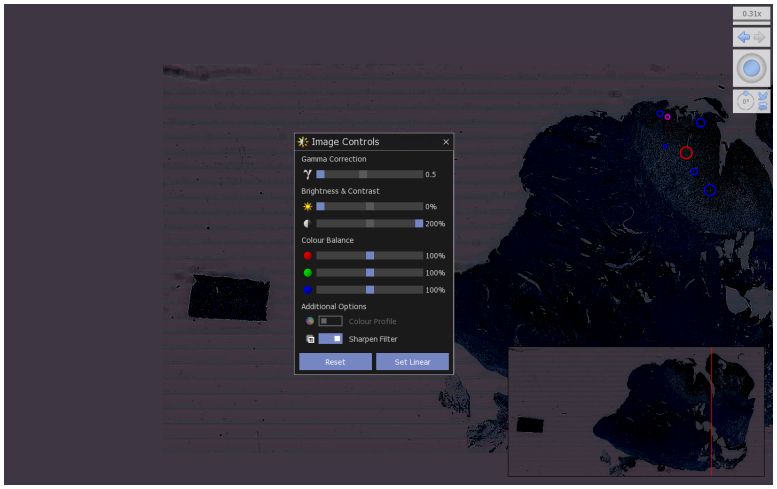


Figura B.1: Aplicación herramienta *image controls* de NDP.view2

Como se muestra en la figura B.1 se agregaron 4 pines en las cuatro esquinas de la imagen y se analizaron las coordenadas ndpi resultantes desde los archivos *ndpa*. En retrospectiva, estábamos buscando una regla de conversión de un sistema de coordenadas a otro. Entonces, teniendo 4 puntos de referencia resultantes de esta transformación se estima la transformación, como se muestra en la ecuación 4.3.

$$x_{pixeles} = \lfloor \left(\frac{x_{ndpi} - x_{offset}}{1000 * mpp_x} + \frac{widthl_0}{2} \right) * \frac{10}{40} \rfloor, \tag{B.1}$$

$$x_{pixeles} = \lfloor \left(\frac{y_{ndpi} - y_{offset}}{1000 * mpp_y} + \frac{lengthl_0}{2} \right) * \frac{10}{40} \rfloor, \tag{B.2}$$

$$radio_{pixeles} = \lfloor \frac{radio_{ndpi}}{227} \rfloor. \tag{B.3}$$

Las variables significan:

(1) x_{ndpi} e y_{ndpi}: son las coordenadas x e y en que el patólogo realiza la anotación, están en el sistema de coordenadas *ndpi*. Una unidad de estas variables significa un nanómetro.

(2) $radio_{ndpi}$: radio que le da el patólogo al círculo que dibuja. Esta en nanómetros.

(3) mpp_x y mpp_y: micrómetros por píxel, representas cuantos micrómetros de la realidad mide un píxel, en el eje x y en el eje y. Son casi el mismo número (diferenciados en la millonésima). En este trabajo se usa mpp_x con el valor de mpp_y por simplicidad.

Para explicar las variables x_{offset}, y_{offset}, *heightl*0 y *widthl*0, es necesario explicar que los archivos *ndpi* guardan dos tipos de imágenes en la más mínima resolución, es decir, en las cuales se ve toda la biopsia a baja resolución. La primera se llama imagen principal y es una imagen que contiene un código QR y la parte de la placa de petri en que los investigadores escriben identificadores de la muestra. La segunda se llama imagen fuente y es una imagen que solamente contiene la muestra. Entonces:

(4) x_{offset} e y_{offset}: son las distancias desde el centro de la imagen principal, que incluye el código QR de la muestra, al centro de la imagen fuente. Los números se guardan como float y están en nanómetros.

(5) *widthl*0 y *heightl*0: son el ancho (horizontal) y el alto (vertical) de la imagen fuente, es decir, la imagen que solo contiene la muestra. Están medidos en píxeles, en la resolución más alta (40x).

En el sistema *ndpi* al dar un número se refiere a cuantas *nanómetros* de la realidad representa. Por ejemplo, si la imagen tiene una anotación en (100, 200) significa que la anotación está hecha en el punto (100 nanómetros, 200 nanómetros).

Las variables: x_{offset}, y_{offset}, mpp_x, mpp_y, *widthl*0 y *heightl*0, se obtienen usando el paquete *OpenSlide*. Son *metadatos* que el microscopio guarda en los archivos *ndpi*. También, se pueden obtener manualmente abriendo la imagen *ndpi* con *ndpi.viewer2*, en la sección:

Información de la imagen.

Anexo C

Uso de *ndpisplit*

NDPIsplit es un paquete ejecutable producto de la investigación de Deroulers *et al* [31] del año 2013. *NDPIsplit* se obtiene descargando el archivo ejecutable para instalarlo. Se utiliza porque los archivos *ndpi* obtenidos de microscopios *Hamamatsu* no siempre pueden ser procesados con los software de procesamiento científico de imágenes como *ImageJ* o *Python*, por su gran tamaño. Entonces, con *NDPIsplit* se logran procesar los archivos *ndpi*.

NDPIsplit se utiliza desde un *command line interface* (CLI). En *Linux* una terminal es una interfaz de línea de comandos. Para procesar un archivo *ndpi* se abre una *Terminal* en la carpeta donde se encuentra el archivo. Luego, *NDPIsplit* se ejecuta ingresando un comando de la forma: *ndpisplit optionsflags nombrearchivo.ndpi*. Luego, la *Terminal* muestra las magnificaciones y dimensiones encontradas para la imagen dentro del archivo *ndpi*, además, muestra el nombre del archivo *TIFF* que es el parche cortado. A modo de referencia, recortar una imagen de 100 x 100 píxeles de un archivo *ndpi* de 2,19 GB toma 0.049 segundos medido con el comando time de *Linux*.

Para cortar todos los parches necesarios se ejecutan cientos de comandos *ndpisplit*. Esto se automatiza escribiendo un *script* en *Python* que genera un archivo *txt* en que cada línea es un comando. Luego, en *MAC OS*, se ejecutan todos los comandos del *txt* usando: *zsh nombre-archivo.txt*. Lo que se está haciendo es usar el intérprete de comandos llamado *zsh* para ejecutar los comandos. En el *cluster Patagón* se realiza de la misma forma, dado que el programador también se comunica con el cluster mediante una *Terminal* de *Linux*.