UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CLASIFICACIÓN AUTOMATIZADA DE SOBREEXPRESIÓN DE PROTEÍNA HER2 EN BIOPSIAS DIGITALIZADAS DE CÁNCER GÁSTRICO TEÑIDAS INMUNOHISTOQUÍMICAMENTE

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

JUAN JOSÉ ALEGRÍA FUENTES

PROFESOR GUÍA:
MAURICIO CERDA VILLABLANCA

MIEMBROS DE LA COMISIÓN:
JUAN BARRIOS NUÑEZ
FELIPE BRAVO MARQUEZ
VICENTE ACUÑA AGUAYO

Este trabajo ha sido parcialmente financiado por el Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT 11611033) y la ICM (P09-015-F).

SANTIAGO DE CHILE
2020

Resumen

Chile es el séptimo país del mundo donde el cáncer gástrico es más frecuente, siendo además el segundo tipo de cáncer más mortal en el país. En 2010, se demostró que en casos de cáncer gástrico avanzado con sobreexpresión de la proteína HER2, un tratamiento con un anticuerpo llamado Trastuzumab puede prolongar la sobrevida del paciente de manera estadísticamente significativa. Para determinar la sobreexpresión de HER2, se realiza un examen inmunohistoquímico sobre una biopsia gástrica, la cual es luego analizada por un patólogo. Este proceso, si bien está estandarizado mediante guías clínicas, es inherentemente subjetivo, lo cual genera que exista variabilidad en el diagnóstico entregado por distintos especialistas. Por otro lado, durante los últimos años los campos de aprendizaje de máquinas y aprendizaje profundo han experimentado una explosión, en especial en lo relativo a clasificación de imágenes. Por ello, con el fin de proveer una herramienta de apoyo en el proceso de evaluación llevado a cabo por los patólogos, en el presente trabajo se estudió cómo clasificar una biopsia digitalizada de cáncer gástrico con tinción inmunohistoquímica de acuerdo a su sobreexpresión HER2, utilizando técnicas de procesamiento de imágenes y modelos de *Deep Learning*. Además, se buscó que los modelos generados siguieran en lo posible los procesos recomendados en las guías clínicas correspondientes.

Tras probar varias configuraciones experimentales, se construyó un algoritmo de clasificación que logra buenos resultados, con una recuperación de 100 % de las clases *Equívoco* y *Positivo*, las más relevantes en términos clínicos. Las clasificaciones de biopsias realizadas por el mejor modelo construido obtienen una concordancia de 89 % con respecto a lo evaluado por el patólogo que generó el conjunto de datos utilizado, con un κ de Cohen de 0.61. Este grado de concordancia es catalogado como considerable, y está en el mismo rango que la concordancia obtenida entre patólogos al evaluar el examen de sobreexpresión HER2. De todas maneras, esta concordancia es menor a la obtenida en trabajos similares que utilizan *deep learning* para clasificar la sobreexpresión de esta proteína en cáncer de mama. Además, a nuestro conocimiento no existen trabajos que utilicen aprendizaje de máquinas y un enfoque de replicación del proceso diagnóstico para clasificar la sobreexpresión de HER2 en biopsias de cáncer gástrico.

Por otro lado, se dieron los primeros pasos para la construcción de un sistema de visualización de sobreexpresión HER2 sobre una biopsia, lo cual permitiría que un patólogo pudiera fácilmente identificar cómo el algoritmo clasificó cada zona de la muestra, paliando en parte el hecho de que las redes neuronales convolucionales suelen ser consideradas como un sistema de caja negra, y proveyendo así un grado de interpretabilidad imprescindible en la práctica médica.

A mi viejo: tu ausencia fue la que me impulsó a investigar sobre cáncer.
A mi mami: tu presencia durante todos estos años hizo posible que pudiera llegar hasta acá.

Agradecimientos

A mi mami, por su amor, dedicación y apoyo incondicional, por sacarme adelante y por estar ahí siempre. De todo corazón, sé que no me alcanzará la vida para poder devolverle todo lo que ha hecho por mí, y aún así sé que eso no le importa. Ojalá algún día ser una persona tan bacán como usted. Y a mi viejo, porque pese a tu ausencia, de una u otra forma siempre estás. Porque por ti decidí investigar sobre cáncer gástrico, porque me gusta creer que estarías orgulloso de mí y porque cada vez que pensé en tirar la toalla, pensar en ti me ayudó a continuar.

A mi tía Judy y mi tío Pato, por ser una segunda pareja de padres para mí. No siempre lo digo, pero ustedes son pilares fundamentales en mi vida. A la Pame y a la Natty, por ser casi mis hermanas, y bueno, por soportarme. Sé que, bien en el fondo, igual me tienen cariño. A toda mi familia de Peñalolén, en especial a mi tío Juan y mi tía Rosa, porque pese a la distancia, sé que siempre puedo contar con ustedes. A tod@s mis amig@s, de la U, del liceo, a los viejos y a los nuevos; gracias por todos los momentos de diversión y/o de conversaciones serias. En especial al Berro, por ser casi un hermano para mí. A mis compañer@s de la FONGSP y Trabajos Voluntarios FECh, por hacerme ver que es posible trabajar por nuestra gente y darle sentido a todos estos años en la universidad. A todas y todos los animalitos que han pasado por mi vida (Minino, Blu, Blacky, Reina, Niña, Blanca, Martín, Minina, Lulú, Bengie, Dandy, Domitila, Toby, Arumi, Yuki, Yoru, Rosado, Negro Matapacos, etc.) que si bien, por motivos obvios, nunca leerán esta tesis, siempre me han ayudado a recordar que hay cosas buenas en el mundo. Y eso es algo que debe ser agradecido.

A Mauricio Cerda, por ser un excelente profesor guía y brindar su apoyo en los momentos difíciles, que fueron muchos. A la Dra. Bettina Müller, por facilitarnos las biopsias usadas en este trabajo. Al doctor Pablo Zoroquiain, por su generosa e invaluable ayuda en el momento más complicado de este proceso. Este proyecto no se podría haber llevado a cabo sin cada uno de ustedes, muchísimas gracias.

A todos los pacientes del estudio PRECISO que dieron su consentimiento para la determinación de HER2 en sus muestras; esto también es por ustedes y gracias a ustedes.

A toda la gente que está luchando por, entre otras cosas, el derecho a un sistema de salud digno. Espero que este proyecto sea un aporte (minúsculo, lo sé) en esa línea.

Finalmente, pero por ningún motivo menos importante, a Tamara, por acompañarme siempre, por darme infinito amor y nunca perder la fe en mí. Por pasar toda esta etapa tan difícil a mi lado, y por todas las próximas etapas que nos quedan por vivir.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

El cáncer gástrico (CG) es el quinto cáncer más común en el mundo, y el tercero con mayor mortalidad. Si bien su incidencia en el mundo occidental ha disminuido, sigue siendo un cáncer muy prevalente en Asia Oriental, Europa Oriental y Sudamérica. En particular, Chile es el séptimo país del mundo con mayor incidencia de esta enfermedad, y dentro del país, es el segundo cáncer que más muertes produce. Esta alta mortalidad se debe principalmente a que el CG no presenta sintomatología temprana y es detectado en estadios muy tardíos, cuando el cáncer ya está en una etapa avanzada [13] [80].

En 2010, un estudio clínico demostró que en pacientes que presentan sobreexpresión de la proteína HER2, un tratamiento con quimioterapia y el anticuerpo monoclonal Trastuzumab puede aumentar de manera estadísticamente significativa la sobrevida de los pacientes [9]. Si bien este tratamiento se llevaba practicando desde hace años en cáncer de mama, existen varias diferencias en la manera en que la proteína HER2 se expresa en ambos cánceres; por ello, fue necesario generar una guía clínica específica detallando cómo evaluar la sobreexpresión de HER2 en CG [10].

En dicha guía clínica, se mencionan dos exámenes posibles para tal fin: el primero es un análisis inmunohistoquímico (IHC, por sus siglas en inglés), donde una biopsia es sometida a la acción de ciertos anticuerpos que generan una reacción específica y con ello una tinción particular en aquellas zonas donde la proteína HER2 está sobreexpresada. Luego, con la biopsia ya teñida y dependiendo de los patrones que se formaron en ésta, un patólogo debe analizar la biopsia y clasificarla en 0 (negativo), 1+ (negativo), 2+ (equívoco) o 3+ (positivo); sólo si es que el resultado es positivo se procede a administrar el tratamiento con Trastuzumab al paciente. Por otro lado, si es que el resultado es equívoco se debe llevar a cabo el segundo examen: hibridación fluorescente in situ (FISH, por sus siglas en inglés), en el cual ya no se estudia la sobreexpresión de la proteína HER2, sino que se analiza la amplificación del gen *HER2* responsable de la codificación de la proteína del mismo nombre.

La expresión de HER2 en cáncer gástrico suele ser más heterogénea que en cáncer de mama [11], lo cual sumado a la subjetividad inherente del análisis IHC, donde el patólogo debe analizar la imagen e interpretarla según su propia experiencia, puede conducir a que distintos patólogos entreguen distintos resultados para una misma muestra.

Por otro lado, durante los últimos años, el área de aprendizaje de máquinas (en inglés *machine learning*, ML) ha cobrado gran relevancia. Esto se debe al desarrollo de nuevos algoritmos, la mejora constante en aspectos de *hardware* y la masiva cantidad de datos disponibles. Así, diversas técnicas de ML son aplicadas en tareas de reconocimiento de imágenes, procesamiento de lenguaje natural, análisis de series de tiempo, etc., y en áreas tan diversas como astronomía, economía y psicología. Del mismo modo, ML tiene también muchas aplicaciones en el campo de medicina [47] [23], y en particular en cáncer, siendo utilizado por ejemplo para predecir tiempo de sobrevida de pacientes y recurrencia de la enfermedad [48].

Algunos de los algoritmos que más atención han concitado dentro de ML son aquellos que se suelen englobar en el concepto de aprendizaje profundo (o *Deep Learning*, DL). Este conjunto de algoritmos tiene la capacidad de aprender múltiples niveles jerárquicos de información, y de extraer características relevantes a partir de datos en bruto; así, ofrecen la ventaja de que los investigadores no necesitan realizar ingeniería de características (*feature engineering*) sobre los datos y los algoritmos pueden ser alimentados con datos en bruto o poco procesamiento previo, y además han demostrado buenos resultados en diversos campos [21]. Esto es especialmente relevante si es que se trabaja con imágenes, dado que extraer características relevantes de cada imagen puede ser un trabajo complejo. De esta forma, los modelos de DL, en especial las Redes Neuronales Convolucionales (*Convolutional Neural Networks*, CNN) suelen ser ampliamente utilizados para tareas de clasificación de imágenes.

1.1. Motivación

Dado todo lo anterior, se tiene por un lado que existen algoritmos de DL que han alcanzado excelentes resultados en tareas de procesamiento de imágenes, y por otro lado, un proceso de diagnóstico médico complejo y subjetivo como es el de clasificar la sobreexpresión de HER2 en biopsias de cáncer gástrico. Dado que estas biopsias pueden ser digitalizadas y tratadas como imágenes, la pregunta que surge naturalmente es si los algoritmos de DL pueden ayudar a los patólogos a llevar a cabo su tarea de forma más estandarizada y eficiente, reduciendo en parte la subjetividad del análisis.

Algo a tener en cuenta es que debido a la importancia que tiene el diagnóstico médico, cualquier resultado entregado por un algoritmo debe ser fácilmente interpretable. Dado que las CNN en particular son reconocidas como algoritmos de caja negra, donde se entrega una entrada y el algoritmo produce una salida sin explicar el porqué de ésta, es que se debe tener especial cuidado en cómo son empleados estos modelos, buscando mantener los buenos resultados que las redes neuronales convolucionales suelen producir, pero a la vez siendo capaz de entregar un resultado que el cuerpo médico tratante pueda interpretar y usar para generar el diagnóstico final.

1.2. Hipótesis

Existen trabajos en la literatura sobre cómo aplicar redes neuronales para clasificar la sobreexpresión de HER2 en cáncer de mama siguiendo el protocolo señalado en las guías clínicas [89]. Por otra parte, la investigación en cáncer gástrico para este mismo problema no necesariamente sigue las guías clínicas estandarizadas y no proveen un resultado fácilmente interpretable por el equipo médico [81] [65]. De esta forma, la hipótesis del presente trabajo es que es posible generar un sistema automatizado de procesamiento de imágenes para la clasificación de sobreexpresión de la proteína HER2 en muestras de cáncer gástrico con tinción inmunohistoquímica, con resultados comparables a los entregados por patólogos. En particular, se plantea utilizar técnicas de procesamiento de imágenes digitales en conjunto con algoritmos de ML, para clasificar correctamente imágenes de biopsias de cáncer gástrico en las correspondientes clases (negativo, equívoco, positivo), intentando replicar el proceso de diagnóstico efectuado por médicos especialistas, y proveyendo un resultado interpretable por el equipo médico.

1.3. Metodología

Para llevar a cabo este trabajo se utilizarán biopsias anonimizadas de cáncer gástrico provenientes del estudio *PRECISO*, efectuado por el Grupo Oncológico Cooperativo Chileno de Investigación [61]. Las biopsias de este estudio cuentan con tinción de hematoxilina-eosina (H&E, utilizada para detectar la presencia de cáncer) y con tinción inmunohistoquímica, junto a la clasificación global de HER2 de cada biopsia. Además, se le solicitó a un equipo de patólogos que realizaran marcajes sobre las biopsias digitalizadas, identificando así regiones de interés y anotando la clasificación de cada zona.

Las biopsias digitalizadas son imágenes de gran tamaño (entre 800 MB a 2 GB en disco, pudiendo ocupar hasta 30 GB en memoria RAM) y poseen un esquema piramidal en distintas magnificaciones; en particular, las biopsias del estudio *PRECISO* están almacenadas en el formato propietario `ndpi`. Por ello, se utilizará el software `ndpisplit` [24] para particionar cada biopsia y generar parches de un tamaño manejable.

Utilizando los parches extraídos de las biopsias, y teniendo la clasificación realizada por el equipo de patólogos, se construirá una base de datos para entrenar algoritmos de Deep Learning. Para ello, se usará el software de programación `Python` y los frameworks `TensorFlow` [1] y `Keras` [17]. Finalmente, se construirá un prototipo de visualizador, también en `Python`.

1.4. Objetivos

Los objetivos del presente trabajo de tesis son detallados a continuación.

1.4.1. Objetivo general

El objetivo de esta tesis es la creación de un sistema computacional que permita clasificar automatizadamente la sobreexpresión de proteína HER2 en biopsias digitalizadas de cáncer gástrico, de tal manera de proveer a los patólogos de una herramienta de apoyo para el proceso de diagnóstico médico.

1.4.2. Objetivos específicos

Los objetivos específicos del presente trabajo son:

- Utilizando las biopsias provenientes del estudio *PRECISO* y las anotaciones realizadas por patólogos, construir un *dataset* apto para la aplicación de algoritmos de aprendizaje de máquinas.
- Construir un modelo de DL que sea capaz de clasificar imágenes, extraídas de una biopsia de cáncer gástrico con tinción IHC, de acuerdo a su clase correspondiente. Dicho modelo debe cumplir con:
  - Replicar el proceso diagnóstico llevado a cabo por los médicos especialistas.
  - Ser interpretable por dichos médicos.
  - Obtener resultados comparables a los entregados por los patólogos.
- Evaluar qué magnificación de imagen obtiene mejores resultados.
- Con el algoritmo generado, construir un prototipo de sistema de visualización para patólogos.

Capítulo 2

Antecedentes

En este capítulo se explicará qué es el cáncer gástrico, su epidemiología y sus distintas taxonomías, además de la relevancia de la proteína HER2 y cómo los patólogos reconocen la sobreexpresión de dicha proteína. Posterior a eso, se presentarán conceptos básicos de imágenes digitales y cómo éstas son usadas en el campo de patología digital. Luego, se hará una introducción al aprendizaje de máquinas, pasando por redes neuronales, aprendizaje profundo y transferencia de aprendizaje. Posterior a eso, se describirán otros modelos de aprendizaje de máquinas que son ampliamente mencionados en la literatura, además de las métricas de evaluación utilizadas en el presente trabajo. Finalmente, al término del capítulo se presentará una revisión del estado del arte respecto al análisis de sobreexpresión de HER2 utilizando técnicas de *machine learning*.

2.1. Cáncer gástrico

De acuerdo al Instituto Nacional del Cáncer de Estados Unidos (NCI, por sus siglas en inglés), una *neoplasia* o *tumor* es una "masa anormal de tejido que aparece cuando las células se multiplican más de lo debido o no se destruyen en el momento apropiado". Una neoplasia puede ser *benigna* o *maligna*, siendo benigna cuando las células afectadas no se diseminan a otras partes del cuerpo, y maligna cuando dichas células tienen la capacidad de invadir tejidos cercanos o de propagarse por el torrente sanguíneo o linfático. Cuando una neoplasia es maligna, es llamada *cáncer* [62].

En particular, el cáncer gástrico es aquel que se origina en el estómago, órgano que forma parte del aparato digestivo, y que ayuda a digerir los alimentos al mezclarlos con jugos digestivos [62]. El estómago está formado por cinco componentes: cardias, fondo, cuerpo, antro y píloro, tal como se ve en la figura 2.1.1.

Además, la pared del estómago tiene varias capas, las cuales cobran gran relevancia a la hora de determinar el estadio del cáncer. Desde la más interna hasta la más externa, las capas del estómago son (ver figura 2.1.2):

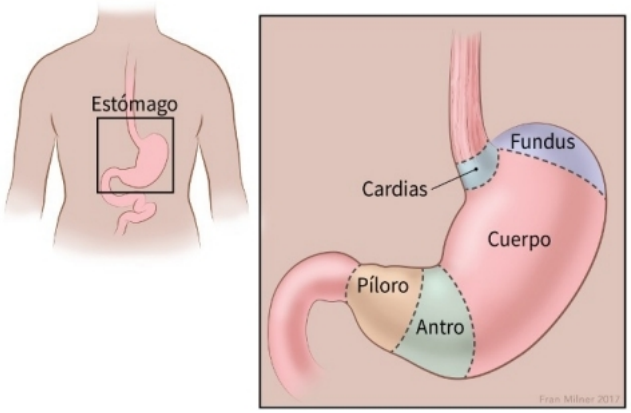


Figura 2.1.1: Partes del estómago. Fuente: American Cancer Society [6].

- *Mucosa*; aquí se producen las enzimas digestivas y el ácido estomacal.
- *Submucosa*.
- *Muscular propia*; capa de músculo que mueve y mezcla el contenido del estomago.
- *Subserosa*.
- *Serosa*; junto a la subserosa, actúan como recubrimiento del estómago.

2.1.1. Epidemiología del cáncer gástrico

El cáncer gástrico es el quinto cáncer más común en el mundo, después del cáncer de pulmón, mamas, próstata y colon [13], con aproximadamente 1.033.000 casos nuevos durante el 2018 en todo el mundo, que representan el 5,7 % de todos los casos de cáncer aparecidos durante ese año. Además, representa el tercer cáncer con mayor número de muertes asociadas, sólo después del cáncer de pulmón y el colorrectal. Uno de los principales problemas del cáncer gástrico es que no suele presentar sintomatología temprana, lo cual se suele traducir en una

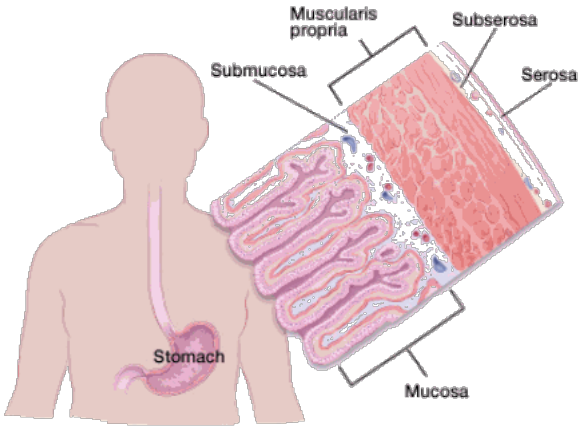


Figura 2.1.2: Capas de la pared estomacal. Fuente: American Cancer Society [6].

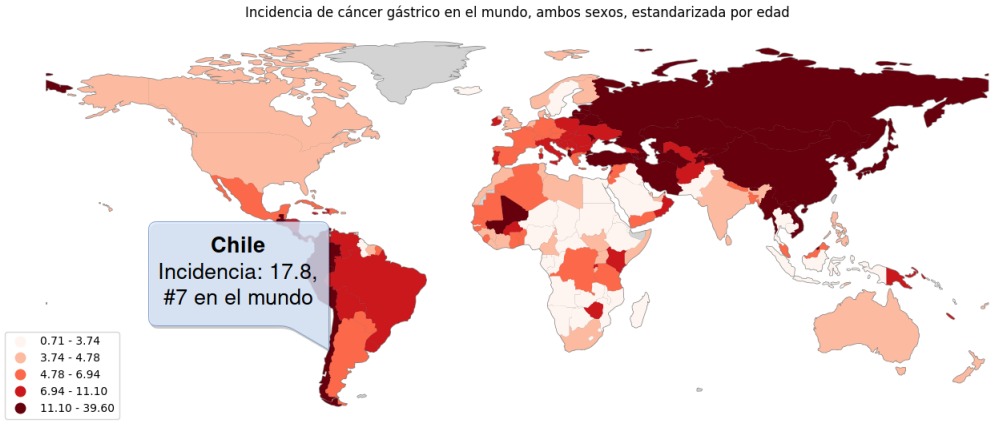


Figura 2.1.3: Incidencia del cáncer gástrico en el mundo, para ambos sexos, estandarizada por edad; resaltada está la incidencia de Chile. Fuente: Elaboración propia con datos extraídos de Ferlay y col. [28].

demora en el diagnóstico; así, se estima que cerca de un 80 % de los pacientes es diagnosticado cuando el cáncer ya está en un estado avanzado [80].

Usualmente, la cantidad de nuevos casos ocurridos en un periodo y lugar específico es conocida como *incidencia*, y se mide ya sea en números absolutos o en un ratio de 100.000 personas por año. Del mismo modo, la *mortalidad* es la cantidad de muertes ocurridas en un tiempo y lugar específicos, y la tasa de mortalidad es el número de muertes por cada 100.000 personas en un año. Dado que los indicadores de incidencia y mortalidad dependen fuertemente de la edad y esto dificulta la comparación entre poblaciones con diferentes composiciones etáreas, se incorporan también *tasas estandarizadas por edad* (ASR, por sus siglas en inglés), donde se asume una población con una distribución estándar y se calculan indicadores ponderados por edad [3] [41].

Si los datos previos se desglosan por sexo, se tiene que el cáncer de estómago es el cuarto con mayor incidencia a nivel mundial para hombres (7,2 % del total) y el séptimo para mujeres (4,1 %). Por otro lado, la mortalidad en hombres es la tercera más alta (9,5 %), mientras que en mujeres ocupa el quinto lugar (6,5 %).

Sin embargo, la incidencia no es uniforme a través del mundo; como se ve en la figura 2.1.4, el cáncer gástrico es mucho más común en Asia Oriental, Europa Oriental y Sudamérica. Factores que explican esta variabilidad son las diferencias en la prevalencia de *Helicobacter Pylori* (principal factor de riesgo para el CG) y distintos tipos de dieta, además del consumo de alcohol y tabaco.

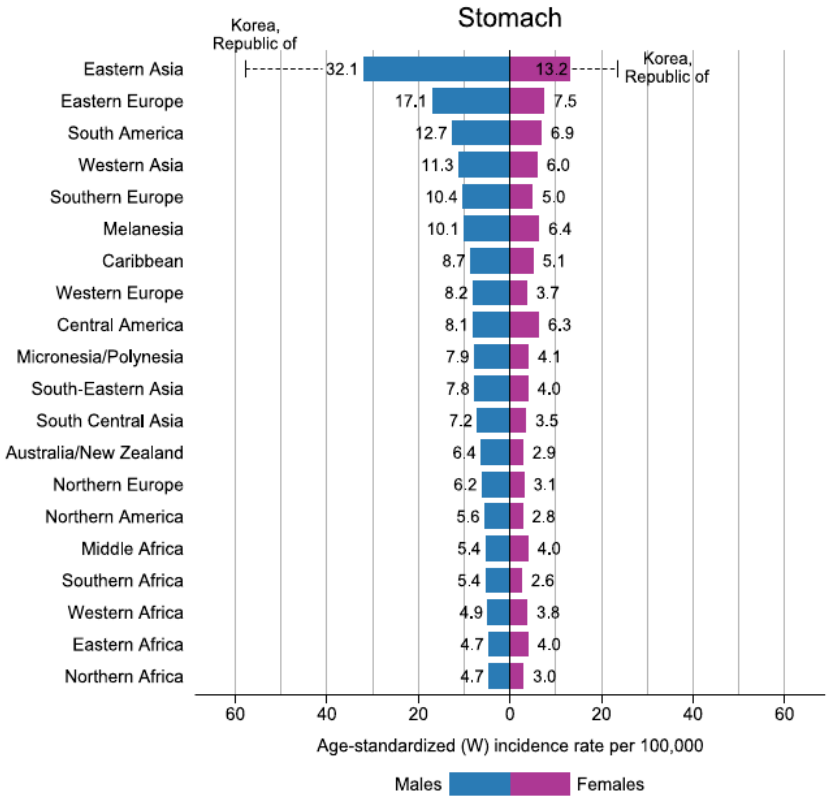


Figura 2.1.4: Incidencia estandarizada por edad del cáncer gástrico en distintas partes del mundo para ambos sexos. Fuente: Bray y col. [13].

2.1.1.1. Cáncer gástrico en Chile

Chile es uno de los países donde el cáncer gástrico es más común. De acuerdo a datos de Globocan [28], durante el 2018 en Chile la incidencia estandarizada por edad para ambos sexos fue de 17,8/100.000 habitantes, mientras que la mortalidad estandarizada por edad para el mismo periodo fue de 11,5/100.000 habitantes. De esta forma, Chile es el séptimo país con mayor incidencia de cáncer gástrico en el mundo, y el número catorce con mayor mortalidad por esta enfermedad. Además, dentro del país, es el cuarto cáncer más común, sólo después del cáncer de próstata, colorrectal y de mama, y el segundo más mortal, solamente precedido por el cáncer de pulmón, tal como se aprecia en la figura 2.1.5. La incidencia y mortalidad desglosada por sexo se detalla en la tabla 2.1.1. Por otra parte, se debe mencionar que la distribución de incidencia y mortalidad dentro del país tampoco es uniforme, con zonas de alta mortalidad (Aysén, Araucanía, Los Lagos, Los Ríos, Bio-Bío y Maule) y zonas de mortalidad intermedia (zona central) [59] [35].

Si bien estas cifras son altas, y pese a que programas de diagnóstico temprano han sido llevados a cabo con relativo éxito en Japón y Corea, el Ministerio de Salud de Chile no recomienda realizar un tamizaje poblacional masivo para la pesquisa precoz de la enfermedad, debido a que no es costo efectivo y a limitaciones logísticas. Aún así, sí se recomienda un tamizaje selectivo en adultos con síntomas tempranos de CG o con historial familiar de la enfermedad [59].

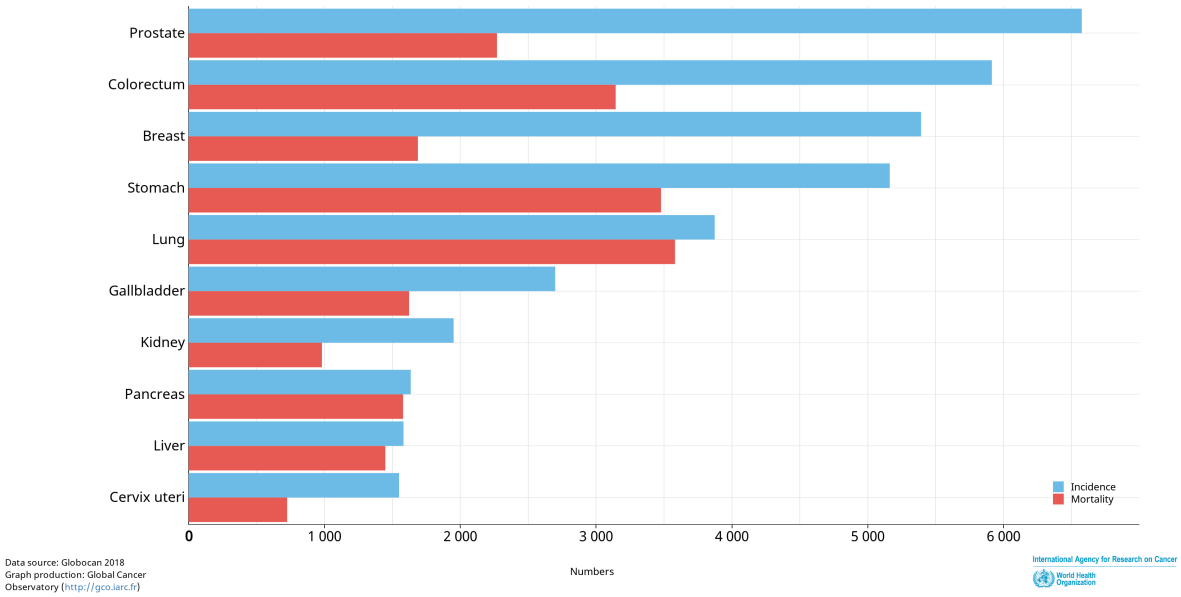


Figura 2.1.5: Incidencia y mortalidad estimada para distintos tipos de cáncer en Chile. Fuente: Ferlay y col. [28].

	Ambos sexos	Hombres	Mujeres
Incidencia, estandarizada por edad, por cada 100.000 habitantes	17,8	26,9	10,3
Mortalidad, estandarizada por edad, por cada 100.000 habitantes	11,5	17,9	6,4

Tabla 2.1.1: Incidencia estandarizada por edad y mortalidad estandarizada por edad para Chile durante el año 2018, para ambos sexos, sólo hombres y sólo mujeres. Fuente: Elaboración propia con datos extraídos de Ferlay y col. [28].

2.1.2. Clasificación de cáncer gástrico

El cáncer gástrico (CG) se puede clasificar siguiendo varios esquemas. En particular, el CG se puede clasificar en base a su histología, y también puede ser categorizado en base a su estadio clínico.

2.1.2.1. Clasificación histológica

La histología es el estudio de las células y los tejidos bajo un microscopio. Así, dependiendo de las características microscópicas del tejido, se pueden establecer distintas clasificaciones histológicas. Dos de las clasificaciones más usadas son la clasificación de la Organización Mundial de la Salud y la clasificación de Lauren.

2.1.2.1.1 Clasificación de Organización Mundial de la Salud

La Organización Mundial de la Salud (OMS) publicó en 2010 una detallada clasificación de los tumores del sistema digestivo [12]. Referente a los tumores del estómago, las cuatro grandes clasificaciones corresponden a:

- *Tumores epiteliales*: tumores que aparecen en el epitelio estomacal. El epitelio es una capa delgada de tejido que cubre los órganos, las glándulas y otras estructuras dentro del cuerpo; en el caso del estómago, el epitelio forma parte de la mucosa estomacal.
- *Tumores mesenquimales*: aquellos tumores que se desarrollan en el tejido conectivo del estómago. Aquí se encuentran los sarcomas, y los tumores del estroma gastrointestinal (GIST, por sus siglas en inglés).
- *Linfomas*: aquellos cánceres que se desarrollan en las células del sistema inmunitario.
- *Tumores secundarios*.

De acuerdo a la *Encyclopedia of Cancer*, el 95 % de los cánceres gástricos corresponden a tumores de origen epitelial, los cuales son conocidos como *carcinomas*. Estos se originan en la capa más interna del estómago, la mucosa, la cual a su vez está dividida en tres partes: *epitelio superficial*, *lámina propia* y *muscularis mucosa*.

Dentro de los carcinomas, y también de acuerdo a la OMS, podemos encontrar los *adenocarcinomas*, un tipo de cáncer que comienza en las células glandulares (secretoras). Estas células "se encuentran en el tejido que reviste ciertos órganos internos; producen y liberan sustancias en el cuerpo, como el moco, los jugos digestivos u otros líquidos" [62]. Un cáncer de estómago suele ser la mayoría de las veces un adenocarcinoma. [6]. Dentro de los adenocarcinomas, se distinguen los subtipos papilar, tubular, mucinoso, carcinoma pobremente cohesivo (incluyendo adenocarcinomas con células en anillo de sello y otras variantes) y mixto. Otros carcinomas reconocidos según esta clasificación son carcinoma adenoescamoso, carcinoma con estroma linfoide, adenocarcinoma hepatoide, carcinoma de células escamosas y carcinoma indiferenciado; todos estas últimas variantes histológicas son raras y representan alrededor de un 5 % de los casos de cáncer gástrico.

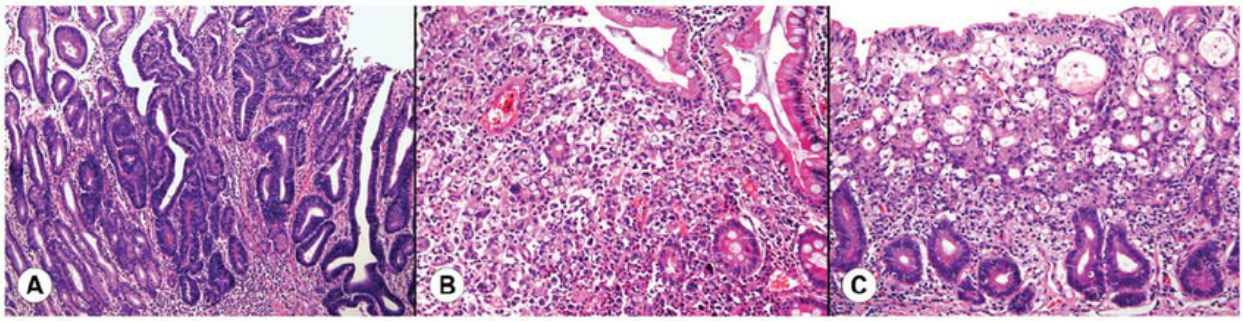


Figura 2.1.6: Clasificación de Lauren para carcinomas gástricos tempranos: A) tipo intestinal, B) tipo difuso y C) tipo mixto. Fuente: Chong y col. [18].

Clasificación de Organización Mundial de la Salud (2010)	Clasificación de Lauren (1965)
Adenocarcinoma papilar Adenocarcinoma tubular Adenocarcinoma mucinoso	Tipo intestinal
Carcinoma con células en anillo de sello y otros carcinomas pobremente cohesivos	Tipo difuso
Adenocarcinoma mixto	Tipo indeterminado
Carcinoma adenoescamoso Carcinoma de células escamosas Adenocarcinoma hepatoide Carcinoma con estroma linfoide Coriocarcinoma Carcinosarcoma Carcinoma de células parietales Tumor rabdoide maligno Carcinoma mucoepidermoide Carcinoma de células de Paneth Carcinoma indiferenciado Carcinoma mixto adenoneuroendocrino Tumor del seno endodérmico Carcinoma embrionario Tumor del saco de yolk puramente gástrico Adenocarcinoma oncocitico	

Tabla 2.1.2: Relación entre clasificaciones de OMS y Lauren. Fuente: Hu y col. [39].

2.1.2.1.2 Clasificación de Lauren

La clasificación de Lauren categoriza los adenocarcinomas gástricos en dos tipos: *intestinal* y *difuso* [53]. El tipo *intestinal* forma glándulas con varios niveles de diferenciación, mientras que el tipo *difuso* consiste de células pobremente cohesionadas, con poca o nula formación de glándulas [18]. Además, tumores que tienen una cantidad aproximadamente igual de tipo difuso e intestinal son catalogados como tumores *mixtos*, y el resto queda en la categoría de tumores *indeterminados*.

2.1.2.2. Clasificación en base al estadio clínico

La estadificación del cáncer se refiere a su extensión, su tamaño y si es que se ha expandido a otras partes del cuerpo. Esta información es usada por el equipo clínico para determinar el tratamiento a seguir y hacer un pronóstico de la enfermedad.

El sistema de estadificación más usado para el cáncer gástrico es el sistema propuesto por el Comité Conjunto Americano del Cáncer (AJCC, por sus siglas en inglés), y es también conocido como *TNM* [7]. Éste es usado para clasificar todos los cánceres de estómago, excepto aquellos que se originan en la unión gastroesofágica o aquellos que se originan en el cardias. El significado de cada una de las letras es el siguiente:

- la T se refiere al tumor primario, su tamaño y extensión,
- la N se refiere a la extensión de cáncer que se ha diseminado a los ganglios (o nódulos) linfáticos cercanos,
- y la M se refiere a la metastatización del tumor; es decir, si es que el cáncer ha hecho *metástasis* y se ha diseminado a otras partes del cuerpo, como por ejemplo, el hígado o el pulmón.

A cada letra va asociado un número, expresando la característica correspondiente. Por lo general, un número más alto indicará un peor diagnóstico (tumor más extendido, mayor número de nodos linfáticos afectados, o presencia de metástasis). Así, una descripción en el sistema TNM puede ser T1N1M1 o T4aN3bM0, por ejemplo. La explicación de los números asociados es:

- **Tumor primario (T):**
  - TX: no se puede determinar la presencia de un tumor primario.
  - T0: sin evidencia de tumor primario.
  - Tis: carcinoma in situ; se refiere a un tumor intraepitelial (capa más interna del estómago), sin invasión de la lámina propia.
  - T1: Tumor invade la lámina propia, la muscularis mucosa o la submucosa.
    - T1a: Invade la lamina propia o la muscularis mucosa (aún en la mucosa).
    - T1b: Invade la submucosa (siguiente capa después de la mucosa).
  - T2: Tumor invade la muscularis propia (tercera capa del estómago).
  - T3: Tumor penetra en el tejido conectivo de la subserosa (cuarta capa del estómago), sin invasión del peritoneo visceral ni de las estructuras adyacentes.
  - T4: Tumor invade la serosa (quinta capa del estómago) o estructuras adyacentes.
    - T4a: Tumor invade la serosa.
    - T4b: Tumor invade las estructuras adyacentes.
- **Ganglios regionales linfáticos (N):**
  - NX: No se puede medir el cáncer en los nodos linfáticos regionales.
  - N0: No hay cáncer en los ganglios linfáticos cercanos.
  - N1: Metástasis en uno o dos ganglios linfáticos regionales.
  - N2: Metástasis en tres a seis ganglios linfáticos regionales.
  - N3: Metástasis en siete o más ganglios linfáticos regionales.
    - N3a: Metástasis en siete a quince ganglios.
    - N3b: Metástasis en dieciséis o más nodos linfáticos.
- **Metástasis distante (M):**
  - M0: el cáncer no se ha diseminado a otras partes del cuerpo.

– M1: metástasis en partes distantes del cuerpo.

Luego de realizar la caracterización TNM, esta información se agrupa y se determina en qué *etapa* está el tumor. A mayor etapa, peor es el pronóstico correspondiente. En cáncer gástrico, existen al menos dos tipos de etapas: la *patológica* y la *clínica*. La etapa patológica (también llamada etapa quirúrgica) es determinada usando el tejido extraído en una operación, mientras que la etapa clínica es utilizada cuando no se puede realizar una cirugía, y es llevada a cabo usando exámenes físicos, biopsias no quirúrgicas y estudios por imágenes. De todas maneras, es usual que la etapa clínica sea menos precisa que la etapa patológica. Las tablas 2.1.3 y 2.1.4 detallan las distintas etapas patológicas y clínicas del cáncer gástrico.

Por otro lado, además de reportar el tamaño del tumor y su posible extensión a ganglios y órganos distantes, también se suele reportar su grado histológico, que corresponde a una descripción de cuán anormales se ven las células y tejidos cancerosos bajo un microscopio. Así, las células con un grado bajo se dice que son *bien diferenciadas*, ya que tienen estructuras y funciones especializadas. Por su parte, las células con un grado alto suelen ser *indiferenciadas*; es decir, no están especializadas ni maduras. Mientras menor sea el grado de un tumor, más se parecerán sus células a las células normales del tejido sano, y tienden a multiplicarse y diseminarse más lentamente que células de grado alto. En el caso de cáncer gástrico, la AJCC indica que, si ningún sistema de gradación es especificado, es porque se hace uso del siguiente sistema [7]:

- GX: no es posible asignar un grado.
- G1: bien diferenciado (grado bajo).
- G2: moderadamente diferenciado.
- G3: escasamente diferenciado.
- G4: indiferenciado (grado alto).

2.1.3. Proteína HER2

HER2 (también conocida por los nombres de *HER2/neu*, *ErbB-2* o *receptor 2 del EGF humano*) es una proteína transmembrana receptora de tirosina quinasa, con un peso molecular de 185 kDa, y es parte de la familia de receptores para el factor de crecimiento epidérmico (EGF, por sus siglas en inglés). Esta proteína es codificada por el gen *HER2*, localizado en el cromosoma 17 [34] [62], y es reconocido como un proto-oncogen; es decir, un gen celular normal que, cuando es activado por mutaciones, adquiere funciones oncogénicas [80] (i.e, tiene la habilidad de estimular el crecimiento celular). Esto, porque una alta amplificación del gen *HER2* induce una sobreexpresión de la proteína HER2 en la membrana celular, lo cual genera muchos receptores transmitiendo señales de proliferación celular al núcleo [34] [20].

En 2001, un estudio demostró que, en casos de cáncer de mama con sobreexpresión de la proteína HER2, un tratamiento con quimioterapia y Trastuzumab (un anticuerpo monoclonal que se une a HER2 e inhibe su vía de señalización) provoca que transcurra un mayor tiempo hasta la progresión de la enfermedad (periodo de tiempo que pasa desde el diagnós-

Etapa patológica	Agrupamientos TNM posibles
0	Tis, N0, M0
IA	T1, N0, M0
IB	T1, N1, M0 T2, N0, M0
IIA	T1, N2, M0 T2, N1, M0 T3, N0, M0
IIB	T1, N3a, M0 T2, N2, M0 T3, N1, M0 T4a, N0, M0
IIIA	T2, N3a, M0 T3, N2, M0 T4a, N1-2, M0 T4b, N0, M0
IIIB	T1-2, N3b, M0 T3-4a, N3a, M0 T4b, N1-2, M0
IIIC	T3-4a, N3b, M0 T4b, N3a-b, M0
IV	Cualquier T, Cualquier N, M1

Tabla 2.1.3: Etapas patológicas del cáncer gástrico. Fuente: González [33].

Etapa clínica	Agrupamientos TNM posibles
0	Tis, N0, M0
I	T1-2, N0, M0
IIA	T1-2, N1-3, M0
IIB	T3-4a, N0, M0
III	T3-4a, N1-3, M0
IVA	T4b, N1-3, M0
IVB	Cualquier T, Cualquier N, M1

Tabla 2.1.4: Etapas clínicas del cáncer gástrico. Fuente: González [33].

tico o comienzo del tratamiento hasta que el cáncer comienza a empeorar), mejores tasas de respuesta, mayor tiempo de sobrevida y reducción en la probabilidad de muerte por la enfermedad [83]. Dado que este tratamiento es útil sólo en casos de cáncer con sobreexpresión de HER2 (también llamados HER2 positivos), en 2007 la Sociedad Americana de Oncología Clínica (ASCO) y el Colegio Americano de Patólogos (CAP) generaron guías clínicas con recomendaciones sobre como evaluar la sobreexpresión de dicha proteína en cáncer de mama [90].

En el caso de cáncer gástrico, se estima que entre un 9-38 % de los tumores son HER2 positivos, presentando variaciones dependiendo de la ubicación anatómica (mayor frecuencia de sobreexpresión en tumores de la unión gastroesofágica en comparación con tumores del estómago), del tipo histológico (tipo intestinal tiene más frecuencia que tipo difuso) y la diferenciación (bien y medianamente diferenciado con mayor frecuencia que pobremente diferenciado) [4]. Fue en 2010 cuando un estudio clínico (conocido como ToGA) demostró que un tratamiento con trastuzumab y quimioterapia prolonga de forma estadísticamente significativa la sobrevida de los pacientes [9]. Así, y debido a que existen importantes diferencias en la expresión y clasificación de HER2 entre cáncer gástrico y cáncer de mama, en 2017 ASCO, CAP y la Sociedad Americana de Patología Clínica (ASCP) publicaron una

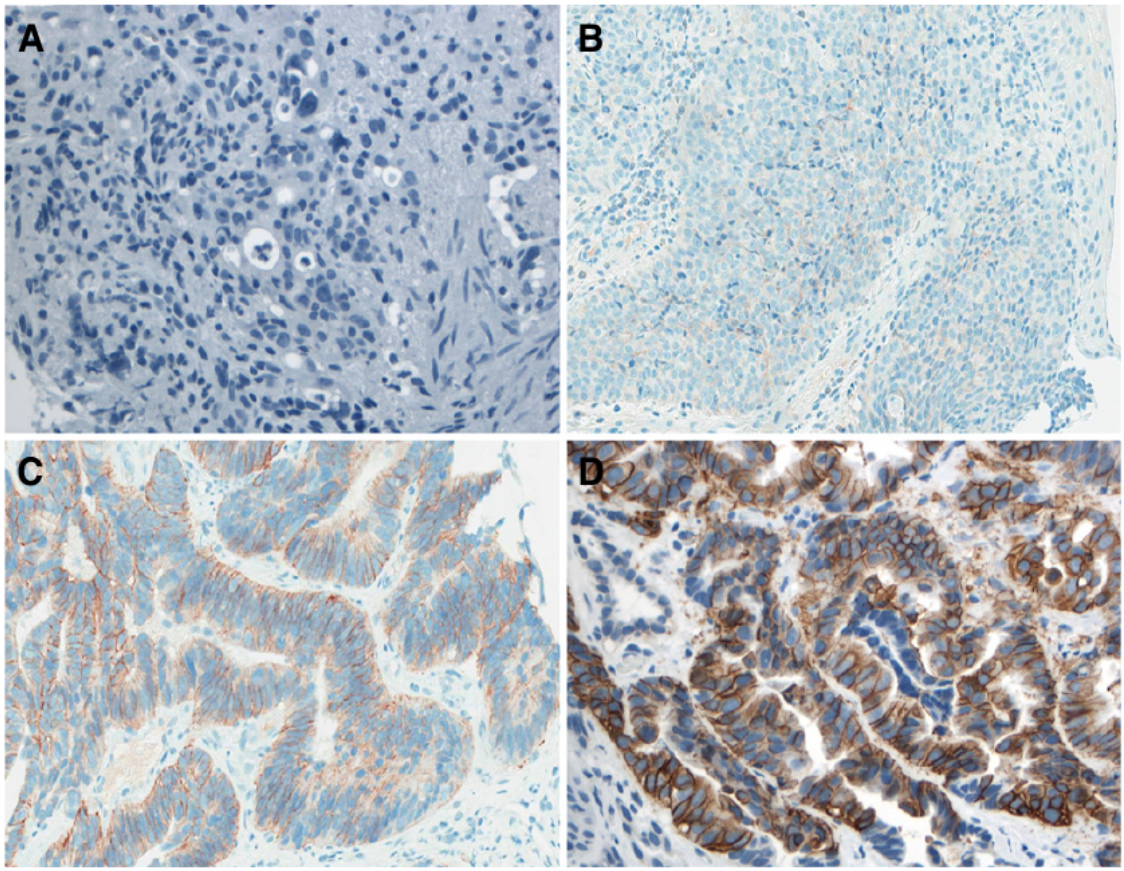


Figura 2.1.7: Análisis inmunohistoquímico en muestras representativas de expresión de HER2 en cáncer gástrico. A) 0, negativo. B) 1+, negativo. C) 2+, equívoco. D) 3+, positivo. Fuente: Bartley y col. [10].

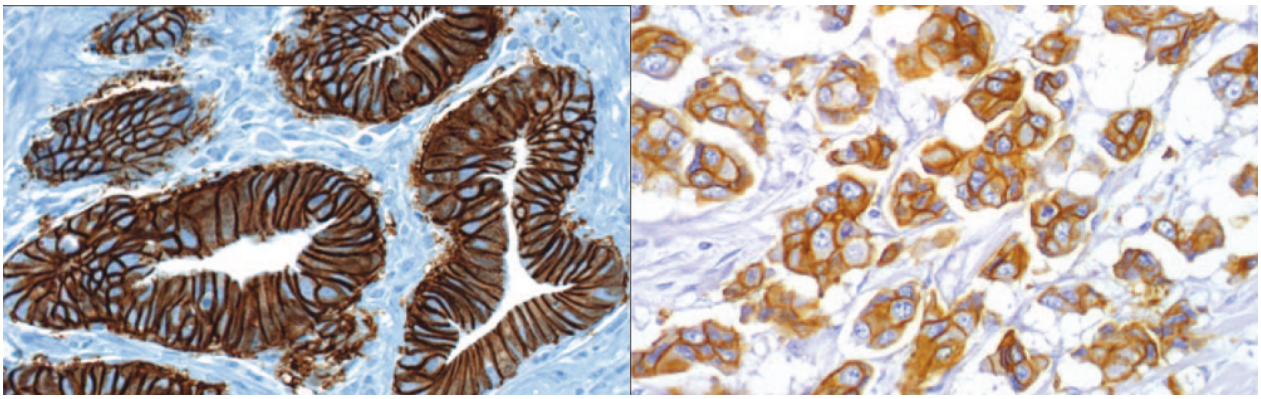
guía con recomendaciones para la evaluación de HER2 en cáncer gástrico; específicamente, para adenocarcinomas gastroesofágicos [10].

En dicha guía clínica, se establece que para pacientes con adenocarcinomas gastroesofágicos avanzados, se debe analizar la presencia de HER2 en tejido canceroso. Este tejido puede ser extraído mediante biopsia endoscópica (es decir, una muestra extraída durante la realización de una endoscopia) o mediante resección (extirpación quirúrgica del tumor que es luego analizada). Además, como alternativa también se pueden utilizar especímenes obtenidos mediante aspiración con aguja fina. Luego, si es que en el tejido se detecta sobreexpresión de la proteína, se debe comenzar el tratamiento enfocado en HER2.

Para detectar la presencia de HER2, se recomienda realizar en primera instancia un análisis *inmunohistoquímico* (IHC, por sus siglas en inglés), el cual consiste en generar una reacción antígeno-anticuerpo en una biopsia, con lo cual se puede visualizar la presencia de la sustancia buscada; en este caso, la proteína HER2. De esta forma, la biopsia queda teñida con un color azul en las células y café en las zonas con sobreexpresión de HER2. Para clasificar la sobreexpresión de HER2, la guía clínica sugiere usar el método de Ruschoff/Hofmann, el cual se detalla en la tabla 2.1.5 [76] [38]. Además, la figura 2.1.7 muestra zonas con distinta sobreexpresión HER2, clasificadas siguiendo este método.

Pese a que existen similitudes en cómo se debe evaluar el nivel de HER2 en cáncer gástrico y cáncer de mama, también existen diferencias mayores, tales como (i) la mayor heterogeneidad de la sobreexpresión HER2 en cáncer gástrico, (ii) la presencia de membranas no completamente teñidas, localizadas en la porción apical de la membrana celular en cáncer gástrico, lo cual raramente se da en cáncer de mama (Figura 2.1.8), y (iii) una mayor discordancia entre FISH e IHC en cáncer gástrico [73]. Además, esta heterogeneidad en cáncer gástrico se puede manifestar de variadas formas, tales como la expresión de HER2 variando entre 1+ a 3+ en el mismo tumor o diferentes expresiones en componentes morfológicamente distintas del tumor [5]. Debido a esto es que se recomienda que el análisis de HER2 sea realizado en el tumor primario obtenido por resección, o en múltiples muestras obtenidas mediante biopsia [10]. De esta manera, la evaluación de HER2 en cáncer gástrico es diferenciada dependiendo de si la muestra es endoscópica o resección, a diferencia de lo que ocurre en cáncer de mama, donde no se hace distinción respecto al método de obtención del tejido. Además, la misma guía clínica señala que la evaluación de HER2 mediante IHC es una técnica subjetiva, pues dependerá de la interpretación que cada patólogo haga de la biopsia; esta variabilidad interobservador puede surgir por errores al calcular el porcentaje de células teñidas o errores al estimar la intensidad de la tinción, lo cual dificulta la reproducibilidad del análisis [11].

En caso de que el análisis IHC entregue un resultado equívoco (es decir, se obtenga una puntuación 2+), se debe proceder a realizar un análisis de *hibridación in situ* (ISH, por sus siglas en inglés). Este tipo de técnicas se utilizan para localizar y detectar secuencias de ADN o ARN en células, cromosomas o tejidos. En el caso de HER2, lo que se busca es determinar la amplificación del gen *HER2*, siendo la *hibridación in situ fluorescente* (FISH, por sus siglas en inglés) el método más utilizado. Como se aprecia en la figura 2.1.9, si bien tanto IHC como ISH son pruebas que sirven para establecer si es que un tumor es HER2 positivo o no, ambos métodos tienen distintos enfoques: mientras IHC detecta la sobreexpresión de la proteína HER2, ISH detecta la amplificación del gen *HER2*.



(a) Cáncer gástrico (b) Cáncer de mama

Figura 2.1.8: Imágenes de cáncer gástrico y cáncer de mama teñidas inmunohistoquímicamente. Ambos tumores son HER2-positivos (IHC 3+), pero se aprecian diferencias importantes: mientras en el cáncer de mama se verifica que cada célula inmunopositiva presenta una tinción completa de su membrana, en el cáncer gástrico esto no ocurre, teniéndose que muchas membranas no están teñidas completamente y quedan "abiertas". Fuente: Ross y Mulcahy [73].

Patrón de teñido en espécimen quirúrgico	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, basolateral o 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.

Tabla 2.1.5: Pauta de clasificación para interpretación de inmunohistoquímica HER2 en carcinoma gástrico. *Un *cluster* de células tumorales se define como un grupo de 5 o más células neoplásicas. Traducido desde Bartley y col. [10].

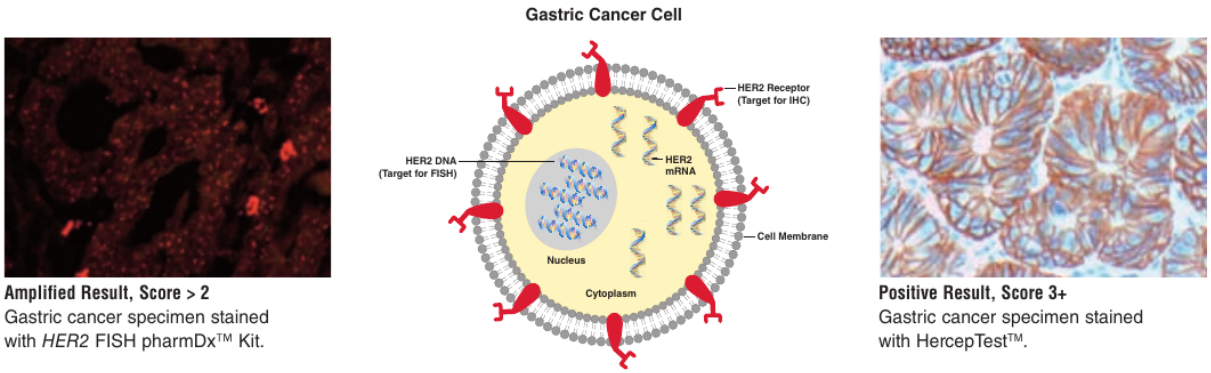


Figura 2.1.9: ISH e IHC y sus respectivos objetivos de análisis; ejemplo en cáncer gástrico. Fuente: Dako [20].

2.2. Imágenes digitales y patología digital

Una *imagen digital*, en escala de grises, puede ser definida como una función $f(x, y)$, donde x, y son coordenadas espaciales, con la particularidad de que tanto x, y y los valores de f son todas cantidades discretas y finitas. Además, la amplitud de f en cualquier par de coordenadas es conocida como *intensidad de gris*. Así, una imagen digital está conformada por un número finito de elementos, cada uno de los cuales tiene una ubicación e intensidad; a cada uno de estos elementos se les conoce como *pixel*.

Para ampliar esta definición a imágenes en colores, primero es necesario definir un modelo de colores; un sistema de coordenadas donde cada color es representado como un punto del espacio. Algunos de los modelos más utilizados son RGB (rojo, verde y azul), CMY (cian, magenta y amarillo) y HSI (matiz, saturación e intensidad). Luego, en cada uno de estos sistemas, un color es representado por tres valores, con lo cual una imagen puede ser representada como un conjunto de tres matrices, una para cada componente del modelo (Figura 2.2.1).

Algo importante de considerar es que, siguiendo la descripción previa, para un computador una imagen es sólo un conjunto de números. Tal como se describe en Hare y col. [37], "las representaciones que se pueden computar desde imágenes en bruto no se pueden transformar fácilmente a descripciones de alto nivel de los conceptos semánticos que la imagen posee". Así, si bien para un humano es fácil reconocer que en una imagen hay un animal o una fruta, lograr que un computador realice ese reconocimiento no es una tarea trivial; a este problema se le suele conocer como *brecha semántica* (o *semantic gap* en inglés).

En el caso particular de patología digital, las placas físicas (o portaobjetos) son digitalizadas utilizando escáneres especialmente diseñados para tal tarea. Las imágenes generadas son de alta resolución, con una magnificación de hasta 40x y una cantidad del orden de 10 millones de pixeles; a resolución completa y sin comprimir, estas imágenes pueden pesar decenas de gigabytes, excediendo la memoria principal normalmente disponible en un computador. Para salvar estas limitaciones, las imágenes suelen ser almacenadas en un formato piramidal, siendo cada capa de la pirámide una magnificación distinta. Así, con software especial, se

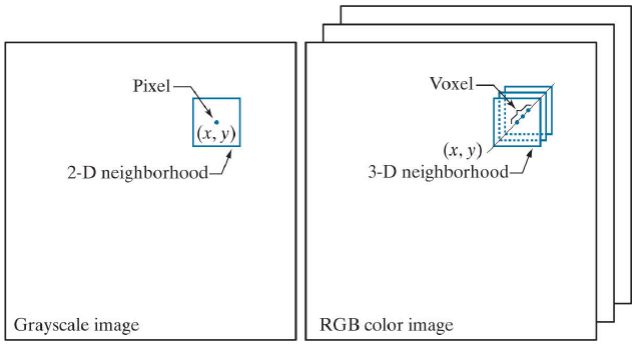


Figura 2.2.1: Representación de una imagen digital como un conjunto de tres matrices. Fuente: Gonzalez y Woods [32].

puede observar toda la imagen en forma panorámica (en las magnificaciones de menor resolución) y a la vez acercarse a cada región de la imagen y observar los detalles en mayor resolución, de forma análoga a como se hace con la visualización de mapas.

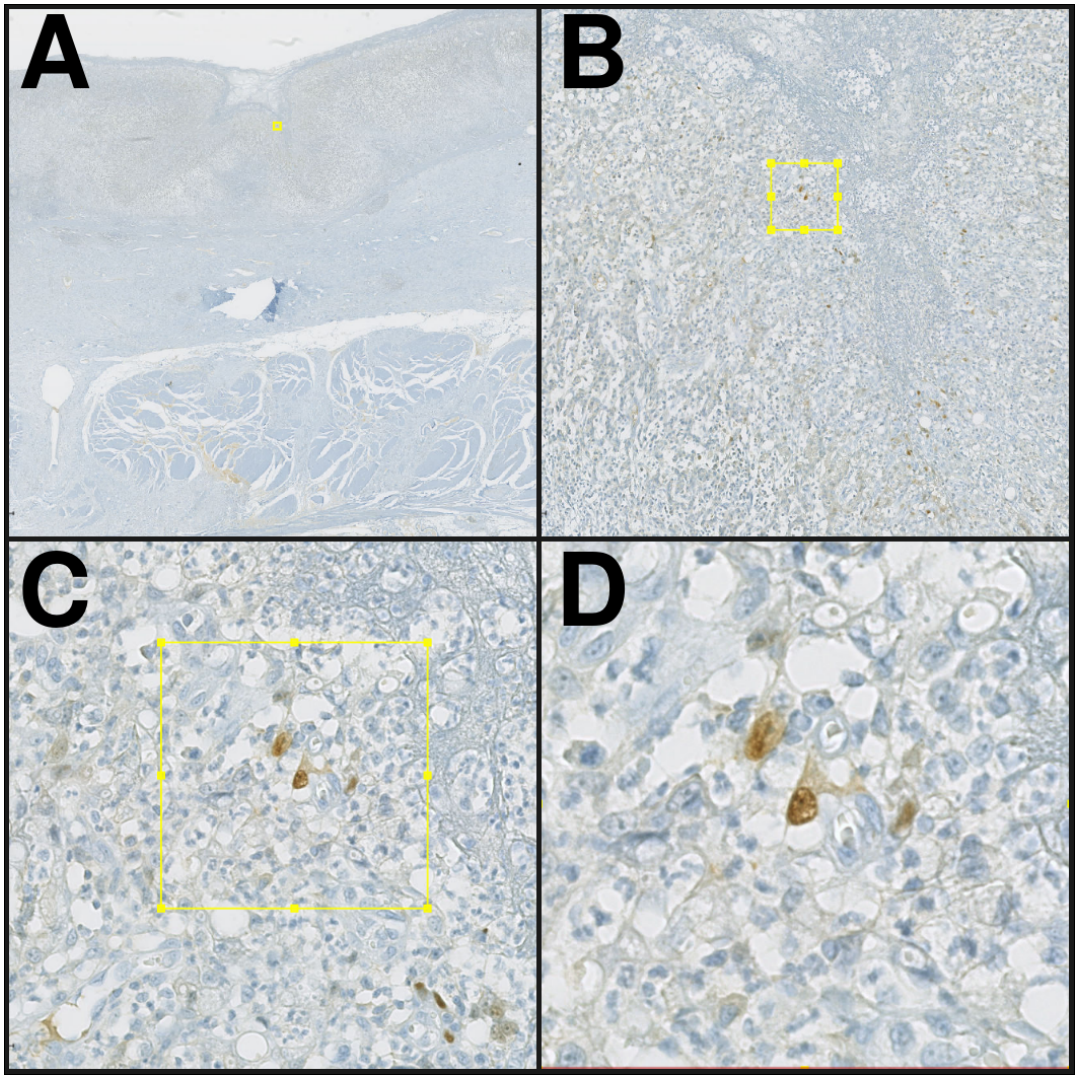


Figura 2.2.2: Ejemplo de imagen piramidal con múltiples magnificaciones. En este caso, biopsia de cáncer gástrico con tinción inmunohistoquímica. A) 0.5x, B) 5x, C) 20x, D) 40x. Fuente: Elaboración propia con datos provenientes del estudio PRECISO [61].

2.3. Aprendizaje de máquinas

El aprendizaje de máquinas, más conocido por su nombre en inglés *machine learning* (ML), es un concepto que no tiene una única definición. Para Tom Mitchell, la mejor forma de definir un campo científico es mediante su pregunta central; en el caso de ML, esta pregunta sería «¿Cómo podemos construir sistemas computacionales que automáticamente mejoren al adquirir experiencia, y cuáles son las leyes fundamentales que gobiernan todos los procesos de aprendizaje?». De acuerdo al mismo autor, el aprendizaje de máquinas se construye

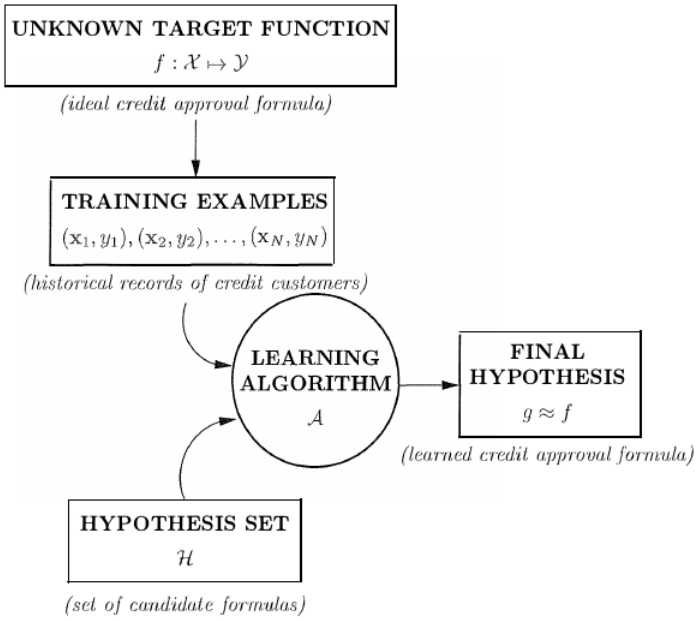


Figura 2.3.1: Esquema básico de un problema típico de aprendizaje de máquinas, con ejemplos relativos a la pregunta de si aprobar o rechazar créditos bancarios a potenciales clientes. Fuente: Abu-Mostafa, Magdon-Ismail y Lin [2].

como una intersección entre los campos de la computación y la estadística. Así, mientras que la Ciencia de la Computación tradicionalmente se ha enfocado en cómo programar manualmente una máquina, el área de ML se enfoca en cómo lograr que los computadores se programen a sí mismos, utilizando experiencia -o datos- y una estructura inicial. Y asimismo, mientras la estadística busca responder qué se puede inferir de un conjunto de datos y con cuánta confianza, ML incorpora preguntas sobre complejidad de los algoritmos, capacidad de almacenamiento, etc. [60].

Por otra parte, en el libro de Abu-Mostafa, Magdon-Ismail y Lin [2], el aprendizaje de máquinas es definido como el campo dedicado al estudio del aprendizaje mediante datos (y su nombre hace directa contraposición al aprendizaje en humanos). Así, el aprendizaje de máquinas es usado en situaciones donde no existe (o no se puede obtener) una solución analítica, pero donde sí se tiene un conjunto de datos para generar una solución empírica. Formalizando esta intuición, se plantea que en un problema de ML existe un espacio de entrada $\mathcal{X}$, un espacio de salida $\mathcal{Y}$ y una función objetivo desconocida $f : \mathcal{X} \longrightarrow \mathcal{Y}$, donde la tarea es encontrar una función $g : \mathcal{X} \longrightarrow \mathcal{Y}$ que sea parecida a f. Además, pese a que f es desconocida, sí se posee un conjunto de datos (o *dataset*) $\mathcal{D}$ formado por varios pares $(x_\mathrm{i}, y_\mathrm{i})$, tales que $f(x_\mathrm{i}) = y_\mathrm{i}$. Finalmente, se define un conjunto de funciones *candidatas* llamado $\mathcal{H}$ y un algoritmo de entrenamiento $\mathcal{A}$ que a partir de $\mathcal{D}$ escogerá la función $g \in \mathcal{H}$ [2]. Este proceso se encuentra esquematizado en la figura 2.3.1.

Esta definición es útil para pensar el aprendizaje de máquinas en términos matemáticos y comprender las distintas componentes que toman parte del proceso: un conjunto de datos, una función desconocida que se desea aproximar y un algoritmo que genera dicha aproximación a partir de los datos.

2.3.1. Aprendizaje supervisado y no supervisado

El esquema descrito en la subsección anterior, e ilustrado en la figura 2.3.1, corresponde a un caso de aprendizaje supervisado, donde los datos disponibles son tuplas de la forma $(x_{\mathrm{i}}, y_{\mathrm{i}})$ y $f(x_y) = y_{\mathrm{i}}$; es decir, para cada entrada sabemos cuál es la salida correcta. Se le llama aprendizaje supervisado, dado que algún "supervisor" ha categorizado cada valor con su salida correspondiente. Muchos métodos de ML suelen ser usados en este esquema supervisado, tales como regresión lineal, bosques aleatorios (*random forest*), máquinas de soporte vectorial (*support vector machines*) y redes neuronales, entre otros. [42].

Por otra parte, existen casos donde el conjunto de datos sólo contiene datos de tipo (x_{i}); es decir, se dispone de un conjunto de mediciones, pero no del valor asociado a ellas. En estos casos, ya no se trata de aprender una función $g : \mathcal{X} \longrightarrow \mathcal{Y}$ que relaciona entradas y salidas, sino de aprender patrones dentro de los datos, y agruparlos de acuerdo a dichos patrones; este tipo de métodos corresponde a aprendizaje no supervisado.

2.3.2. Clasificación y regresión

Dentro del esquema de aprendizaje supervisado, suelen distinguirse dos tipos de problemas: aquellos donde la variable a predecir es *cuantitativa* (continua) y aquellos donde es *cualitativa* (discreta). Ejemplos de variables cuantitativas son la altura de una persona, el precio de algún bien o la temperatura de una habitación. Por otra parte, variables cualitativas, o categóricas, son el color de un semáforo (verde, amarillo o rojo), raza de un perro (dálmata, beagle, quiltro) o la presencia de tejido canceroso en una biopsia (sí/no). De esta forma, cuando la variable que se desea predecir (los y_{i} en nuestro conjunto de datos) es categórica, se dirá que se trata de un problema de *clasificación*, mientras que si es una variable cualitativa, se tratará de un problema de *regresión*.

Algo importante de destacar es que es la variable de respuesta la que determina si se trata de un problema de clasificación o regresión; es decir, depende de si y_{i} es continua o discreta, y no de si x_{i} es cuantitativa o categórica. En general, muchos métodos de ML se pueden aplicar sin importar el tipo de variable que sea x_{i}, siempre y cuando estos vectores hayan sido debidamente codificados antes de realizar cualquier análisis [42].

2.3.3. Teoría de aprendizaje de máquinas

Como se mencionó previamente, en el contexto de aprendizaje de máquinas supervisado, lo que se busca es encontrar una función $g : \mathcal{X} \longrightarrow \mathcal{Y}$ que sea "parecida" a la función objetivo f, y para eso se dispone de un conjunto de datos $\mathcal{D}$. No obstante, es necesario recordar que lo importante es obtener una función g que *generalice* bien; es decir, que tenga un buen rendimiento en datos nuevos, fuera de $\mathcal{D}$. Así, surge naturalmente la pregunta de si es realmente posible *aprender* una función con potencialmente infinitas entradas distintas desde un conjunto finito de datos.

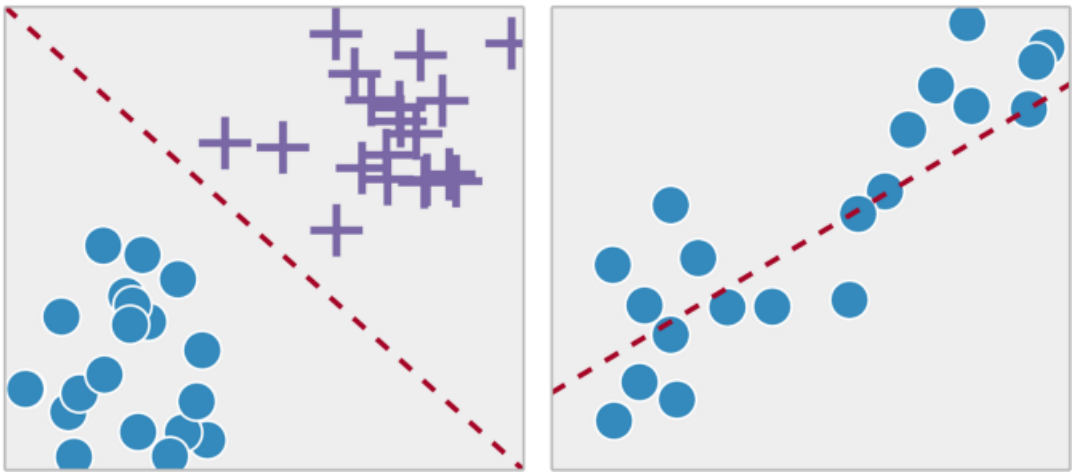


Figura 2.3.2: Izquierda: ejemplo de problema de clasificación binaria; el objetivo es encontrar una función que permita separar los datos correctamente. Derecha: ejemplo de problema de regresión; el objetivo es encontrar una función que se aproxime a la distribución de los datos. Fuente: Soni [84].

Desde un punto de vista probabilístico, es posible dar una respuesta afirmativa a esa pregunta, pero para ello es necesario formalizar antes algunas intuiciones. En primer lugar, es necesario establecer cómo se genera el conjunto de datos $\mathcal{D}$; para ello, se asumirá que los datos son escogidos independientemente de acuerdo a una distribución de probabilidad P sobre el espacio $\mathcal{X}$. Luego, para formalizar la idea de funciones "parecidas", se introduce el concepto de función de error, la cual cuantifica cuán discordantes son dos funciones. Si bien esta función de error suele escogerse de manera *ad hoc* de acuerdo a las especificidades del problema, por ahora se definirá el *error dentro de la muestra* E_{in} para una función candidata h cualquiera como

$$E_{\text{in}} = \frac{1}{N}\sum_{n=1}^{N}[\![h(x) \neq f(x)]\!], \tag{2.3.1}$$

donde $[\![z]\!]$ es una función que toma como entrada una expresión lógica z, y retorna 1 si es que z es verdadera y 0 en caso contrario. Del mismo modo, se puede definir el error fuera de la muestra E_{out} como

$$E_{out} = \mathbb{P}[h(x) \neq f(x)], \tag{2.3.2}$$

donde $\mathbb{P}$ está basada en la distribución P usada para muestrear el conjunto de datos $\mathcal{D}$ desde $\mathcal{X}$. Es importante recordar que si bien E_{in} es un valor medible (dado que h es una función que el algoritmo de ML escogerá, y $f(x)$ es conocida para todos los $x \in \mathcal{D}$), E_{out} es desconocida, ya que depende de los valores de h y f en todos los puntos del espacio $\mathcal{X}$.

No obstante, usando las formalizaciones recién introducidas, es posible demostrar mediante teoremas estadísticos que el error fuera de la muestra se aproxima al error dentro de la muestra; es decir, $E_{out} \approx E_{\text{in}}$. En otras palabras, esto significa que si el dataset $\mathcal{D}$ está bien muestreado (es decir, es una buena representación de los datos que se utilizarán para evaluar), entonces lo aprendido con ese conjunto de datos sí es extrapolable a datos nuevos. De esta forma, podemos utilizar E_{in} como un *proxy* para estimar E_{out}, y así, el objetivo de obtener una función con E_{out} cercano a 0 se reduce a minimizar E_{in}.

Ahora bien, en ML existen muchas situaciones en que los datos no están generados de una

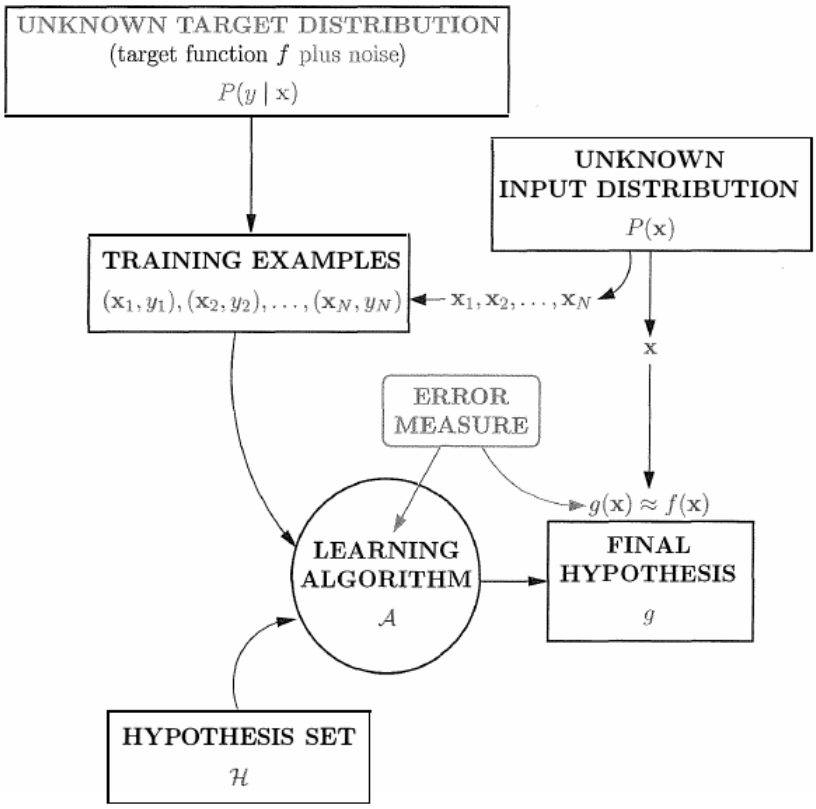


Figura 2.3.3: Esquema general de un problema de aprendizaje de máquinas supervisado, considerando distribuciones de probabilidad y ruido. Fuente: Abu-Mostafa, Magdon-Ismail y Lin [2].

forma completamente determinista, sino que existe un *ruido* asociado a los datos. Para modelar esto dentro del esquema de aprendizaje de máquinas, basta con hacer algunas pequeñas modificaciones: en lugar de tener que $y = f(x)$, se dirá que y es una variable aleatoria que está influenciada (mas no determinada) por x. En términos formales, ahora se buscará una distribución de probabilidad objetivo $P(y|x)$ en lugar de una función $y = f(x)$. Del mismo modo, cada dato (x_i, y_i) en D estará generado por la distribución conjunta $P(x_\text{i})P(y_\text{i}|x_\text{i})$. Bajo esta perspectiva, es posible pensar que un objetivo con ruido es simplemente una función determinista con ruido añadido; así, por ejemplo, el valor esperado de y es efectivamente $f(x)$, e $y - f(x)$ es solamente ruido presente en los datos.

La figura 2.3.3 muestra el esquema general de un problema de ML con ruido añadido y distribuciones de probabilidad. Sin embargo, al añadir el concepto de ruido al modelo, se abre la posibilidad de que el algoritmo $\mathcal{A}$ que seleccionará la función candidata g desde $\mathcal{D}$ también "aprenda" el ruido presente en los datos, en especial cuando el algoritmo tiene muchos parámetros que aprender en comparación al tamaño del dataset utilizado. Así, es posible obtener un g que obtenga un E_{in} muy bajo, ya que el algoritmo está aprendiendo (o más bien, memorizando) el ruido en la muestra, pero que a la vez tenga un E_{out} muy alto, dado que el ruido no es generalizable a datos nuevos. Esta situación es conocida como *sobreajuste* (u *overfitting* en inglés) y en el próximo apartado se verá cómo se puede subsanar este problema.

2.3.4. Conjuntos de entrenamiento, validación y evaluación

Como se mostró en el apartado previo, el problema de obtener una función g que generalice bien se puede reducir a obtener un g que tenga un E_{in} cercano a 0. Ahora bien, sin detallar aún cómo es que un algoritmo de ML opera para minimizar el error, existen algunas consideraciones que deben hacerse previamente, en especial respecto al *sobreajuste*.

Por lo general, un algoritmo de ML tiene varios *hiperparámetros*; parámetros que controlan ciertas características del modelo y que son independientes de los datos utilizados para entrenar. Así, utilizando el mismo algoritmo $\mathcal{A}$ y el mismo conjunto de datos de entrenamiento $\mathcal{D}$, es posible obtener funciones g muy distintas, tan sólo variando los hiperparámetros del modelo. Ante esto, surge la pregunta de cómo evaluar efectivamente qué modelo es mejor; es decir, qué función g tiene un menor E_{out}.

Una técnica común en ML es dividir el conjunto de datos en tres subconjuntos: *entrenamiento*, *validación* y *evaluación*. De esta forma, los distintos modelos son entrenados, usando distintos hiperparámetros, en el dataset de entrenamiento y se mide su comportamiento en el dataset de validación. Luego, se reajustan los hiperparámetros, se vuelve a entrenar y se itera. Finalmente, se escoge el mejor modelo (aquel que obtuvo el menor error en el conjunto de validación) y se evalúa en el conjunto de evaluación, siendo este resultado el que generalmente se reporta. Dentro del área de ML, se recomienda fuertemente que el conjunto de evaluación sea "ciego" a la hora de entrenar. Por ello, el ajuste de parámetros se realiza en el conjunto de validación, ya que al medir el desempeño en ese conjunto de datos y utilizar esa retroalimentación, parte de la información de ese subconjunto está siendo indirectamente introducida al modelo.

El esquema de subconjuntos de entrenamiento, validación y evaluación suele ser muy usado en ML y en particular en competencias de ML, donde los organizadores ponen a disposición de los participantes los subconjuntos ya particionados; de hecho, en algunos concursos, el conjunto de evaluación es liberado solamente cuando la competencia está por finalizar. Sin embargo, en problemas de ML menos estandarizados, donde los datos fueron recogidos de manera ad hoc, la partición en subconjuntos no está hecha, y queda a arbitrio del investigador decidir cómo llevarla a cabo. Es en este punto dónde surgen complicaciones: ¿cómo particionar de buena manera los datos? En especial si se considera que: 1) se puede producir sobreajuste en los conjuntos de entrenamiento y validación, y 2) siempre es ideal entrenar con la mayor cantidad de datos posible, por lo cual tampoco es conveniente tener un dataset de entrenamiento pequeño en desmedro de conjuntos de validación y evaluación grandes.

Una alternativa que surge para resolver este problema es la validación cruzada (o *cross-validation*), una estrategia que se basa en realizar múltiples instancias de un experimento y evaluarlas sobre distintos conjuntos de validación, para así promediar los resultados obtenidos y obtener un mejor estimador de E_{out}. Una técnica en particular de cross-validation es *K-Fold*, la cual consiste en particionar el conjunto de datos en K subconjuntos de igual o similar tamaño. Luego, se realizan K iteraciones de entrenamiento, utilizando como conjunto de validación una partición distinto en cada iteración, y empleando las $(K-1)$ particiones restantes como conjunto de entrenamiento. De esta forma, el error de K-Fold se calcula como $E = \frac{1}{K}\sum_{\mathrm{i}=1}^{K} E_{\mathrm{i}}$, donde E_{i} es el error calculado en el conjunto de validación i (figura 2.3.4).

Figura 2.3.4: Esquema de K-Fold, con $K = 10$. Fuente: Norena [66].

Otra técnica muy utilizada para reducir la posibilidad de *overfitting* es la de aumentar artificialmente el conjunto de datos (*data augmentation*). En el caso de procesamiento de imágenes, esto ocurre bajo la hipótesis de que el contenido de una imagen sigue siendo reconocible si es que ésta es rotada, movida o si se le modifica el brillo o contraste, y por ende, un clasificador debería ser capaz de reconocer esta nueva imagen modificada. De esta forma, es posible alimentar a la red con varias instancias distintas de la misma imagen, ligeramente modificadas y con la misma etiqueta, generándose así un dataset de mayor tamaño y reduciendo efectivamente la posibilidad de *sobreajuste*.

2.4. Redes Neuronales

Uno de los modelos clásicos dentro de ML, en especial en clasificación y aprendizaje supervisado, son las redes neuronales artificiales (*artificial neural networks*, ANN), un modelo computacional inspirado en una versión simplificada del funcionamiento del cerebro humano [71]. Una ANN consiste típicamente de varias capas jerárquicas de unidades o neuronas, donde cada capa procesa cierta información y la propaga hacia la siguiente; este proceso se repite sucesivamente hasta llegar a la capa final que produce una salida. Así, la primera capa de una red neuronal es conocida como capa de entrada (*input layer*), la última es la capa de salida (*output layer*) y las capas intermedias son llamadas capas ocultas (*hidden layers*).

La arquitectura mostrada en la figura 2.4.1 suele conocerse como red neuronal completamente conexa (*fully connected neural network*), ya que cada neurona de una capa está conectada a todas las neuronas de la capa siguiente. Para producir un resultado (o *activación*), cada neurona tiene asociado un vector de pesos; luego, estos pesos se ponderan con las activaciones de la capa anterior y son pasados por una función de activación. Así, si se fija la atención en la neurona j de la capa l, se tiene que el resultado que produce esta neurona es

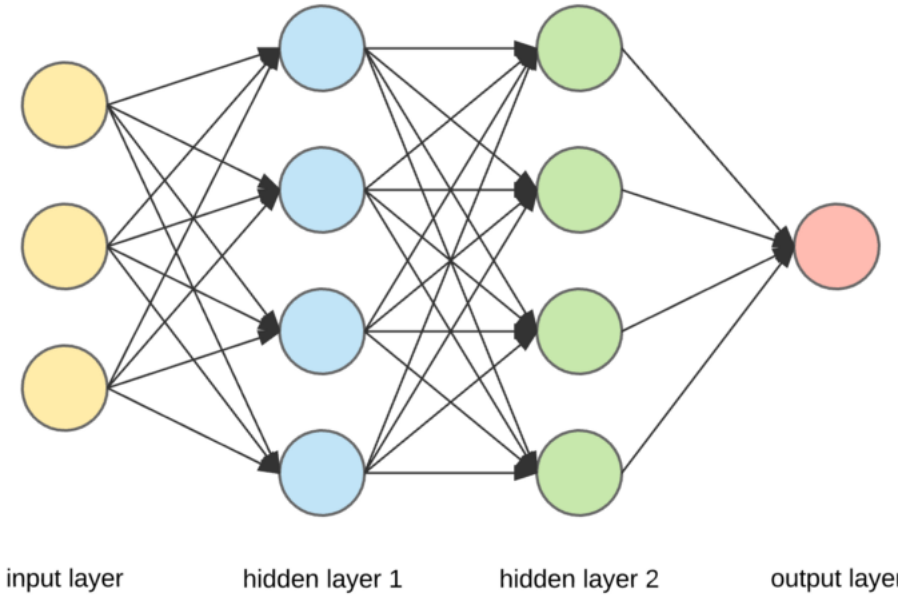


Figura 2.4.1: Arquitectura típica de una red neuronal artificial completamente conexa. Fuente: Dertat [25].

$$a_j^l = \sigma(\sum_k w_{jk}^l a_k^{l-1} + b_j^l), \tag{2.4.1}$$

donde a_j^l es la activación de la neurona j de la capa l, w_{jk}^l es el peso k de la neurona j de la capa l, a_k^{l-1} es la activación de la neurona k de la capa $l-1$, b_j^l es un valor de sesgo (o *bias*) de la neurona j de la capa l, y σ es la función de activación.

Para simplificar esta ecuación, se puede definir w^l como la matriz de pesos de la capa l, donde en cada fila j se encuentren los pesos de la neurona j de la capa l; así, la entrada en la fila j y columna k de la matriz es el peso w_{jk}^l de la ecuación anterior. Del mismo modo, se puede definir un vector de sesgos b^l, donde la entrada j es el bias de la neurona j. Además, el vector de activación de la capa l se puede escribir como a^l y σ es ahora una función vectorial, donde $\sigma(v)_j = \sigma(v_j)$; es decir, se aplica dicha función en cada entrada del vector. Luego, la activación de una capa de una red neuronal puede ser reescrita como

$$a^l = \sigma(w^l a^{l-1} + b^l) \tag{2.4.2}$$

Con estas definiciones, es fácil ver que son las matrices de pesos w^l y vectores de sesgo b^l los que definen la función aprendida por la red neuronal. Pero previo a explicar cómo es el proceso de aprendizaje y ajuste de pesos en una ANN, es necesario introducir una función de costo C, cuyo rol es cuantificar qué tan bueno o malo es el desempeño de la red sobre el conjunto de datos de entrenamiento. Si bien lo que intuitivamente se busca es que la red logre clasificar correctamente la mayoría de los datos, las matemáticas involucradas requieren tener una función objetivo que sea continua y diferenciable respecto a los parámetros w y b de la red. Así, una función básica que puede ser utilizada es el error cuadrático medio, definido como

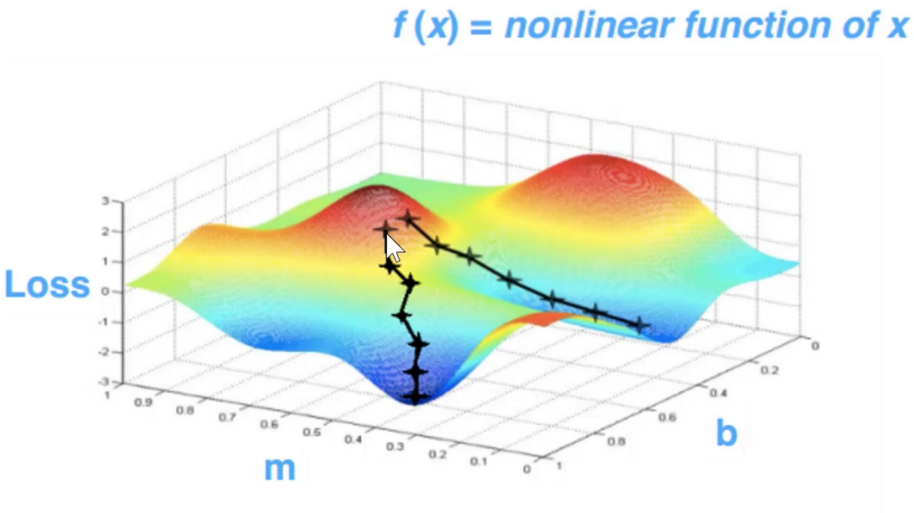


Figura 2.4.2: Esquematización de la heurística de descenso de gradiente. Es posible apreciar que dependiendo del punto de partida es posible alcanzar un mínimo local, mas no necesariamente el global. Fuente: Zhang [93].

$$C = \frac{1}{2N} \sum_x \|y(x) - a^L(x)\|^2, \tag{2.4.3}$$

donde $\|z\|$ es la norma del vector z, $y(x)$ es la clasificación correcta de x, L es el número de capas de la red, y $a^L(x)$ es la salida producida por la red cuando x es la entrada. Así, si es que la función clasifica bien la mayoría de los datos, C es cercano a 0, mientras que C es mayor si es que muchos datos son mal clasificados. Algo importante de recalcar es que la función C previamente definida es sólo un ejemplo de función de costo que puede ser utilizada. Así, en la práctica sólo se requiere de una función C que sea derivable respecto a w y b y que pueda ser calculada como un promedio de los errores individuales; es decir, $C = \frac{1}{N} \sum_x C_x$, donde C_x es el error producido con la entrada x.

Con estas definiciones, es posible establecer que el objetivo del proceso de entrenamiento de una ANN es minimizar la función de costo $C(w, b)$; así, el problema de aprendizaje se reduce a un problema de optimización. Dado que la función objetivo es compleja y no necesariamente tiene una solución analítica, se hace necesario utilizar una técnica iterativa conocida como *descenso del gradiente*. Si bien la matemática involucrada en esta técnica escapa el alcance del presente trabajo, es posible explicar el descenso del gradiente recordando que una función multivariable crea una superficie en el espacio; así, cada iteración del proceso consiste en dar un paso corto en la dirección de la pendiente generada por la superficie de la función hasta alcanzar un valle o mínimo local, tal como se ejemplifica en la figura 2.4.2. Es importante mencionar que este proceso no asegura que se alcance un mínimo global de la función, por lo cual esta técnica es considerada como una heurística. Por otra parte, dada la manera en que fue descrito el descenso de gradiente, cada iteración del proceso requiere calcular la salida de la red con cada valor del conjunto de entrenamiento, lo cual puede ser muy costoso en términos computacionales. Por ello, una variante muy utilizada es el *descenso estocástico del gradiente* (*Stochastic Gradient Descent*, SGD) el cual consiste en seleccionar un subconjunto aleatorio de datos en cada iteración y sobre ellos realizar el cálculo correspondiente [74].

Ahora bien, la pregunta que surge naturalmente es cómo calcular el gradiente de la función $C(w, b)$. Sin embargo, la respuesta a esta pregunta es compleja, ya que calcular eficientemente un gradiente que depende de miles de parámetros puede ser costoso en términos computacionales. En particular, es necesario calcular los valores $\partial C/\partial w^l_{jk}$ y $\partial C/\partial b^l_j$; es decir, las derivadas parciales de la función de costo respecto a cada uno de los pesos y sesgos de la red. Un algoritmo que soluciona este problema es el de *backpropagation*, el cual fue propuesto durante la década de 1970, pero no fue hasta 1986 que fue apreciado en el contexto de redes neuronales, gracias al trabajo de Rumelhart, Hinton y Williams [75]. Este algoritmo consiste en aplicar inteligentemente la regla de la cadena desde la última capa de la red hacia la primera, e ir propagando los resultados obtenidos hacia atrás; de ahí el nombre del algoritmo [64].

Con todos estos componentes, es posible definir el algoritmo de entrenamiento de una red neuronal artificial, tal como es posible apreciar en formato pseudocódigo en el algoritmo 2.4.1. Es importante notar que deliberadamente se ha dejado de forma ambigua cuál es la condición de detención del algoritmo; para esto existen distintas estrategias, tales como verificar convergencia hasta cierta tolerancia o entrenar por un número fijo de épocas (o *epochs*), cada una de las cuales requiere diferentes hiperparámetros.

Algoritmo 2.4.1 Entrenamiento de una red neuronal artificial, usando SGD.

```
1:  function EntrenamientoANN(𝒟)
2:      while !condición de detención do
3:          𝒟_k ← subconjunto aleatorio de 𝒟                          ▷ SGD
4:          for x ∈ 𝒟_k do
5:              Calcular a^L (salida de la red)                       ▷ Feedforward
6:              Calcular C_x(w, b)                                    ▷ El error con la entrada x
7:              Propagar error hacia atrás y calcular derivadas parciales   ▷ Backpropagation
8:          end for
9:          Actualizar pesos y sesgos de acuerdo a gradientes calculados
10:         Verificar condición de detención
11:     end while
12: end function
```

2.5. Aprendizaje profundo

El modelo de redes neuronales completamente conexas descrito en la sección anterior tiene muchas aplicaciones prácticas y es un pilar dentro de ML. Aún así, este modelo tiene varias falencias, algunas de las cuales son:

- Al intentar añadir más capas ocultas y "profundizar" la red, el gradiente de la función de costo se vuelve inestable, lo que generalmente se traduce en que las primeras capas de la red aprenden exponencialmente más lento que las capas finales; este fenómeno es conocido como *problema de desvanecimiento de gradiente* (*vanishing gradient problem*). Así mismo, en condiciones muy particulares, puede ocurrir el caso contrario: que las primeras capas de la red aprendan exponencialmente más rápido que las finales,

situación conocida como *problema de gradiente explosivo* (*explosive gradient problem*). Ambas situaciones implican obstáculos graves en el entrenamiento de una red neuronal.

- Dado que cada neurona de una capa l está conectada con todas las neuronas de la capa $l+1$, añadir nuevas capas intermedias implica agregar muchos nuevos parámetros (pesos y sesgos) que la red debe ajustar, aumentando considerablemente la complejidad computacional del algoritmo de aprendizaje.
- En caso de trabajar con imágenes, el hecho de que la entrada a la red sea un vector genera un problema, dado que una imagen suele representarse como una matriz. Así, dos opciones posibles para subsanar este problema son:
  - Aplanar la imagen y convertirla en un vector; de esta forma, por ejemplo, una imagen de 28x28 pixeles se transformaría en un vector de 784 valores. Aún así, esta solución implica perder la relación espacial entre los pixeles de la imagen y con ello, desechar información potencialmente relevante para la clasificación.
  - Calcular características (o *features*) de la imagen y usar dichos valores como entrada de la red. El problema con este enfoque es que decidir qué features utilizar es un proceso difícil, demoroso y que requiere conocimiento experto [63].

Algunos métodos que abordan estos problemas son aquellos que se engloban dentro del *aprendizaje profundo* (*Deep Learning*, DL). Si bien existen varias definiciones sobre qué es *Deep Learning*, en general dos conceptos claves son transversales a todas ellas: 1) se trata de modelos con varias capas de procesamiento no lineal, y 2) son métodos que permiten un aprendizaje de características de forma jerárquica, cada vez más abstracto en capas más profundas. Estos algoritmos de aprendizaje profundo se han masificado durante los últimos años debido, en parte, a las mejoras en hardware y capacidad de cómputo, la mayor disponibilidad de datos para entrenar los modelos y los avances en el campo de ML [22].

Uno de los modelos de DL más utilizados para el análisis de imágenes son las *redes neuronales convolucionales* (*convolutional neural networks*, CNN), las cuales están inspiradas vagamente en el funcionamiento de la corteza visual del cerebro humano. En esta arquitectura es útil pensar en los valores de entrada de la red ya no como un vector, sino como una matriz o superficie (en el caso de imágenes en escala de grises) o como un volumen (en el caso de imágenes en colores). Además, conceptualmente una CNN puede ser dividida en dos fases: una de extracción de características y otra de clasificación, tal como se aprecia en la figura 2.5.1. Así, si bien la etapa de clasificación es prácticamente idéntica al modelo de redes neuronales completamente conexas, es en la fase de extracción de *features* donde aparecen nuevos conceptos; a saber, el uso de campos receptivos locales, pesos compartidos entre neuronas y *pooling*. Todo esto presenta la ventaja de que una CNN puede ser alimentada por imágenes con poco o nulo preprocesamiento, ya que es la red, y no un humano, quien aprende y define cuáles son las características o *features* relevantes de las imágenes.

Campos receptivos locales: a diferencia de las redes neuronales completamente conexas, donde cada neurona está conectada a todas las neuronas de la capa previa, en la fase de extracción de features de una CNN cada neurona está conectada sólo a una pequeña parte de la capa anterior. El proceso consiste en definir una ventana (por ejemplo, de 5x5 neuronas), la cual será el campo receptivo de cada unidad de la siguiente capa. De esta forma, la entrada para cada neurona es sólo un subconjunto de datos de la capa previa, con lo cual cual la activación de dicha neurona dependerá solamente de los datos presentes en dicho subconjunto.

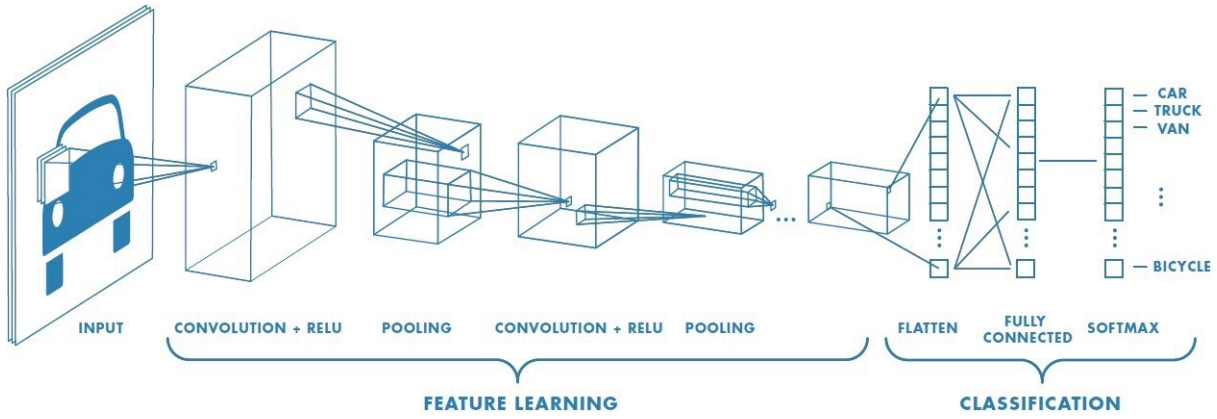


Figura 2.5.1: Arquitectura típica de una red neuronal convolucional. Fuente: Saha [79].

Luego, esta ventana se desliza hasta barrer todas las neuronas de la capa de entrada, tal como se ilustra en la figura 2.5.2.

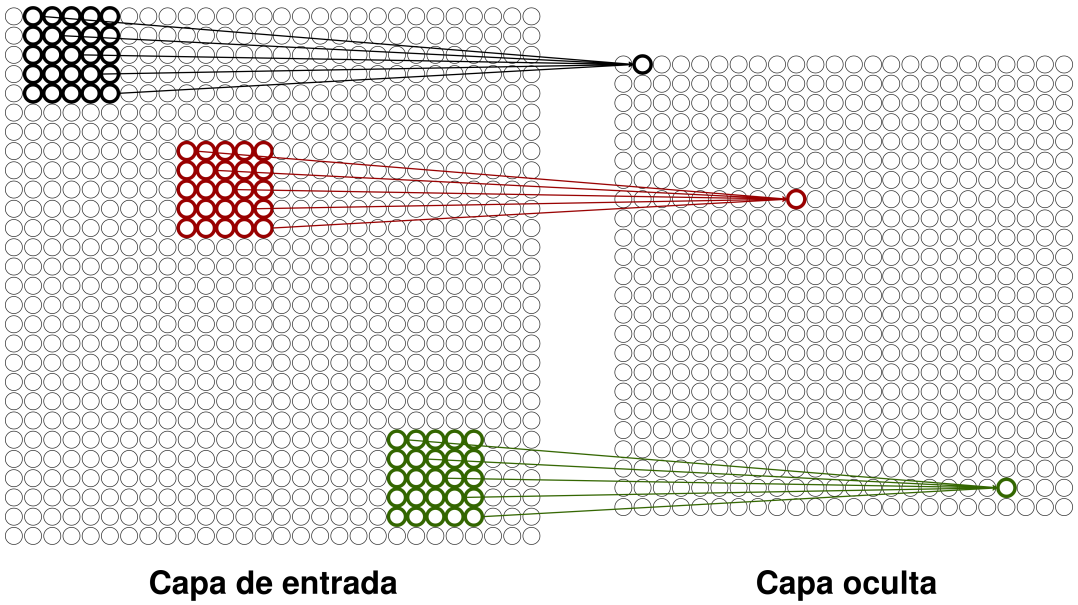


Figura 2.5.2: Campos receptivos locales en una red neuronal convolucional.

Pesos compartidos: otro concepto fundamental de las CNN es que todas las neuronas de una misma capa comparten los mismos pesos y sesgos. Así, todas las unidades de dicha capa detectan la misma *feature* o característica, sólo que en distintas partes de la imagen de entrada; por ello, se dice que las redes neuronales convolucionales están bien adaptadas a la invarianza traslacional de las imágenes. De esta forma, se dice que el conjunto formado por los pesos y sesgos compartidos de una capa forman un *kernel* o *filtro*; es más, la fórmula que aplica cada filtro sobre una capa de entrada es equivalente a la operación matemática conocida como *convolución*, lo cual da nombre a esta arquitectura de redes neuronales, y en particular, a este tipo de capas. Además, para aprender un mayor número de *features*, en cada nivel de la red se suele apilar varias capas, cada una de las cuales comparte pesos entre sus neuronas y aprende distintas características.

Pooling: posterior a una capa convolucional, se suele ubicar una capa de *pooling* (cuya

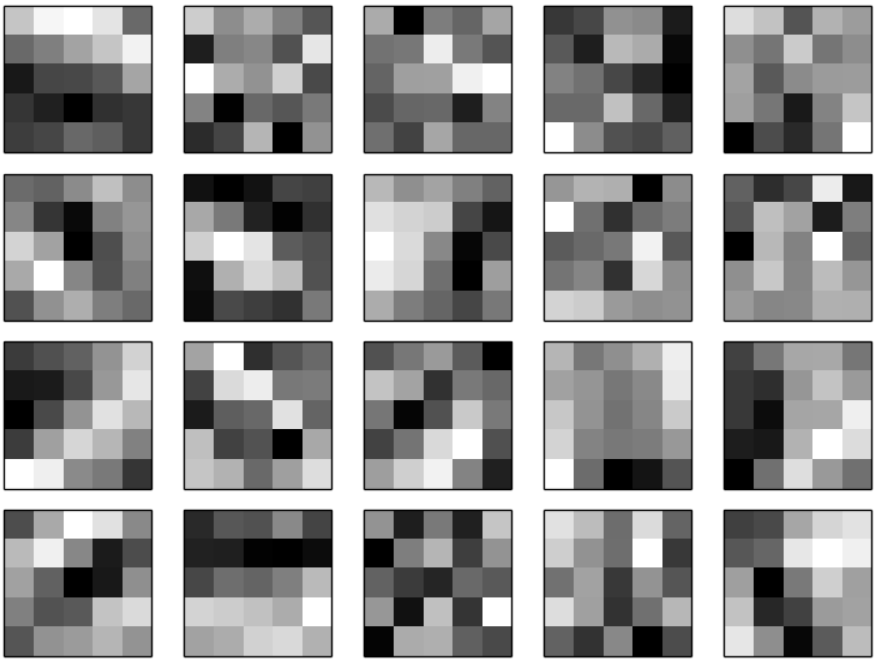


Figura 2.5.3: Ejemplos de distintos filtros de tamaño 5x5 aprendidos por una red neuronal convolucional. Bloques más oscuros representan un mayor peso, lo cual implica que dicho filtro responde con mayor fuerza a los pixeles correspondientes. Fuente: Nielsen [64].

traducción más cercana es "agrupamiento"), la cual cumple el rol de resumir la información proveniente de la capa convolucional. Así, por ejemplo, la información contenida en una ventana de 2x2 neuronas puede ser condensada en sólo una neurona, aplicando alguna función que sirva para sintetizar dicha información. Dos de las funciones más utilizadas son *max pooling*, la cual retorna el valor más alto contenido en la ventana, y *average pooling*, que calcula el promedio de dichos valores, las cuales se ejemplifican en la figura 2.5.4 [64] [79]. El uso de esta técnica de agrupamiento presenta dos ventajas: por un lado, se reduce la cantidad de parámetros que la red debe aprender en capas posteriores, y por otra parte, las capas de *pooling* funcionan como supresoras de ruido, descartando aquellos valores que no son relevantes para el aprendizaje.

Además, sobre cada capa convolucional suele añadirse una función de activación no lineal, tales como *ReLU* (*rectified linear unit*, $\sigma(x) = max(0, x)$) o la función sigmoide ($\sigma(x) = \frac{1}{1+e^{-x}}$). De esta forma, una red neuronal convolucional típica suele estar compuesta de varias capas convolucionales (con o sin función de activación) y capas de *pooling* en la etapa de extracción de características; posteriormente, los resultados de la última capa de esta fase son aplanados para formar un vector y se entra así a la etapa de clasificación, la cual es idéntica a una red neuronal completamente conexa.

Si bien los primeros prototipos de redes neuronales convolucionales se remontan a 1980, con el "*neocognitron*" de Fukushima [30], y ya en 1998, Yann LeCun mostró cómo entrenar una CNN utilizando *backpropagation* [54], no fue hasta 2012 que las CNN adquirieron popularidad, tras el trabajo de Alex Krizhevsky, quien utilizó redes neuronales convolucionales entrenadas en GPU para alcanzar un rendimiento de vanguardia en el desafío de clasificación

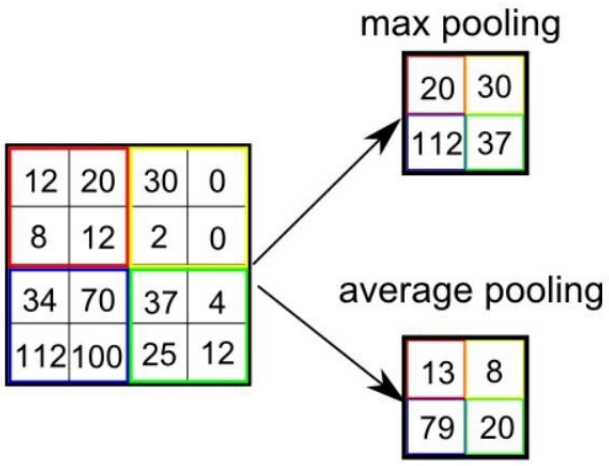


Figura 2.5.4: Ejemplos de aplicación de *max pooling* y *average pooling*, con una ventana de 2x2 neuronas. Fuente: Saha [79].

de imágenes de *ImageNet* [50]. Tras eso, las CNN han sido usadas extensivamente, y con excelentes resultados, en distintos contextos, como clasificación de imágenes naturales, análisis de imágenes médicas y reconocimiento facial [26].

2.5.1. Transferencia de aprendizaje

Un gran problema que aparece con las redes neuronales convolucionales es la gran cantidad de recursos necesarios para su entrenamiento; tanto en términos de hardware como de datos necesarios para que la red no se sobreajuste (el problema de *overfitting* previamente descrito). Como ejemplo, la CNN conocida como *AlexNet* fue entrenada con un subconjunto de la base de datos *ImageNet*, formado por aproximadamente 1,2 millones de imágenes de entrenamiento, 50.000 imágenes de validación y 150.000 imágenes de evaluación [50]. Además, dicha red neuronal está compuesta por más de 62 millones de parámetros entrenables [57], mientras que otras redes posteriores tales como *VGG-19* contienen más de 140 millones de parámetros [82]. Por ello, una alternativa popular es la de *transferencia de aprendizaje* (*transfer learning*), que busca mejorar el aprendizaje en una nueva tarea a través de la transferencia del conocimiento de una tarea aprendida previamente [85]. En el caso particular de imágenes y redes neuronales convolucionales, se ha observado que las primeras capas de la red tienden a detectar patrones generales (e.g. detectores de bordes), los cuales pueden servir en imágenes de dominios distintos a aquel en que fue originalmente entrenada la red [44]. Así, opciones que han surgido son:

- Utilizar la red convolucional como un extractor de características.
- Tomar una red pre-entrenada y reentrenar sólo la última capa (la de clasificación).
- Utilizar una red pre-entrenada como punto de partida para un nuevo entrenamiento, en vez de inicializar todos los pesos y sesgos en valores aleatorios. Esto puede significar reentrenar la red completa o sólo parte de la red (en especial las últimas capas). Esta técnica es conocida como *micro ajuste* (*fine tuning*), ya que las variaciones realizadas sobre la red tienden a ser pequeñas.

Por ello, se han creado repositorios para alojar redes neuronales ya entrenadas, tales como el *Model Zoo* de *Caffe* [16] o el *Hub* de *TensorFlow* [87], de tal manera de promover la transferencia de aprendizaje y facilitar la tarea de los investigadores.

2.6. Otros modelos de aprendizaje de máquinas

En esta sección, se describirán algunos modelos de ML que, si bien no son directamente empleados en este trabajo, sí son mencionados en la literatura de ML aplicado a clasificación de sobreexpresión de HER2, y que además serán usados como punto de comparación en la discusión de este proyecto.

2.6.1. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (*support vector machines*, SVM) son un conjunto de algoritmos de ML ampliamente utilizados en clasificación [19]. Para comprender su funcionamiento, es útil pensar primero en un caso de clasificación binaria, donde $x_{\mathrm{i}} = [x_{\mathrm{i}1}, x_{\mathrm{i}2}, ..., x_{\mathrm{i}P}]^T$ es un vector de *features* en el espacio $\mathbb{R}^P$, $y_{\mathrm{i}} \in \{-1, 1\}$ es su etiqueta asociada y $\mathcal{D}$ es un conjunto de datos de entrenamiento formado por pares $(x_{\mathrm{i}}, y_{\mathrm{i}})$. Luego, si es que las clases son separables linealmente, es posible calcular un *hiperplano* que divida el espacio en dos, quedando a un lado del hiperplano los datos que pertenecen a la clase 1 y al otro los de la clase -1. Así, por ejemplo, si es que $P = 2$, cada $x_{\mathrm{i}} = [x_{\mathrm{i}1}, x_{\mathrm{i}2}]$ es un vector con dos valores, y el hiperplano que se busca es una recta; del mismo modo, si es que $P = 3$, el hiperplano correspondiente será un plano. Dado que, cuando los datos efectivamente son linealmente separables, existen infinitos hiperplanos que permiten generar una clasificación, es que surge la idea de buscar un hiperplano óptimo. Para ello, es útil definir el concepto de *margen*: si se calcula la distancia desde cada punto hacia un hiperplano, el margen corresponderá a la mínima de esas distancias. Luego, se define como *hiperplano de margen máximo* a aquel que maximiza el margen con respecto a los datos de entrenamiento, tal como se ilustra en la figura 2.6.1.

Para hallar el hiperplano de margen máximo se debe resolver un problema de optimización cuadrática, el cual tiene solución analítica. Algo importante de mencionar es que el hiperplano resultante está determinado sólo por los datos que se encuentran sobre el margen, y no por aquellos más distantes del hiperplano. De esta forma, estos datos o vectores "soportan" el hiperplano, lo cual da nombre al modelo. Sobre esta idea de un clasificador de margen máximo se construyen los SVM, con dos conceptos adicionales:

- Dado que no siempre los datos son separables por la presencia de valores extremos (*outliers*), y también para reducir la variabilidad del modelo, se introduce el concepto de *margen blando*. Así, se permite que algunos valores crucen el margen o incluso sean clasificados de forma incorrecta. Este comportamiento es regulado por un hiperparámetro $C \geq 0$, de tal forma que mientras mayor sea C, se tendrá mayor tolerancia a errores.

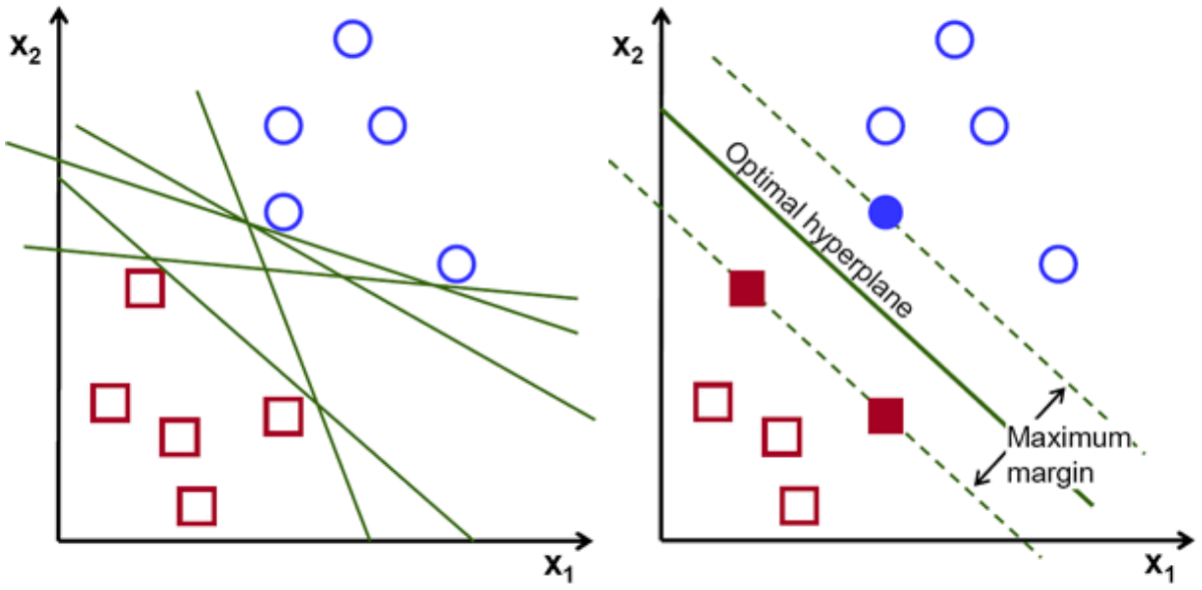


Figura 2.6.1: Izquierda: múltiples hiperplanos que dividen un conjunto de datos en base a sus clases. Derecha: hiperplano de margen máximo sobre el mismo conjunto de datos. Fuente: Gandhi [31].

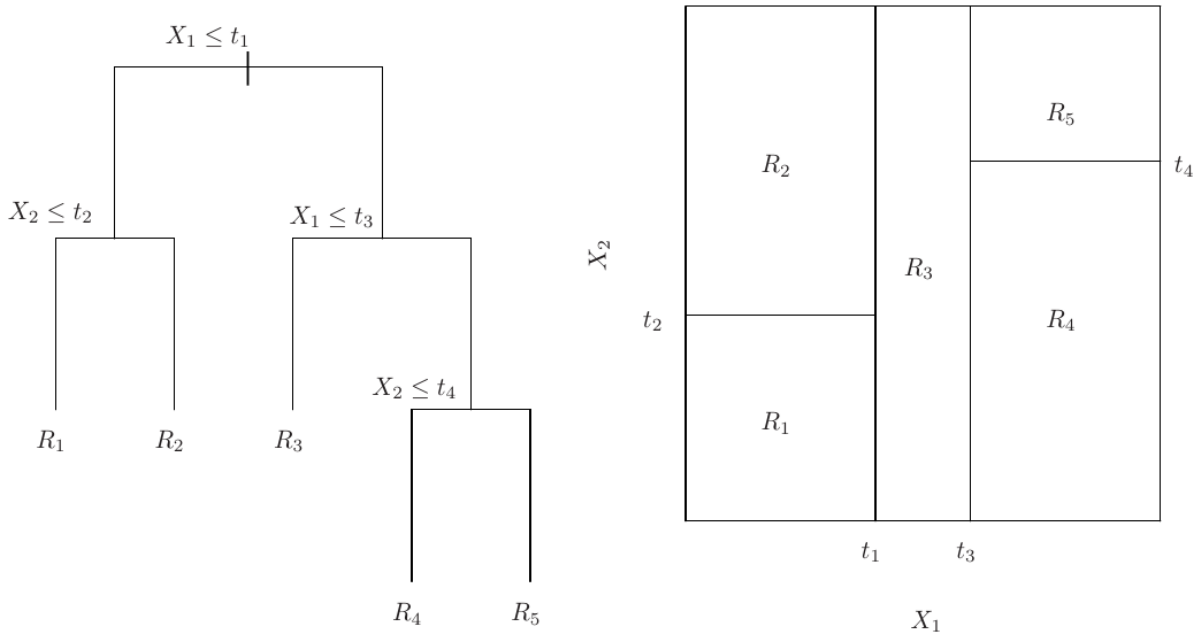
- Para problemas donde los datos no son lineales, se añade el concepto de *kernel*, el cual permite extender el espacio de características a dimensiones más altas y así acomodar una separación no lineal entre las clases.

Finalmente, para extender el modelo de SVM a clasificación no binaria, se han desarrollado estrategias consistentes en entrenar varios clasificadores binarios y luego agregar sus resultados [42].

2.6.2. Árboles de decisión y bosques aleatorios

Los árboles de decisión (*decision trees*) son modelos utilizados en aprendizaje supervisado, tanto para clasificación como regresión. En este modelo, el conjunto de datos es particionado recursivamente hasta alcanzar algún criterio de detención. Así, si es que cada x_{i} del conjunto de entrenamiento está formado por P *features* (es decir, $x_{\mathrm{i}} = [x_{\mathrm{i}1}, x_{\mathrm{i}2}, ..., x_{\mathrm{i}P}]^T$), en cada iteración del algoritmo se selecciona aquella característica que genera una "mejor" partición del subconjunto de datos. De esta forma, el espacio queda particionado en varias regiones, los cuales representan las *hojas* del árbol, mientras que cada división realizada por el algoritmo corresponde a un *nodo interno* de dicho árbol. La figura 2.6.2 ilustra un árbol construido en un espacio bidimensional (donde los datos contienen sólo dos *features*), junto a sus regiones correspondientes.

Una de las grandes ventajas de los árboles de decisión es su interpretabilidad. Sin embargo, de la forma en que han sido descritos hasta ahora, suelen obtener peores resultados que otros métodos de ML. Por ello, una alternativa que ha surgido para subsanar este problema es la de los bosques aleatorios (*random forests*), los cuales consisten en un ensamble o combinación de varios árboles de decisión decorrelacionados. Para lograr esto, se utilizan principalmente



(a) Árbol de decisión (b) Regiones generadas

Figura 2.6.2: Ejemplo de árbol de decisión en un espacio bidimensional, junto a las particiones generadas en dicho espacio. Fuente: James y col. [42].

dos técnicas:

- Para entrenar varios árboles de decisión desde un conjunto finito de N datos, se utiliza un método de remuestreo conocido como *bootstrap*, que consiste en generar nuevos conjuntos de datos, también de tamaño N, extrayendo con reemplazo datos desde el dataset original.
- Además, aunque los datos de entrenamiento contienen P características, cada árbol de decisión tiene acceso sólo a un subconjunto de $m < P$ features. Así, los distintos árboles deben generar particiones basándose en distintos criterios, lo cual en la práctica significa que haya una baja correlación entre ellos.

De esta forma, si bien se pierde interpretabilidad, los bosques aleatorios obtienen mejores resultados que un sólo árbol de decisión, además de reducir la posibilidad de sobreajuste y aumentar la capacidad de generalización del modelo [42].

2.7. Métricas de evaluación

En campos como estadística, ciencia de datos (*data science*) y aprendizaje de máquinas, muchas veces se busca cuantificar o medir un fenómeno, como puede ser, por ejemplo, la concordancia entre observadores o el desempeño de un modelo. Por ello, se han desarrollado diversas métricas que permiten evaluar numéricamente alguna tarea, facilitando la comparación entre distintas configuraciones o la selección de modelos. A continuación, se detallan las métricas de clasificación utilizadas en el presente trabajo:

- **Matriz de confusión:** no es una métrica por sí misma, pero sí una visualización gráfica del desempeño de un clasificador y que sirve como base para definir y explicar otras métricas. Es definida como una matriz C, tal que C_{ij} corresponde al número de elementos que pertenecen a la clase i y que fueron clasificados como pertenecientes a la clase j. De esta forma, en aquellas casillas donde $i = j$ (la diagonal de la matriz), se encuentra la cantidad de valores correctamente clasificados. La figura 2.7.1 esquematiza una matriz de confusión, junto a algunas fórmulas que se derivan de ésta.
- **Precisión:** responde la pregunta de qué proporción de los elementos clasificados como pertenecientes a la clase j realmente corresponden a dicha clase. Utilizando la matriz de confusión, se calcula como

$$\text{Precisión}_j = \frac{C_{jj}}{\sum_{i=1}^{n} C_{ij}}. \tag{2.7.1}$$

- **Recuperación (*recall*):** por otra parte, esta métrica sirve para cuantificar qué proporción de los elementos que realmente pertenecen a la clase i fueron clasificados como tal; es decir, fueron "recuperados" por el clasificador. Se calcula como

$$\text{Recuperación}_i = \frac{C_{ii}}{\sum_{j=1}^{n} C_{ij}}. \tag{2.7.2}$$

- **Exactitud (*accuracy*):** una métrica global de la performance de un clasificador. Corresponde al número de elementos correctamente clasificados (de todas las clases) dividido por el número total de elementos. Siguiendo la notación de la matriz de confusión, su fórmula es

$$\text{Exactitud} = \frac{\sum_{i=1}^{n} C_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}}. \tag{2.7.3}$$

- **F1-score:** corresponde a la media armónica entre la precisión y la recuperación para una clase dada. Se calcula como

$$\text{F1-score} = 2 \times \frac{\text{precisión} \times \text{recuperación}}{\text{precisión} + \text{recuperación}}. \tag{2.7.4}$$

Si bien las métricas de precisión, recuperación y F1-score están definidas para una clase en particular, esta definición puede ser extendida para generar una métrica global. Para ello, se puede utilizar un promedio *macro* (consistente en calcular la métrica para todas las clases y luego promediarlas) o un promedio *ponderado* (lo mismo, pero la métrica de cada clase tiene un peso asociado, correspondiente al número de elementos pertenecientes a dicha categoría). Dado que la exactitud no es una métrica adecuada cuando el conjunto de datos es desbalanceado, F1-score suele ser una métrica más apropiada para estos casos [40].

Se debe mencionar también que en el caso específico de clasificación binaria existen métricas y términos particulares. En esta configuración se suele tener una clase *positiva* y otra *negativa*; ejemplo típico de esto es el diagnóstico médico, donde se busca identificar si un paciente tiene o no cierta enfermedad (casos positivo y negativo, respectivamente). Luego, en base a dichas clases se definen las siguientes métricas:

- **Sensibilidad:** también llamada *tasa de verdaderos positivos* (*true positive rate*, TPR), mide la proporción de casos positivos correctamente identificados. En el caso de la

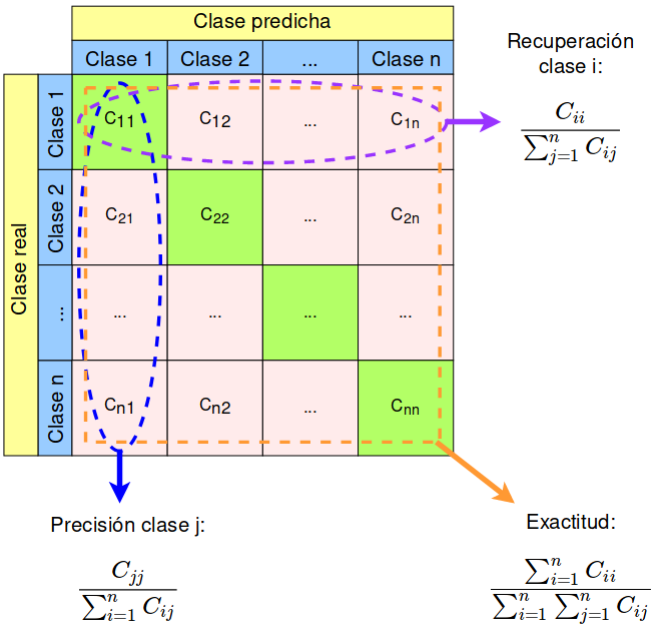


Figura 2.7.1: Esquema de matriz de confusión, junto a fórmulas derivadas.

analogía médica, la sensibilidad de un clasificador mediría la proporción de pacientes enfermos que sí fueron identificados como tales. Esta métrica se define como

$$\text{Sensibilidad} = \frac{TP}{TP + FN}, \tag{2.7.5}$$

donde TP = número de elementos correctamente identificados como positivos y FN = número de elementos incorrectamente clasificados como negativos.

- **Especificidad:** también llamada *tasa de verdaderos negativos*, (*true negative rate*, TPR), mide la proporción de casos negativos correctamente identificados. Continuando con el ejemplo clínico, esta métrica correspondería al porcentaje de personas sanas correctamente identificadas de esa manera. Se define como

$$\text{Especificidad} = \frac{TN}{TN + FP}, \tag{2.7.6}$$

donde TN = número de elementos correctamente identificados como negativos, y FP = número de elementos incorrectamente clasificados como positivos.

Además, muchos clasificadores binarios, en vez de producir directamente una etiqueta *positiva* o *negativa*, generan un valor continuo (e.g, una probabilidad $\in [0, 1]$). Por ello, se introduce un parámetro conocido como *umbral de decisión*, que cumple el rol de dividir el intervalo de valores de salida y mapear la salida desde un valor continuo a uno categórico (e.g., $x < 0{,}5$ es negativo, $x \geq 0{,}5$ es positivo). De esta forma, ajustando el umbral de decisión, es posible variar la sensibilidad y especificidad de un clasificador. Este nuevo parámetro introduce un nuevo grado de libertad en el sistema y con ello la pregunta respecto a cuál es el valor de umbral óptimo. Cuando los falsos positivos son igual de relevantes que los falsos negativos, se puede utilizar el **Índice J de Youden**, definido como

$$\text{J} = \max_{T}(\text{Sensibilidad}_T + \text{Especificidad}_T - 1), \tag{2.7.7}$$

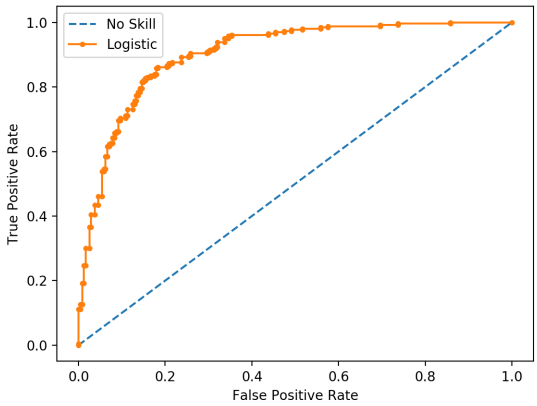


Figura 2.7.2: Ejemplo de curva ROC. Fuente: Brownlee [14].

y que retorna aquel umbral T que maximiza la sensibilidad y especificidad al mismo tiempo. Sin embargo, en múltiples contextos el costo de un falso positivo no es el mismo que el de un falso negativo; en el caso particular de la práctica clínica, identificar erróneamente a un paciente sano como enfermo (falso positivo) es significativamente menos grave que diagnosticar a un paciente enfermo como sano (falso negativo). En dichas situaciones, queda a arbitrio de los desarrolladores del sistema decidir qué umbral utilizar en base a las características propias del problema.

En casos así, se suelen recurrir a las **curvas ROC** (acrónimo de *receiver operating characteristic* o *característica operativa del receptor*), una representación gráfica de cómo varía la tasa de verdaderos positivos (*sensibilidad*) versus la tasa de falsos positivos ($1 -$ *especificidad*) de un clasificador al modificar el umbral de decisión (figura 2.7.2). Uno de los principales beneficios de esta herramienta es que permite visualizar qué tanto mejora la sensibilidad del clasificador al sacrificar un poco la especificidad de éste, facilitando así la decisión que deben tomar quienes implementan el sistema de clasificación. Además, también es posible calcular el área bajo la curva ROC, dando así pie a la métrica **AUC** *(area under the curve)*. Dado que el área bajo esta curva es parte de un cuadrado unitario, AUC toma valores en el rango $[0, 1]$, teniendo mejores resultados un clasificador que tiene un área más cercana a 1. Junto a esto, la métrica AUC cumple con la propiedad de ser equivalente a la probabilidad de que el clasificador produzca una salida mayor para una instancia aleatoria positiva frente a una instancia aleatoria negativa [27].

Existen varios enfoques para extender el análisis de curvas ROC y la métrica AUC al caso de clasificación multiclase (número de clases $= K > 2$). Algunos de estos enfoques se basan en calcular un volumen bajo una curva (VUC) en un espacio con K dimensiones, mientras que otros calculan un promedio ponderado de AUC entre las distintas clases. Ejemplos de esto último son la métrica M de Hand y Till [36], que calcula un promedio de AUC entre todos los pares de clases posibles (estrategia uno-contra-uno), y el trabajo de Provost y Domingos [69], que utiliza una estrategia uno-contra-todos. No obstante, todos estos enfoques sufren de problemas de interpretabilidad y/o costo computacional, además de perder ciertas propiedades deseables que sí cumple AUC en el caso binario. A esto se suma la dificultad para establecer el umbral de decisión; mientras en el caso binario basta con calcular sólo

un valor que divide el intervalo, en la configuración multiclase esta tarea se transforma en encontrar una partición de un $(K-1)$-símplex [46]. Por todos estos motivos, en el presente trabajo se decidió no hacer uso de curvas ROC ni AUC para los clasificadores multiclase y sólo utilizarlos en clasificación binaria.

Ahora bien, también existen métricas o índices para evaluar y comparar la concordancia entre observadores. En el presente trabajo, un grupo de patólogos evaluó un conjunto de biopsias, y algunas de estas métricas fueron utilizadas para analizar la concordancia entre estos especialistas (capítulo 3). Las métricas empleadas fueron:

- **Porcentaje de concordancia:** parecida a la exactitud, pero en un contexto distinto. Para el caso de dos observadores y clases categóricas, se calcula como

$$\% \text{ de concordancia} = 100 \times \frac{n_{\mathrm{i}}}{n_{\mathrm{i}} + n_{\mathrm{d}}}, \tag{2.7.8}$$

  donde n_{i} es el número de muestras que fueron clasificadas de igual manera por ambos observadores, y n_{d} es el resto de las muestras (es decir, aquellas donde los observadores dieron una clasificación distinta). Además, $n_{\mathrm{i}} + n_{\mathrm{d}} = N =$ número total de muestras.
- κ **de Cohen:** Es un índice utilizado para medir la concordancia entre un par de observadores que evalúan un grupo de muestras en clases categóricas. En teoría, este índice toma en cuenta la probabilidad de que ambos observadores hayan entregado la misma evaluación simplemente por azar. La fórmula es

$$\kappa = \frac{p_o - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}}, \tag{2.7.9}$$

  donde p_o es el acuerdo relativo entre ambos patólogos (es decir, el porcentaje de concordancia dividido por 100) y p_{e} es la probabilidad hipotética de concordancia por azar; este valor se calcula como

$$p_{\mathrm{e}} = \frac{1}{N^2} \sum_{k=1}^{K} n_{1k} n_{2k}, \tag{2.7.10}$$

  donde N es el número total de muestras, K es el número de clases posibles y $n_{\mathrm{i}k}$ es el número de elementos clasificados en la categoría k por el observador i.

  Una de las críticas al κ de Cohen es su dificultad de interpretación. En Landis y Koch [52] se recomienda utilizar la tabla 2.7.1 para interpretar el nivel de acuerdo representado por distintos valores de κ, aunque asumiendo que esta división es arbitraria.
- α **de Krippendorff:** es una medida estadística para evaluar el acuerdo entre observadores. Tiene la ventaja de ser una generalización de varios otros índices, pudiendo ser aplicado a cualquier número de observadores, (no sólo dos, como en el κ de Cohen), distintos tipos de datos (categóricos, ordinales, intervalos, etc) y conjuntos de datos con valores faltantes (donde alguno de los observadores no etiquetó todos los datos) [49]. Para este trabajo en particular, resulta interesante que este índice sea aplicable a clases ordinales, como lo son las relativas a clasificación de sobreexpresión HER2. Esto, porque si bien son cuatro clases distintas, existe un orden entre estas categorías; no es lo mismo confundir las clases 0 y 1+ que confundir las clases 0 y 3+. La fórmula general del α de Krippendorff es:

$$\alpha = 1 - \frac{D_o}{D_{\mathrm{e}}}, \tag{2.7.11}$$

Valor de κ	Nivel de concordancia
<0.00	Pobre (*Poor*)
0.00 - 0.20	Leve (*Slight*)
0.21 - 0.40	Aceptable (*Fair*)
0.41 - 0.60	Moderado (*Moderate*)
0.61 - 0.80	Considerable (*Substantial*)
0.81 - 1.00	Casi perfecto (*Almost perfect*)

Tabla 2.7.1: Interpretación de distintos valores de κ de Cohen. Fuente: Landis y Koch [52].

donde D_o es la discordancia observada y D_e es la discordancia esperada por azar.

2.8. Estado del arte de clasificación de sobreexpresión de HER2 mediante ML

Durante los últimos años, el uso de técnicas de ML dentro de la investigación de cáncer ha ido en aumento. Según una revisión realizada el 2014, entre 2010 y 2014 se habían publicado más de 500 artículos que ligaban el cáncer con el aprendizaje de máquinas [48]. De acuerdo a la misma publicación, muchos de estos estudios iban en la línea de generar modelos matemáticos para (i) predecir la susceptibilidad de contraer cáncer, (ii) la recurrencia del cáncer en pacientes que ya habían contraído la enfermedad previamente, y (iii) predecir la supervivencia de pacientes cuyo cáncer ya ha sido diagnosticado.

Parte importante de la investigación actual sobre cáncer trata sobre el análisis y desarrollo de biomarcadores con potencial valor clínico, ya sea para predicción de riesgo, detección temprana, clasificación del cáncer según su etapa y/o grado y selección de terapia personalizada para el paciente, entre otros usos [56]. Uno de dichos biomarcadores es la ya mencionada proteína HER2, la cual se presenta en cáncer de mama, pulmón y estómago. Respecto al análisis computacional de HER2 en biopsias digitalizadas, la literatura muestra dos grandes enfoques, a saber:

- **Correlación de variables *externas* con la clasificación HER2:** este enfoque busca aprovechar las ventajas de las técnicas de análisis de imágenes y aprendizaje de máquinas para encontrar variables que se correlacionen con los valores de HER2 0, 1+, 2+ o 3+, independientemente de si dichas variables son consideradas o no en las guías clínicas correspondientes. Un ejemplo de esto es el software comercial HER2-CONNECT [15], que cuantifica la conectividad de las membranas de las células teñidas inmunohistoquímicamente en muestras de cáncer de mama. Este valor de conectividad, continuo en el intervalo $[0, 1]$ y que no es considerado por la guía clínica correspondiente, puede ser correlacionado con los valores discretos 0/1+, 2+ y 3+, dependiendo del rango donde se encuentra. Los autores del artículo reportan una concordancia de 92.3 % (κ de Cohen = 0.86) entre el algoritmo y lo evaluado por patólogos sobre regiones de interés extraídas de biopsias, utilizando un esquema de tres clases (0/1+, 2+ y 3+).

No obstante, gran parte de la efectividad del algoritmo depende de los valores de corte (*cut-off values*) empleados para generar los rangos de clasificación, además de un valor de sensibilidad que es requerido al usuario al momento de utilizar el software. Si bien en el artículo original se presentan los valores que los autores utilizaron, estos parámetros son dependientes de las muestras y cómo éstas fueron obtenidas. Por ello, los valores de corte y la sensibilidad deben ser ajustados por cada laboratorio de acuerdo a sus propias características (escáneres, componentes químicos utilizados para teñir las muestras, etc), siendo este un proceso complejo que ha sido incluso objeto de estudio en tesis de postgrado [55].

- **Replicación de proceso de clasificación de especialistas:** este enfoque busca simular computacionalmente los algoritmos utilizados por patólogos y otros especialistas a la hora de realizar la clasificación HER2, siguiendo las recomendaciones de las guías clínicas correspondientes. Algunos ejemplos de este enfoque son:
  - En Vandenberghe y col. [89] utilizan redes neuronales convolucionales para la clasificación automatizada de sobreexpresión de HER2 en biopsias de cáncer de mama. Específicamente, emplean técnicas clásicas de procesamiento de imágenes para detectar la presencia de células en cada biopsia, para posteriormente ingresar cada imagen de célula a una CNN y clasificarla de acuerdo a su nivel de HER2. Finalmente, ya habiendo cuantificado las células pertenecientes a cada nivel, la biopsia completa se clasifica utilizando las recomendaciones de CAP/ASCO [90], alcanzando una concordancia de 83 % (κ de Cohen = 0.69) respecto a lo evaluado por patólogos.
  - En Pitkäaho y col. [68] también utilizan redes neuronales convolucionales, pero en vez de trabajar con células, clasifican imágenes de 128x128, que pueden contener una o más células. En este trabajo se alcanza un 97,7 % de exactitud en la clasificación de imágenes, y los autores plantean que es posible utilizar dicho clasificador para evaluar una biopsia completa siguiendo el algoritmo recomendado en la guía clínica. No obstante esto, no reportan resultados para la clasificación global de biopsias.
  - En Saha y Chakraborty [78] proponen una arquitectura de CNN llamada *Her2Net* para segmentar núcleos y membranas celulares y clasificar la sobreexpresión de HER2. Tal como señalan los autores de este último artículo, su método se diferencia de lo propuesto por Vandenberghe y col. [89] en que: i) *Her2Net* es una red más profunda, lo que llevaría a obtener mejores resultados, y ii) esta nueva arquitectura sí toma en cuenta la tinción en las membranas de las células, tal como se señala en las guías clínicas correspondientes. De la literatura revisada, este trabajo es el que presenta mayor concordancia con la evaluación llevada a cabo por patólogos (0: 100 %, 1+: 98.13 %, 2+: 97.98 %, 3+: 99.97 %).

  Los métodos descritos, tal como están planteados, no son necesariamente resistentes a la variabilidad de configuraciones que se da entre laboratorios. Además, presentan la desventaja de que las CNNs no proveen una forma clara de interpretar ni explicar las reglas utilizadas durante la clasificación, lo cual puede provocar que un médico especialista desconfíe de los resultados entregados por estas *cajas negras* [47]. No obstante, gracias a que las técnicas de *machine learning* sólo se utilizan para la clasificación de células y/o parches, y en todo el resto del proceso se siguen las recomendaciones de las guías clínicas, se estima que este problema queda parcialmente solucionado.

Por otro lado, Pezoa y col. [67] estudia el problema de segmentación de células con sobreexpresión HER2 en cáncer de mama, el cual corresponde a un paso previo a la clasificación del tejido canceroso en las clases 0, 1+, 2+ y 3+. El método que allí se propone busca generar una imagen binaria segmentada, donde cada píxel corresponderá a la clase sobreexpresado o no sobreexpresado. Para lograr dicha tarea, un especialista debe etiquetar primero una pequeña parte de los píxeles de la imagen, los cuales son utilizados para entrenar una máquina de soporte vectorial (SVM). Finalizado el entrenamiento, se utiliza dicho SVM para clasificar el resto de los píxeles de la imagen y producir así la imagen binaria segmentada. Este método genera un nuevo clasificador para cada biopsia, lo cual permite que sea resistente a problemas de estandarización interlaboratorios, además de aprovechar el conocimiento de cada patólogo; sin embargo, requiere de mayor trabajo humano en el proceso y aún depende en parte de la subjetividad del especialista.

Si bien los métodos previamente detallados abordan la sobreexpresión de la proteína HER2, todos estos estudios tratan el problema específicamente en cáncer de mama. Esto tiene sentido al recordar que la relevancia de este biomarcador en este tipo de cáncer fue observada varios años antes que en cáncer gástrico. No obstante, como ya se mencionó previamente, existen varias diferencias en cómo se manifiesta la sobreexpresión de HER2 entre ambos tipos de cáncer. Por ello, los métodos descritos para cáncer de mama no necesariamente funcionarán *out-of-the-box* con muestras de cáncer gástrico, lo cual refuerza la necesidad de realizar investigación en este tópico.

En dicha línea de investigación, sobre clasificación de HER2 en cáncer gástrico, resaltan tres estudios recientes:

- En Nielsen, Nielsen y Vyberg [65] se recalculan los valores de corte del software HER2-CONNECT, para así aplicar dicho algoritmo en cáncer gástrico. Si bien se obtienen buenos resultados (sensibilidad y especificidad de 100 % con respecto a FISH), este enfoque sigue teniendo la desventaja de que los parámetros óptimos deben ser ajustados por cada laboratorio, en un proceso que dista de ser trivial.
- Sharma y col. [81] utiliza CNNs para clasificar imágenes de una biopsia en las clases HER2+, HER- y no-tumor, obteniendo una exactitud igual a 0.69. Sin embargo, las muestras que ellos utilizan no están teñidas inmunohistoquímicamente, sino que emplean una tinción de hematoxilina-eosina, lo cual los aleja de los otros trabajos revisados.
- En Zakrzewski y col. [92], en vez de clasificar la sobreexpresión de la proteína HER2, utilizan algoritmos de DL para clasificar la amplificación del gen HER2. Para ello, utilizan imágenes digitalizadas del examen FISH. Además, pese a que los autores mencionan que sus modelos serían aplicables tanto a cáncer gástrico como a cáncer de mama, las muestras de FISH utilizadas para el entrenamiento corresponden sólo a cáncer de mama.

A nuestro conocimiento, no existe aún literatura que trate el problema de clasificación de sobreexpresión de HER2 en cáncer gástrico utilizando un enfoque computacional que replique el proceso de diagnóstico llevado a cabo por especialistas. Por ello, y debido a la preponderancia de este tipo de cáncer en Chile, se plantea que es necesario llevar a cabo investigación que aborde este problema.

Capítulo 3

Bases de datos

En este capítulo se presentan los datos originales (o *en bruto*) utilizados para llevar a cabo los experimentos. En particular, es descrito el estudio *PRECISO*, de donde provienen las biopsias con tinción inmunohistoquímica usadas, además de dos etiquetados distintos, consistentes de clasificaciones de sobreexpresión HER2 en biopsias y anotaciones de regiones de interés.

3.1. Estudio PRECISO

Las imágenes utilizadas en el presente proyecto provienen del estudio *Prospective Observational Study of Patients With Locally Advanced Gastric Cancer Treated With Perioperative Chemotherapy and Surgery, (PRECISO)* [61]. Dicho estudio tiene como fin "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, Chile". El cohorte del estudio estuvo compuesto por pacientes con carcinomas gástricos de tipo T3-T4 y/o N+ M0, de acuerdo a la clasificación TNM. Además, el cáncer debía ser resecable y localmente avanzado, y los pacientes tratados con quimioterapia perioperatoria. Bajo estas condiciones, 61 pacientes fueron incluidos en el estudio.

De estas 61 personas enroladas en el estudio, 48 autorizaron que les fuera determinada la sobreexpresión de HER2; la información del resultado HER2 y las muestras digitalizadas de 40 de estos pacientes fueron utilizadas en el presente trabajo. Es importante mencionar que los datos utilizados consisten exclusivamente en biopsias digitalizadas y anonimizadas y de su correspondiente evaluación HER2, sin hacer uso de información adicional de cada persona. Además, se debe notar que varios pacientes tienen dos biopsias asociadas: una endoscópica, realizada antes de la extirpación del tumor, y una por resección.

Así, los datos utilizados consisten en 61 pares de biopsias provenientes de 40 pacientes, donde cada par está formado por una biopsia con tinción H&E (utilizada para detectar la presencia de cáncer) y otra con tinción IHC (utilizada para detectar sobreexpresión de HER2), como se puede apreciar en la figura 3.1.1. Se incluyen tanto biopsias endoscópicas

como biopsias por resección. Todas estas biopsias fueron digitalizadas utilizando el escáner de placas histológicas Nanozoomer XR (Hamamatsu, Japón) y almacenadas en el formato propietario `.ndpi`, con una magnificación máxima de 40x. Además, las biopsias IHC incluyen un tejido de control, utilizado para comparar la reactividad HER2 frente a una muestra conocida.

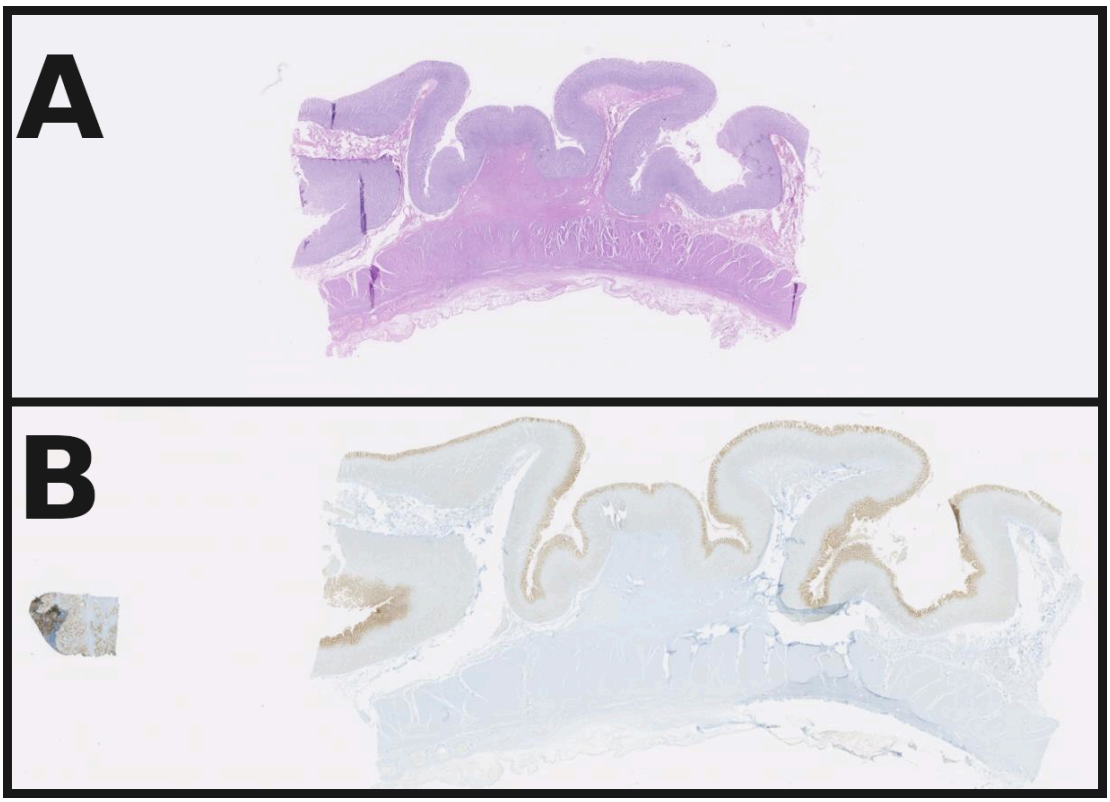


Figura 3.1.1: Ejemplo de biopsias del estudio *PRECISO*. A) Tinción H&E, magnificación 0.21x B) Tinción IHC, magnificación 0.25x. Es posible apreciar que en la biopsia IHC, en el lado izquierdo se encuentra el tejido de control.

Junto a las imágenes, también fueron provistas las evaluaciones realizadas originalmente por la patóloga asociada al estudio *PRECISO*, de ahora en adelante llamada *patólogo 0*. La tabla 3.1.1 detalla las estadísticas de la evaluación realizada.

De acuerdo al algoritmo de evaluación de la sobreexpresión de proteína HER2 en cáncer gástrico (tabla 2.1.5), un patólogo debe estimar la cantidad de células que cumplen cierto patrón y en base a eso clasificar la biopsia completa. De esta forma, que una biopsia sea clasificada como 0, 1+, 2+ o 3+, no implica que todas las células del tejido sean de dicha clase; en particular, en el caso de las biopsias por resección, basta con que sólo un 10 % de las células tengan algún patrón para clasificar así toda la biopsia. Es más, no todas las zonas del tejido de una biopsia son cancerosas, y sólo sobre el tejido canceroso debe evaluarse la sobreexpresión HER2. Por ello, para construir una base de datos de imágenes apta para ser utilizada en un algoritmo de ML, es necesario contar con anotaciones de regiones de interés (*regions of interest, ROIs*), donde cada zona sea relativamente homogénea y contenga preferentemente células de un solo tipo. No obstante, los datos originales del estudio *PRECISO* no contaban con anotaciones de este tipo.

	0	1+	2+	3+	Total
Biopsias endoscópicas	17	4	2	3	26
Biopsias por resección	28	4	0	3	35
Total	45	8	2	6	61

Tabla 3.1.1: Resumen de la evaluación realizada por patólogo 0, original del estudio PRECISO, desagregada por tipo de biopsia.

3.2. Etiquetado previo

Para estudios previos, le fue solicitado a dos patólogos que realizaran anotaciones de diversas regiones de las biopsias con tinción IHC, etiquetando cada ROI con su clasificación HER2 correspondiente (0, 1+, 2+, 3+). Además, se le solicitó a los patólogos que evaluaran la sobreexpresión HER2 de cada biopsia completa, de acuerdo al algoritmo recomendando por la guía clínica de evaluación de HER2 en cáncer gástrico [10].

No obstante, dichos patólogos no anotaron la totalidad de las biopsias. De esta forma, uno de los patólogos (de ahora en adelante, patólogo 1) etiquetó 12 biopsias, mientras que el otro (patólogo 2) etiquetó 38 biopsias. Estas anotaciones fueron realizadas en la plataforma NDP.Serve (Hamamatsu, Japón), generando cada uno de ellos 3 ROIs en cada biopsia en la que trabajaron, además de la clasificación total de la placa. Todas los ROIs anotados fueron de tipo circular, y se le asignaron colores distintos a las anotaciones de cada patólogo.

	Patólogo 1				
	0	1+	2+	3+	Total
Biopsias endoscópicas	0	3	2	2	7
Biopsias por resección	1	1	2	1	5
Total	1	4	4	3	12
	Patólogo 2				
	0	1+	2+	3+	Total
Biopsias endoscópicas	2	3	7	2	14
Biopsias por resección	2	2	15	5	24
Total	4	5	22	7	38

Tabla 3.2.1: Resumen de las evaluaciones de biopsias realizadas por patólogos 1 y 2, desagregadas por tipo de biopsia.

La tabla 3.2.1 detalla las clasificaciones de biopsias que los patólogos 1 y 2 llevaron a cabo, mientras la tabla 3.2.2 muestra las estadísticas de los ROIs anotados por ambos patólogos. En total, se obtuvieron 150 ROIs, de tamaños relativamente uniformes y provenientes de 38 placas distintas. Por otra parte, la tabla 3.2.3 muestra la concordancia entre patólogos relativa a la clasificación HER2 de biopsias. De esta forma, para cada par de patólogos se calculó el porcentaje de concordancia y el coeficiente κ de Cohen. Además, para el grupo de tres

	Número de anotaciones				
	0	1+	2+	3+	Total
Patólogo 1	3	13	10	10	36
Patólogo 2	12	18	57	27	114
Total	15	31	67	37	150
	Radio promedio (Desviación estándar) (µm)				
	0	1+	2+	3+	Total
Patólogo 1	564.2 (0)	564.3 (0.1)	564.3 (0.1)	564.3 (0.1)	564.3 (0.1)
Patólogo 2	567.1 (4.7)	434.5 (189.6)	504.4 (142.2)	480.5 (164.0)	494.3 (151.2)
Total	566.5 (4.3)	488.9 (156.9)	513.3 (132.8)	503.1 (144.4)	511.1 (135.0)
	Radio mínimo (µm) / Radio máximo (µm)				
	0	1+	2+	3+	Total
Patólogo 1	564.2 / 564.2	564.2 / 564.4	564.2 / 564.4	564.2 / 564.4	564.2 / 564.4
Patólogo 2	564.3 / 574.8	167.6 / 565.4	178.5 / 576.2	179.1 / 573.3	167.6 / 576.2
Total	564.2 / 574.8	167.6 / 565.4	178.5 / 576.2	179.1 / 573.3	167.6 / 576.2

Tabla 3.2.2: Estadísticas de anotaciones realizadas por patólogos 1 y 2.

patólogos se calculó la concordancia utilizando el coeficiente α de Krippendorff, utilizando una métrica ordinal. Así, se puede observar que los patólogos 1 y 2 suelen concordar en su evaluación, pero difieren mayoritariamente de lo observado por el patólogo 0, el experto asociado al estudio *PRECISO*.

Debido a esta alta discrepancia entre los tres patólogos, se debió aplicar un proceso de filtrado de datos. Así, se utilizó una estrategia de *voto de mayoría*, trabajando sólo con aquellas biopsias donde al menos dos de los tres patólogos hayan entregado la misma evaluación de sobreexpresión de HER2. De esta manera, de las 38 biopsias que fueron evaluadas por al menos dos patólogos, sólo 17 cumplieron el criterio ya mencionado. Además, dado que el patólogo 1 sólo evaluó 12 biopsias, y que sólo 3 de ellas entraron en el conjunto filtrado, es que se decidió utilizar sólo los ROIs anotados por el patólogo 2. La tabla 3.2.4 detalla las biopsias y ROIs utilizados tras este proceso de filtrado.

	Endoscopias		
	N°muestras	% concordancia	κ de Cohen
Patólogos 0-1	7	29 %	0.22
Patólogos 0-2	14	36 %	0.22
Patólogos 1-2	7	86 %	0.86
α de Krippendorff (todos los patólogos)	0.39		
	Resecciones		
	N°muestras	% concordancia	κ de Cohen
Patólogos 0-1	5	20 %	0.0
Patólogos 0-2	24	21 %	0.12
Patólogos 1-2	5	80 %	0.71
α de Krippendorff (todos los patólogos)	0.08		
	Total		
	N°muestras	% concordancia	κ de Cohen
Patólogos 0-1	12	25 %	0.16
Patólogos 0-2	38	26 %	0.15
Patólogos 1-2	12	83 %	0.77
α de Krippendorff (todos los patólogos)	0.19		

Tabla 3.2.3: Concordancia entre patólogos 0, 1 y 2, desglosada por tipo de biopsia. α de Krippendorff fue calculado utilizando las clasificaciones realizadas por los tres patólogos.

Clasificación HER2	Número de biospias			Número de ROIs		
	Endoscopias	Resecciones	Total	Endoscopias	Resecciones	Total
0	2	2	4	5	6	11
1+	2	0	2	7	0	7
2+	3	2	5	6	6	12
3+	2	4	6	9	12	21
Total	9	8	17	27	24	51

Tabla 3.2.4: Estadísticas de clasificaciones de biopsias y ROIs, tras aplicar proceso de filtrado basado en voto de mayoría y eliminar ROIs de patólogo 1.

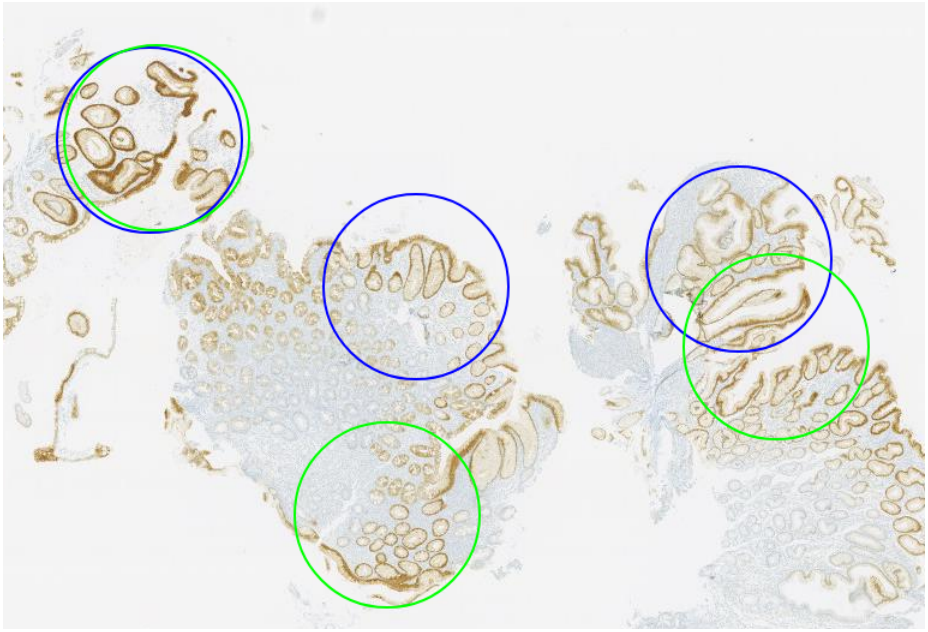


Figura 3.2.1: Ejemplo de anotaciones realizadas por patólogos 1 (azul) y 2 (verde) sobre biospia endoscópica. Magnificación 1.25x.

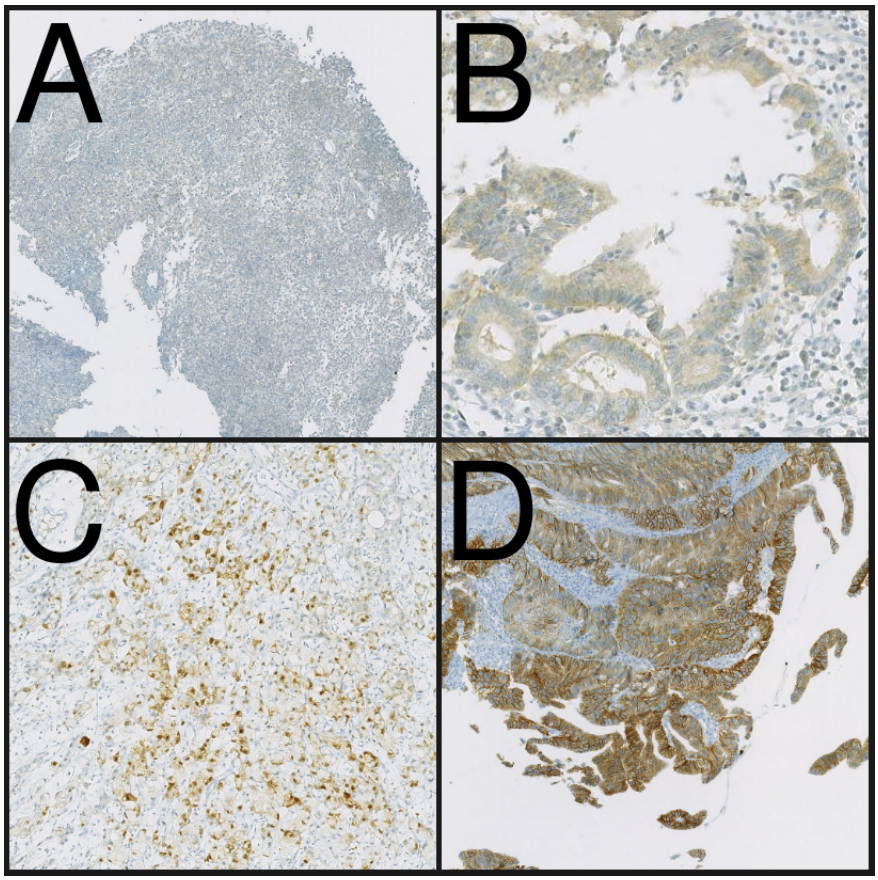


Figura 3.2.2: Ejemplo de ROIs resultantes tras aplicar proceso de filtrado, magnificación x10. A) 0, B) 1+, C) 2+, D) 3+.

3.3. Etiquetado generado para esta tesis

La gran discrepancia de las evaluaciones de los patólogos 1 y 2 respecto a la clasificación original del patólogo 0 es problemática, en especial al tener en cuenta que la evaluación de HER2 es necesaria para determinar si un paciente accede o no a un tratamiento que puede aumentar su sobrevida. Estos problemas clínicos con las anotaciones fueron confirmados por otros patólogos expertos, quienes señalaron que incluso zonas evaluadas por el patólogo 1 y 2 como HER2 positivas no eran parte del tejido canceroso.

Por otra parte, y desde un punto de vista meramente computacional y algorítmico, aún si las anotaciones de los patólogos 1 y 2 fueran correctas, estas seguirían sin ser suficientes para evaluar la sobreexpresión de HER2 sobre toda una biopsia y dar una clasificación global de una muestra. Esto, porque de acuerdo a la guía clínica de cáncer gástrico y lo expresado por otros patólogos expertos, la sobreexpresión de HER2 debe ser evaluada sólo en la componente invasiva del cáncer (aquella que invade la lámina propia o siguientes capas del estómago) y no en la componente *in situ* (aquella ubicada en el epitelio). Por ello, para generar un sistema computacional de ML que permita clasificar una biopsia completa, es necesario tener anotaciones de regiones que no deban ser consideradas para la clasificación de sobreexpresión de HER2, de forma similar a lo planteado por Vandenberghe y col. [89].

De esta forma, dados los problemas presentes en las anotaciones disponibles, se hizo necesario conseguir un nuevo etiquetado que fuera clínicamente correcto y que permitiera diferenciar zonas cancerosas de aquellas que no presentan neoplasia. Por ello, se le solicitó a un patólogo especialista y con experiencia en clasificación de HER2 en cáncer gástrico que etiquetara alguna de las biopsias del estudio *PRECISO*, anotando además las regiones de interés.

Dicho patólogo especialista (de ahora en adelante, patólogo 3) generó anotaciones para 34 biopsias con tinción IHC, además de la clasificación HER2 global de cada una de esas muestras. Por recomendación del propio especialista, las anotaciones fueron realizadas sólo en biopsias por resección, dado que allí se aplica la regla del 10 % y ésta es más fácil de aplicar en términos computacionales. Además, también por sugerencia del patólogo 3, las anotaciones de ROIs fueron en base a la reactividad de las células presentes en la región, las cuales se pueden mapear fácilmente a la clasificación HER2 correspondiente (Tabla 3.3.1).

Etiqueta Patólogo 3	Clasificación HER2 correspondiente
No tumor	No aplica
Sin reactividad	HER2 0
Reactividad positiva no lineal	HER2 0
Reactividad lineal casi imperceptible	HER2 1+
Reactividad lineal débil	HER2 2+
Reactividad lineal fuerte	HER2 3+

Tabla 3.3.1: Anotaciones realizadas por patólogo 3 y clasificación HER2 correspondiente, de acuerdo a método de Ruschoff/Hofmann para biopsias por resección.

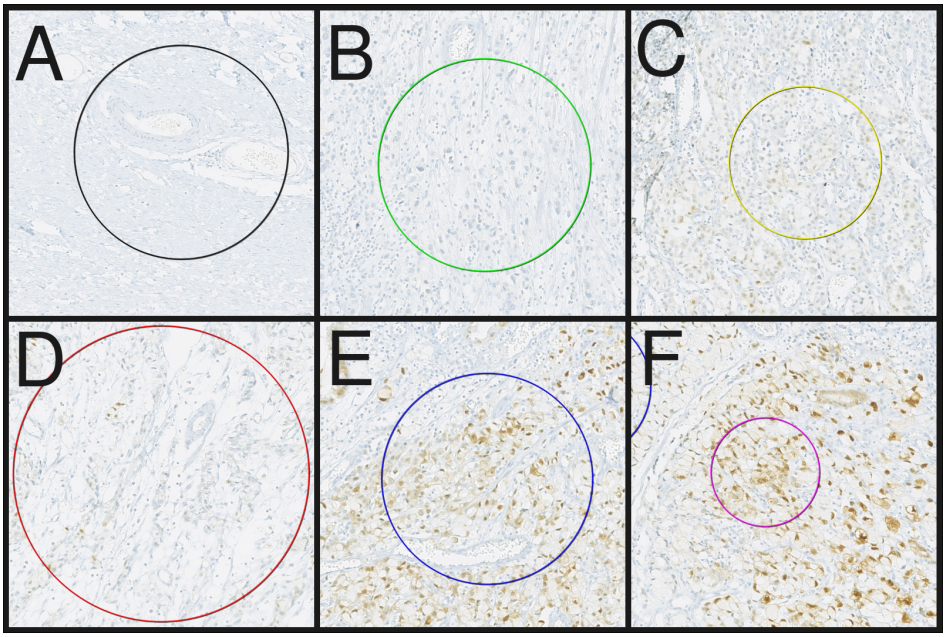


Figura 3.3.1: Ejemplo de anotaciones realizadas por patólogo 3, todas provenientes de la misma biopsia por resección a magnificación 10x. A) No tumor, B) sin reactividad, C) reactividad positiva no lineal, D) reactividad lineal casi imperceptible, E) reactividad lineal débil, F) reactividad lineal fuerte.

Las tablas 3.3.2 y 3.3.3 detallan, respectivamente, las clasificaciones de biopsias y las anotaciones de ROIs realizadas por el patólogo 3. Es posible apreciar que fueron obtenidos más ROIs que en el etiquetado previo (225 vs 150, sin considerar el proceso de filtrado), con mayor variabilidad en el tamaño de cada región. Por otra parte, a diferencia de las evaluaciones de los patólogos 1 y 2, la clasificación de biopsias del patólogo 3 señala que muchas biopsias son de tipo 0 y 1+, y sólo cinco pertenecen a los tipos 2+ y 3+. Esto último se condice con las clasificaciones realizadas por el patólogo 0, original del estudio *PRECISO*. En la tabla 3.3.4 se muestra cómo la concordancia entre el patólogo 0 y el patólogo 3 es mucho mayor a la concordancia entre patólogo 0, 1 y 2, lo cual es un buen indicador respecto a la idoneidad clínica del etiquetado generado para esta tesis. Además, dado que tanto la clase 0 como la clase 1+ corresponden a una sobreexpresión HER2 negativa, si se juntan las biopsias de estas clases bajo una misma etiqueta, se obtiene una concordancia aún más alta entre el patólogo 0 y 3. Esto se explica porque varias biopsias fueron evaluadas como negativas por ambos especialistas, con la diferencia de que uno la etiquetó como 0, mientras que el otro como 1+, y viceversa. Así, considerando sólo tres clases (negativo: 0/1+, equívoco, 2+, positivo: 3+), se tiene que entre las evaluaciones del patólogo 0 y el patólogo 3 existe un κ de Cohen igual a 0.73 y un α de Krippendorff de 0.77.

Por otro lado, desde el punto de vista algorítmico, este nuevo conjunto de anotaciones permite entrenar un clasificador que distinga zonas cancerosas, de tal manera de sólo allí aplicar la clasificación de sobreexpresión de HER2, siguiendo las guías clínicas y replicando el proceso diagnóstico. De todas formas, se debe tener en cuenta que una desventaja de este nuevo etiquetado es que fue realizado por sólo un patólogo, lo cual introduce un sesgo hacia la evaluación realizada por dicho especialista.

	0	1+	2+	3+	Total
Biopsias por resección	23	6	2	3	34

Tabla 3.3.2: Resumen de las evaluaciones realizadas por patólogo 3.

	Número de anotaciones	Radio promedio (std) (µm)	Radio mínimo (µm)	Radio máximo (µm)
No tumor	52	724.1 (362.2)	118.2	1617.8
Sin reactividad	50	480.1 (313.1)	115.0	1682.5
Reactividad no lineal	37	300.4 (152.7)	45.2	633.5
Reactividad lineal casi imperceptible	33	344.4 (167.8)	99.9	875.1
Reactividad lineal débil	33	241.1 (104.8)	79.5	478.6
Reactividad lineal fuerte	20	312.4 (192.3)	93.0	828.6
Total	225	437.1 (307.8)	45.2	1682.5

Tabla 3.3.3: Estadísticas de anotaciones realizadas por patólogo 3.

	Patólogos 0-3	Patólogos 1-3	Patólogos 2-3	Todos
4 clases (0, 1+, 2+, 3+)				
N° de muestras	34	5	23	34 *
% de concordancia	0.79	0.2	0.3	-
κ de Cohen	0.53	0.0	0.18	-
α de Krippendorff	0.63	-0.46	0.07	0.32
3 clases (Negativo, Equívoco, Positivo)**				
% de concordancia	0.94	0.4	0.4	-
κ de Cohen	0.73	0.0	0.22	-
α de Krippendorff	0.77	-0.27	0.11	0.39

Tabla 3.3.4: Concordancia de patólogo 3 con patólogos 0, 1 y 2, utilizando esquemas de 4 clases (0, 1+, 2+ y 3+) y 3 clases (Negativo, Equívoco, Positivo). Todas las muestras evaluadas corresponden a biopsias por resección. * α de Krippendorff puede ser computado aún con datos faltantes; así, el α calculado para el grupo de todos los patólogos corresponde a las 34 biopsias evaluadas por el patólogo 3. ** N°de muestras es igual en ambos esquemas.

Capítulo 4

Solución propuesta y resultados

Como se vio en la sección 2.8, varios trabajos abordan la sobreexpresión de la proteína HER2, especialmente en cáncer de mama. En particular, los trabajos de Vandenberghe y col. [89] y Saha y Chakraborty [78] utilizan DL para clasificar células de biopsias IHC en clases 0, 1+, 2+ y 3+ (además de otras clases como no tumor), y luego siguen las guías clínicas para generar una evaluación global de cada biopsia. No obstante, los datos disponibles al momento de comenzar este trabajo (estudio *PRECISO* y etiquetado de patólogos 1 y 2) no eran adecuados para realizar clasificación de células. Esto, porque i) no se disponía de anotaciones específicas de células, y ii) los ROIs no eran lo suficientemente homogéneos como para suponer que todas las células de un ROI presentaran la misma reactividad inmunohistoquímica.

De esta forma, considerando los datos disponibles, y teniendo en cuenta que se busca que el sistema generado replique el proceso diagnóstico llevado a cabo por los patólogos y que además sea interpretable por dichos médicos, es que se optó por una estrategia ligeramente distinta. Así, en vez de clasificar células como en los trabajos previamente mencionados, este trabajo se enfocó en clasificar *parches*, imágenes de mayor tamaño que en general contienen varias células.

Tal como se mencionó en el capítulo 3, el etiquetado realizado por los patólogos 1 y 2 no es suficiente para construir un modelo que sea capaz de generar una evaluación global de sobreexpresión de HER2 en una biopsia. Por ello, durante el transcurso de este proyecto se consiguió que un nuevo patólogo realizara un nuevo etiquetado para las biopsias del estudio *PRECISO*. Así, con estos dos conjuntos de datos en bruto, fueron definidos dos grandes experimentos (*macroexperimentos* de ahora en adelante); el primero con el etiquetado de los patólogos 1 y 2, y el segundo con el etiquetado del patólogo 3. El primer macroexperimento fue entendido como una exploración del problema y los resultados allí obtenidos fueron utilizados en el segundo macroexperimento.

En el resto del capítulo se detallarán los macroexperimentos I y II, el procesamiento de datos llevado a cabo, los modelos utilizados y los resultados obtenidos. Todos los experimentos fueron llevados a cabo utilizando el lenguaje de programación `Python 3.6`, el framework de redes neuronales `TensorFlow 1.14` con la API `Keras`, el framework de aprendizaje de máquinas `scikit-learn 0.21.3`, la librería matemática `NumPy 1.17.1`, la librería de análisis

y manipulación de datos `pandas 0.25.1` y las librerías de generación de gráficos `matplotlib 3.1.0` y `seaborn 0.9.0`. Además, los experimentos fueron ejecutados en el servidor *Apophis* del Departamento de Ciencias de la Computación de la Universidad de Chile. Este servidor cuenta con el sistema operativo `Fedora 27`, con kernel `Linux 4.18.19`, arquitectura `x86-64`, procesador `Intel Xeon E5-2640 v3` de 32 núcleos y cuatro GPUs `NVIDIA Tesla K40c` de 12 GB, con la plataforma de computación en paralelo `CUDA 10.0` y la librería de primitivas de redes neuronales profundas con aceleración para GPU `cuDNN 7.6.3`.

4.1. Macroexperimento I

El primer macroexperimento fue llevado a cabo utilizando las biopsias del estudio *PRECISO* junto al etiquetado disponible al momento de comenzar este trabajo, realizado por los patólogos 1 y 2.

4.1.1. Procesamiento de datos

Como se vio previamente, el etiquetado de los patólogos 1 y 2 consiste en un conjunto de ROIs circulares, donde cada región está clasificada en las clases 0, 1+, 2+ y 3+, además de la evaluación global de cada biopsia. Además, dada la alta varianza en las evaluaciones, este conjunto de datos fue filtrado, quedando así sólo 51 ROIs, todas provenientes del patólogo 2.

Para generar un dataset a partir de las biopsias y las anotaciones, los ROIs fueron extraídos utilizando el programa `ndpisplit` [24] en las magnificaciones 10x, 20x y 40x. Luego, a partir de cada ROI se extrajeron parches de 300x300 pixeles, con superposición de 50 pixeles tanto en el eje horizontal como el vertical. Un supuesto fuerte de este proceso es que todos los parches provenientes de un ROI comparten la misma clasificación HER2. Así, por ejemplo, si se tiene un ROI de tipo 3+ del cual se extraen 25 parches, todos esos parches también serán de tipo 3+, como se ilustra en la figura 4.1.1.

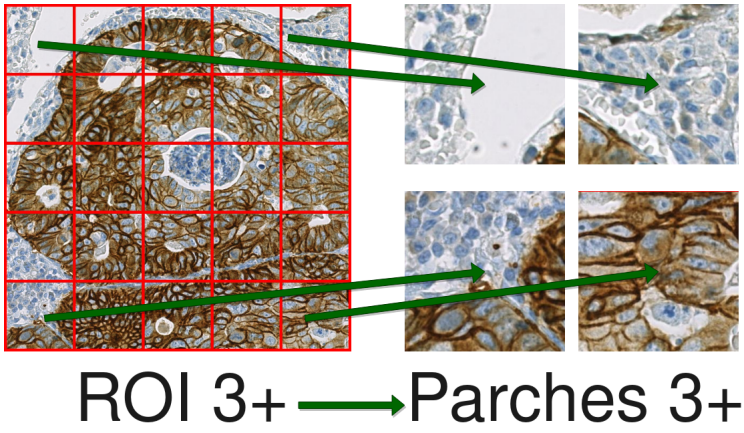


Figura 4.1.1: Esquema de extracción de parches desde un ROI. Todos los parches extraídos de un ROI de tipo 3+ también son de tipo 3+.

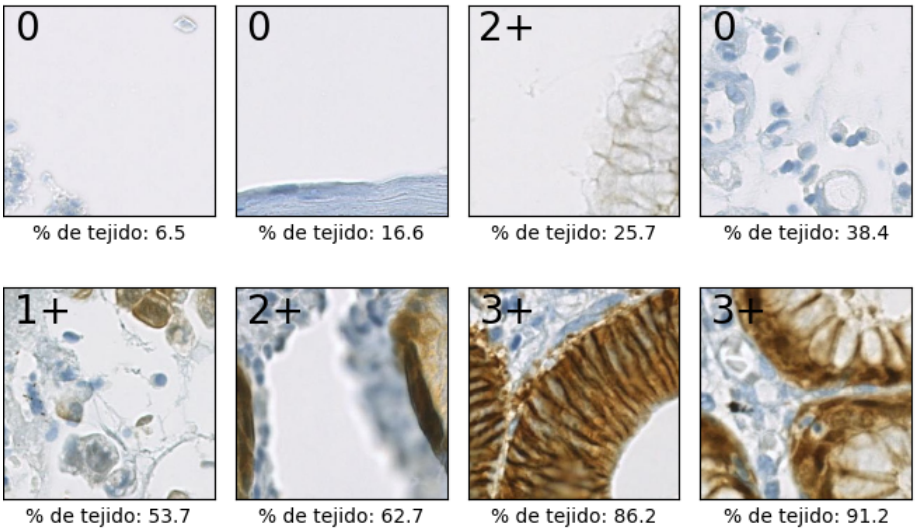


Figura 4.1.2: Ejemplos de parches extraídos de ROIs, junto a la proporción de tejido presente en cada uno de ellos. Además, en la esquina de cada parche, se encuentra anotada la clase a la que pertenece.

Adicionalmente, se calculó la cantidad de tejido presente en cada parche. Para ello, se cuantificó la proporción de fondo (o *background*) de cada parche, y se asumió que todo el resto de los pixeles corresponden a tejido de la muestra. Debido a que el fondo de las biopsias digitalizadas es de un color relativamente constante, este proceso pudo ser llevado a cabo utilizando sólo técnicas simples de procesamiento de imágenes. La figura 4.1.2 muestra ejemplos de parches junto a su proporción de tejido calculada. Dado que los parches con baja cantidad de tejido entregan poca información y pueden confundir a los modelos de ML, se decidió filtrar los datos y trabajar sólo con aquellos parches que tuvieran más de un 20 % de tejido.

Tras este proceso de extracción y filtrado de aquellos parches con baja proporción de tejido, fueron obtenidos tres subdatasets, uno para cada magnificación trabajada (10x, 20x, 40x). Además, dado que las clases 0 y 1+ corresponden a una sobreexpresión negativa de acuerdo a las guías clínicas, se procedió a juntar ambas clases. Así, en vez de clases 0, 1+, 2+ y 3+, en el macroexperimento I se trabajará con clases *negativo* (0 y 1+), *equívoco* (2+) y *positivo* (3+). La tabla 4.1.1 detalla las estadísticas de este conjunto de datos (de ahora en adelante, dataset 1).

4.1.2. Metodología experimental

Fueron diseñados una serie de experimentos, consistentes en clasificación de parches utilizando una red neuronal. La arquitectura escogida para esta tarea fue *Inception-v3* [86], desarrollada por Google para la competición *ImageNet Large Scale Visual Recognition Challenge*, donde alcanzó el segundo lugar en la competencia de clasificación de imágenes [77]. Esta red está formada por varios módulos pequeños, donde se implementa factorización de filtros convolucionales en filtros más pequeños y técnicas de reducción de dimensionalidad, lo cual permite disminuir la cantidad de parámetros entrenables de la red y a la vez mantener

Clasificación HER2	10x		20x		40x	
	Sin filtrar	Filtrado	Sin filtrar	Filtrado	Sin filtrar	Filtrado
Negativo	243	213	1227	1036	5607	4578
Equívoco	147	136	741	631	3357	2701
Positivo	246	216	1239	1010	5631	4334
Total	636	565	3207	2677	14595	11613

Tabla 4.1.1: Estadísticas de dataset 1, formado por los parches extraídos de los ROIs anotados por los patólogos 1 y 2, desagregados por magnificación. Filtro aplicado corresponde a parches con proporción de tejido mayor a 20 %.

un buen desempeño. Así, en la implementación provista por `Keras`, *Inception-v3* está compuesta por 152 capas y 23.851.784 parámetros entrenables, muchos menos que otras redes con desempeño similar, tales como *VGG19* (143.667.240 parámetros) y *ResNet-152* (60.419.944 parámetros) [8]. Además, `Keras` provee una versión de *Inception-v3* pre-entrenada con la base de datos de ImageNet, la cual puede ser utilizada fácilmente para transferencia de aprendizaje.

Luego, utilizando la red Inception-v3 pre-entrenada, se definieron cuatro experimentos, los que se detallan a continuación:

- **Entrenamiento simple:** se removió la capa final de la red *Inception-v3* (1000 neuronas, aquella que realiza la clasificación final) y se agregaron dos capas completamente conexas: una con 1024 neuronas y otra para realizar la clasificación, con 3 neuronas de salida, una para cada clase (negativo, equívoco, positivo). Además, la red fue configurada para que sólo las últimas dos capas nuevas sean entrenables, de tal manera que los parámetros previamente aprendidos por la red al ser entrenada en ImageNet queden fijos. La función de activación usada en la capa de 1024 neuronas fue ReLU ($\sigma(x) = max(0, x)$) y en la capa final se usó la función softmax ($S(y_i) = e^{y_i}/\Sigma_j e^{y_j}$),

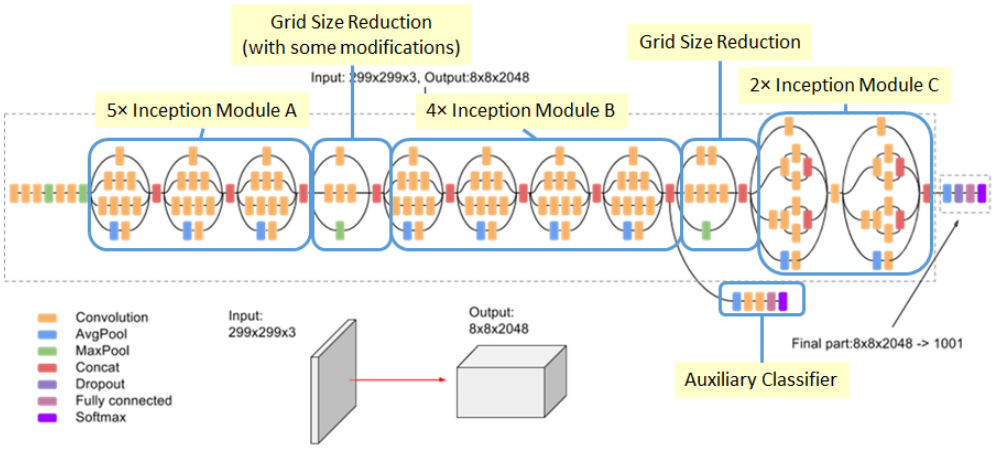


Figura 4.1.3: Arquitectura de *Inception v3*. Fuente: Tsang [88].

la cual tiene como característica producir salidas que pueden ser interpretadas como probabilidades. En esta configuración experimental, la red fue entrenada utilizando sólo los datos originales del dataset 1. El algoritmo de optimización utilizado fue *Adam* [45], una extensión del algoritmo de descenso estocástico del gradiente (SGD) que permite una convergencia más rápida. Además, la función de costo utilizada fue entropía cruzada categórica (*categorical cross entropy*), definida como

$$C = -\frac{1}{N}\sum_{\mathrm{i}}^{N}\sum_{c}^{C} y_{\mathrm{i},c} log(a_c(x_{\mathrm{i}})), \tag{4.1.1}$$

donde C es el número de clases ($C = 3$, en este caso), N es el número de elementos en el mini-conjunto de entrenamiento (*batch*) y x_{i} es un dato de entrada para la red. Además, y_{i} es un vector donde $y_{\mathrm{i}}, c = 1$ cuando c es la clase correspondiente a x_{i}, y todos los otros valores son 0. Finalmente, $a(x_{\mathrm{i}})$ es un vector de tamaño c con los valores predichos por la red. Dado que se usa la función de activación $softmax$ en la capa de salida, se tiene que $\sum_c^C a_c(x_{\mathrm{i}}) = 1$.

- **Entrenamiento con *data augmentation*:** idéntica a la configuración de entrenamiento simple, pero se utiliza *data augmentation* para artificialmente generar un conjunto de datos más grande, y por ende, reducir la posibilidad de sobreajuste. Así, sobre cada parche fueron aplicadas una o varias transformaciones aleatorias dentro de las siguientes posibilidades:
  - Rotación en un ángulo $\theta \in [-10°, 10°]$.
  - Desplazamiento horizontal en un $\Delta \in [-10\,\%, 10\,\%]$ del total de la imagen.
  - Desplazamiento vertical en un $\Delta \in [-10\,\%, 10\,\%]$ del total de la imagen.
  - Acercamiento (*zoom in*) de hasta un 10 %.
  - Alejamiento (*zoom out*) de hasta un 10 %.
  - Reflexión horizontal.
  - Reflexión vertical.

  Como color de fondo para los casos de rotación, desplazamiento y alejamiento se escogió un color similar al *background* de las slides. Se decidió no aplicar distorsiones de color o brillo, ya que estas características son importantes en la clasificación de HER2 y modificarlas podría inducir errores. La figura 4.1.4 muestra un parche original junto a algunas de sus transformaciones aleatorias.
- **Entrenamiento con *data augmentation* + *fine tuning*:** al igual que en las configuraciones previas, se reemplazó la capa final de la red con dos nuevas capas completamente conexas, además de utilizar *data augmentation*. Sin embargo, en vez de entrenar sólo las últimas dos capas completamente conexas, se entrenaron además los tres últimos bloques de la red *Inception v3*. Esto en teoría permite que la red aprenda *features* más adecuadas para este dataset, aún manteniendo los primeros bloques intactos. Además, en vez de utilizar el algoritmo de optimización *Adam*, se empleó SGD con una tasa de aprendizaje pequeña ($\eta = 0{,}0001$), tal como se recomienda en Karpathy [44].
- **Rentrenamiento total, con *data augmentation*:** idéntico al caso previo, pero ahora se procedió a reentrenar toda la red; i.e., todos los parámetros de la red podían ser modificados. Así, los pesos aprendidos en el dataset de ImageNet son sólo utilizados como punto de partida del entrenamiento.

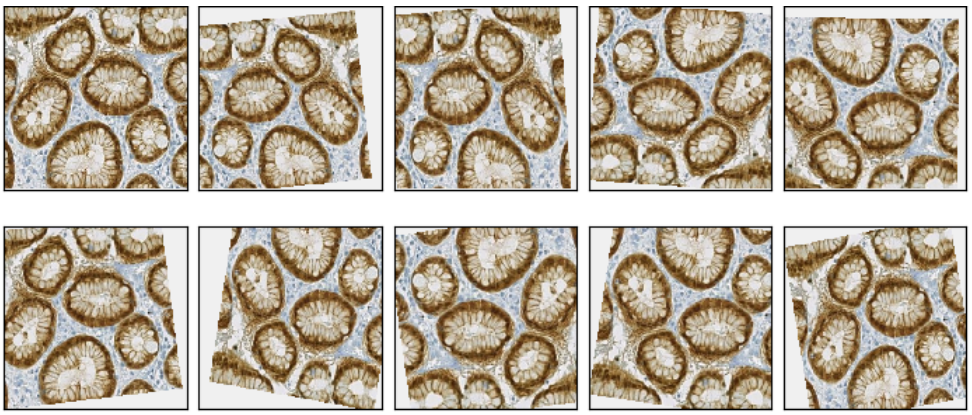


Figura 4.1.4: Parche extraído de ROI (esquina superior izquierda) junto a ejemplos de transformaciones aleatorias aplicadas sobre dicho parche.

La tabla 4.1.2 resume cada una de estas configuraciones experimentales. Además, para contrarrestar el desbalanceo existente entre las clases, en cada experimento se le asignó a cada clase un peso inversamente proporcional al número de datos de dicha etiqueta. Cada una de estas configuraciones fue entrenada utilizando parches a tres magnificaciones distintas (10x, 20x, 40x). Así, el macroexperimento I está conformado por 12 experimentos: 3 magnificaciones $\times$ 4 configuraciones.

Para evaluar el desempeño de cada uno de estos experimentos, se utilizó la estrategia de *k-fold*. No obstante, dado que los parches provenientes de una misma biopsia pueden tener correlación entre sí, fue necesario aplicar una pequeña variación a esta técnica. Así, en vez de dividir directamente el dataset de parches (como se suele hacer en k-fold), se particionó

	Simple	*Data augmentation*	*Data augmentation + fine tuning*	Reentrenamiento total
Arquitectura red	Inception v3	Inception v3	Inception v3	Inception v3
Capas entrenables	Últimas 2 capas	Últimas 2 capas	Últimos 3 bloques de Inception y capas completamente conexas	Todas
Función de activación	Softmax	Softmax	Softmax	Softmax
Algoritmo de optimización	Adam	Adam	SGD con $\eta = 0{,}0001$	SGD con $\eta = 0{,}0001$
Función de costo	Entropía cruzada categórica	Entropía cruzada categórica	Entropía cruzada categórica	Entropía cruzada categórica
Tamaño batch	32	32	32	32
Distorsiones aleatorias	No	Sí	Sí	Sí

Tabla 4.1.2: Resumen de cada configuración experimental del macroexperimento I.

primero el conjunto de biopsias en k subconjuntos, donde cada subconjunto contiene al menos una biopsia negativa (0 o 1+), una equívoca (2+) y una positiva (3+). Luego, utilizando este particionamiento de biopsias, fueron generados k subconjuntos de parches provenientes de las biopsias correspondientes. Además, cada uno de estos subconjuntos está dividido en 3 conjuntos más pequeños, uno para cada magnificación. Dado que el número de biopsias es bajo ($n_b = 17$), y con el fin de asegurar que en cada partición existiera una biopsia de cada tipo, se escogió $k = 5$. La figura 4.1.5 esquematiza la generación de estos subconjuntos. Además, utilizando esta estrategia, se tiene que:

- Nunca los conjuntos de entrenamiento y evaluación contendrán parches provenientes de la misma biopsia. Esto sería problemático, ya que puede existir correlación entre parches extraídos de una misma biopsia o de un mismo ROI, y se estaría violando la norma de que el conjunto de evaluación sea "ciego" al momento de entrenar el modelo.
- La comparación entre distintos modelos es más directa. Esto, porque en cada uno de los 12 experimentos, en la primera iteración de k-fold se utilizaron las biopsias b_1, b_2 y b_3 para evaluar, en la segunda iteración las biopsias b_4, b_5 y b_6, etc.

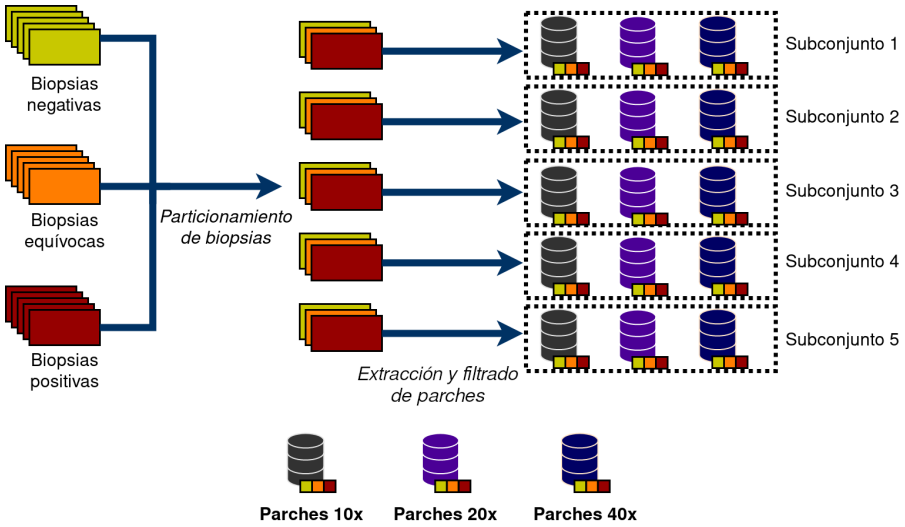


Figura 4.1.5: Esquema de generación de subconjuntos para entrenamiento con técnica de validación cruzada k-fold, con $k = 5$.

Con la técnica de validación cruzada k-fold, es posible obtener una evaluación de un modelo menos sesgada que usando la estrategia de sólo un conjunto de entrenamiento, uno de validación y otro de evaluación. Sin embargo, para la selección de hiperparámetros es necesario utilizar una variante llamada *k-fold anidado*. En esta técnica, en un loop interno se evalúan y seleccionan los mejores hiperparámetros, para luego en un loop externo usar dichos hiperparámetros óptimos y evaluar el modelo. En el caso del presente trabajo, el único hiperparámetro relevante de ajustar fue el número de épocas de entrenamiento, con un máximo de 50 épocas. La figura 4.1.6 esquematiza el funcionamiento del algoritmo de k-fold anidado, mientras que el pseudocódigo 4.1.1 detalla cómo fue usado este algoritmo para seleccionar el número óptimo de épocas de entrenamiento. Se debe mencionar que, si bien el número de iteraciones internas es en general independiente del número de iteraciones externas, en este proyecto se decidió utilizar el mismo K para ambos loops. Además, es

importante notar que, pese a que hasta ahora ha sido descrito como un algoritmo anidado, en el presente trabajo este algoritmo fue adaptado para ser usado de forma secuencial. De esta manera, primero fue ejecutada la sección correspondiente a la búsqueda de hiperparámetros, luego se analizaron los datos y fueron seleccionados los hiperparámetros óptimos, y recién entonces se procedió con la parte "externa" del algoritmo, correspondiente a la evaluación del modelo.

Algoritmo 4.1.1 Variante de k-fold anidado utilizado en este trabajo para encontrar número óptimo de épocas de entrenamiento. Adaptado desde Jin [43].

1: Generar K subconjuntos de datos, utilizando particionamiento previo de biopsias.
2: `max-epochs` ← número máximo de épocas de entrenamiento
3: **for** $\mathrm{i} = 1, 2, .., K$ **do**
4: `trainval` ← subconjuntos $1, ..., \mathrm{i}-1, \mathrm{i}+1, ..., K$
5: **for** $j = 1, ..., \mathrm{i}-1, \mathrm{i}+1, ..., K$ **do** ▷ Loop interno
6: `val` ← subconjunto j
7: `train` ← todos los subconjuntos, excepto i y j
8: Entrenar usando el conjunto `train` por `max-epochs` épocas.
9: Evaluar el desempeño del modelo en el conjunto `val` y guardar el número de épocas que entrega un mejor resultado.
10: **end for**
11: **end for**
12: *Manualmente, analizar los datos y seleccionar los hiperparámetros*
13: $\mathrm{e}_1, \mathrm{e}_2, ..., \mathrm{e}_K$ ← conjunto de mejores números de épocas. El valor e_i corresponde al número de épocas optimizado sobre los subconjuntos $1, ..., \mathrm{i}-1, \mathrm{i}+1, ..., K$.
14: **for** $\mathrm{i} = 1, 2, ..., K$ **do** ▷ Loop externo, mismo K que loop interno
15: `test` ← subconjunto i.
16: `trainval` ← subconjuntos $1, ..., \mathrm{i}-1, \mathrm{i}+1, ..., K$.
17: Entrenar el modelo con el conjunto `trainval` por e_i épocas
18: Evaluar desempeño del modelo con conjunto `test`
19: **end for**
20: Evaluar el desempeño global del modelo, ponderando las K iteraciones externas.

4.1.3. Resultados

Como se detalló previamente, se utilizó la técnica de k-fold anidado (con $k = 5$) para seleccionar los hiperparámetros óptimos y evaluar los modelos. Así, cada experimento consistió en el entrenamiento de $k \times (k-1)$ modelos para selección de hiperparámetros, y k modelos para evaluación. Luego, en cada experimento fueron entrenadas $k^2 = 5^2 = 25$ redes neuronales. Y dado que el macroexperimento I consiste de 12 experimentos (3 magnificaciones $\times$ 4 configuraciones experimentales), en total fueron entrenadas $25 \times 12 = 300$ redes neuronales convolucionales.

De esta forma, primero fueron evaluados los entrenamientos "internos" para seleccionar los hiperparámetros correctos, y luego los entrenamientos "externos".

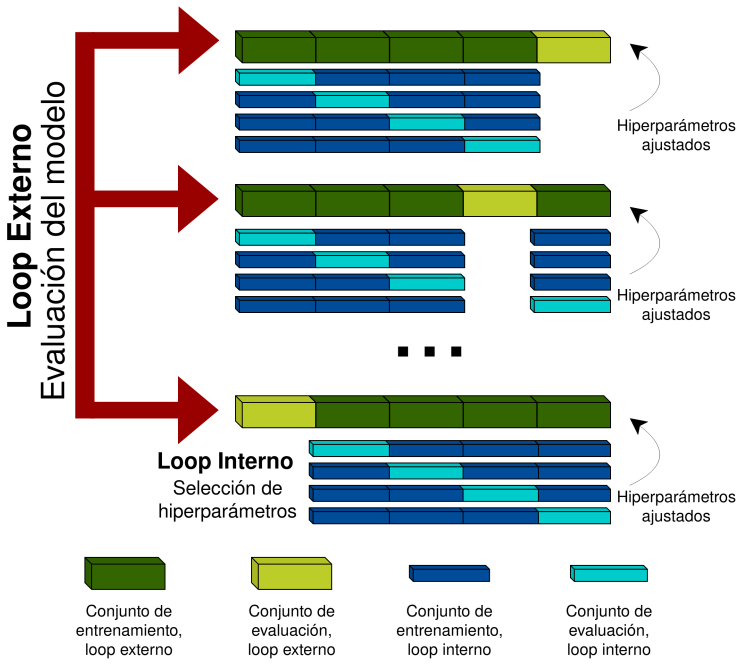


Figura 4.1.6: Esquema de k-fold anidado. Adaptado desde Jin [43].

4.1.3.1. Selección de hiperparámetros

Para cada entrenamiento de cada modelo, se guardó un archivo con los valores de exactitud (*accuracy*) y pérdida para cada época de entrenamiento, tanto para los conjuntos de entrenamiento como validación. Estos datos fueron guardados en formato `csv` y procesados utilizando la librería `pandas`. Luego, para calcular el mejor número de épocas, fueron promediados los resultados de cada iteración interna, y se seleccionó aquel número de épocas en que se consiguió una mejor exactitud en los conjuntos de validación. Las tablas 4.1.3, 4.1.4 y 4.1.5 detallan los hiperparámetros seleccionados para cada subconjunto en cada configuración experimental y cada magnificación.

	Subconjunto 1	Subconjunto 2	Subconjunto 3	Subconjunto 4	Subconjunto 5
Simple	7	18	11	4	38
Data augmentation	5	8	21	7	3
Data augmentation + fine tuning	49	42	27	50	39
Reentrenamiento total	47	45	34	24	28

Tabla 4.1.3: Hiperparámetros seleccionados mediante k-fold anidado, magnificación 10x.

	Subconjunto 1	Subconjunto 2	Subconjunto 3	Subconjunto 4	Subconjunto 5
Simple	6	4	8	43	3
Data augmentation	1	27	5	44	3
Data augmentation + fine tuning	28	7	7	46	12
Reentrenamiento total	30	11	11	40	9

Tabla 4.1.4: Hiperparámetros seleccionados mediante k-fold anidado, magnificación 20x.

	Subconjunto 1	Subconjunto 2	Subconjunto 3	Subconjunto 4	Subconjunto 5
Simple	6	41	5	4	47
Data augmentation	10	50	4	3	44
Data augmentation + fine tuning	17	46	45	14	42
Reentrenamiento total	33	2	50	47	31

Tabla 4.1.5: Hiperparámetros seleccionados mediante k-fold anidado, magnificación 40x.

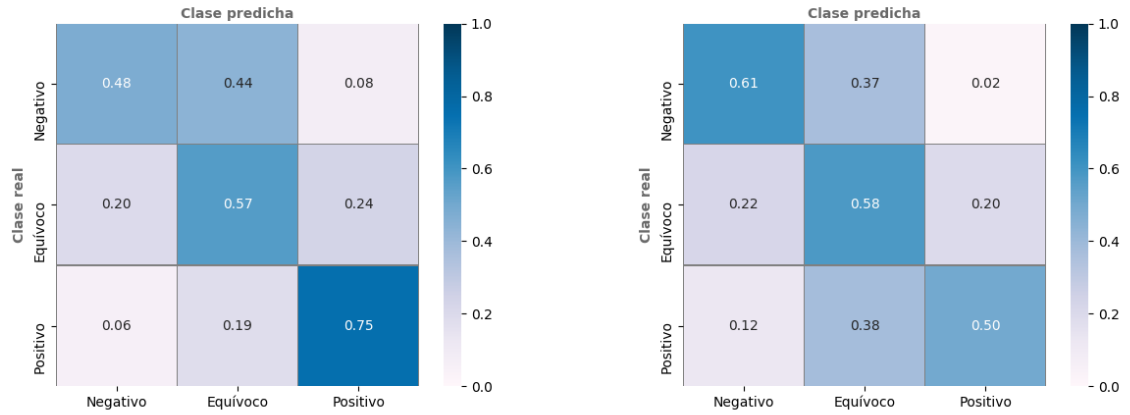
4.1.3.2. Evaluación de modelos

Utilizados los hiperparámetros óptimos escogidos en el loop interno de k-fold anidado, se procedió a ejecutar el loop externo del algoritmo. De esta forma, para cada uno de los doce experimentos se corrieron $k = 5$ subexperimentos. Luego, con los modelos ya entrenados, se procedió a predecir la clasificación de los parches de los conjuntos de evaluación correspondientes. Así, un modelo entrenado con los subconjuntos 1, 2, 3 y 4 fue evaluado en el subconjunto 5; uno entrenado en los subconjuntos 1, 2, 3 y 5 fue evaluado en el subconjunto 4; etc. Posteriormente, las predicciones realizadas por cada modelo de una misma configuración experimental y magnificación fueron concatenadas, para finalmente con estos datos calcular las métricas que se reportan en las tablas 4.1.6, 4.1.7 y 4.1.8. Esta estrategia de evaluación es fuertemente recomendada en Forman y Scholz [29]. Además, las figuras 4.1.7, 4.1.8, 4.1.9 muestran las matrices de confusión normalizadas para los modelos que obtuvieron mejor desempeño en cada magnificación.

Es posible notar que bajo la métrica de F1-score (más adecuada para situaciones donde los datos están desbalanceados), los experimentos que obtuvieron mejor resultado fueron reentrenamiento total con magnificación 10x y configuración simple con magnificación 20x.

	Negativo	Equívoco	Positivo	**Promedio**
Simple				
Precisión	0.76	0.35	0.62	0.61
Recuperación	0.61	0.51	0.56	0.57
F1-score	0.68	0.42	0.59	0.58
Exactitud	-	-	-	0.57
Data augmentation				
Precisión	0.68	0.34	0.76	0.63
Recuperación	0.68	0.56	0.46	0.56
F1-score	0.68	0.43	0.57	0.58
Exactitud	-	-	-	0.56
Data augmentation* + *fine tuning				
Precisión	0.69	0.33	0.77	0.64
Recuperación	0.61	0.58	0.5	0.56
F1-score	0.65	0.42	0.61	0.58
Exactitud	-	-	-	0.56
Reentrenamiento total				
Precisión	0.72	0.36	0.77	0.65
Recuperación	0.48	0.57	0.75	0.61
F1-score	0.57	0.44	0.76	**0.61**
Exactitud	-	-	-	0.61

Tabla 4.1.6: Resumen de resultados conseguidos con magnificación 10x. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.



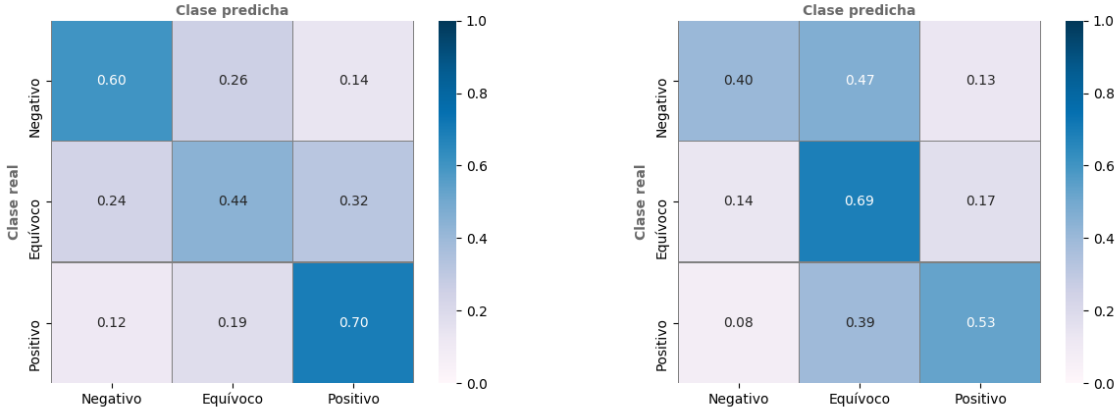
(a) Reentrenamiento total con magnificación 10x.

(b) *Fine tuning* y *data augmentation* con magnificación 10x.

Figura 4.1.7: Matrices de confusión normalizadas de experimentos con mejores resultados para magnificación 10x.

	Negativo	Equívoco	Positivo	**Promedio**
Simple				
Precisión	0.7	0.38	0.67	0.61
Recuperación	0.6	0.44	0.7	0.6
F1-score	0.65	0.41	0.68	**0.61**
Exactitud	-	-	-	0.6
Data augmentation				
Precisión	0.72	0.33	0.63	0.59
Recuperación	0.34	0.58	0.66	0.52
F1-score	0.46	0.42	0.64	0.52
Exactitud	-	-	-	0.52
Data augmentation* + *fine tuning				
Precisión	0.72	0.33	0.69	0.61
Recuperación	0.4	0.69	0.53	0.52
F1-score	0.52	0.45	0.6	0.53
Exactitud	-	-	-	0.52
Reentrenamiento total				
Precisión	0.45	0.24	0.69	0.49
Recuperación	0.23	0.4	0.74	0.46
F1-score	0.31	0.3	0.71	0.46
Exactitud	-	-	-	0.46

Tabla 4.1.7: Resumen de resultados conseguidos con magnificación 20x. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.



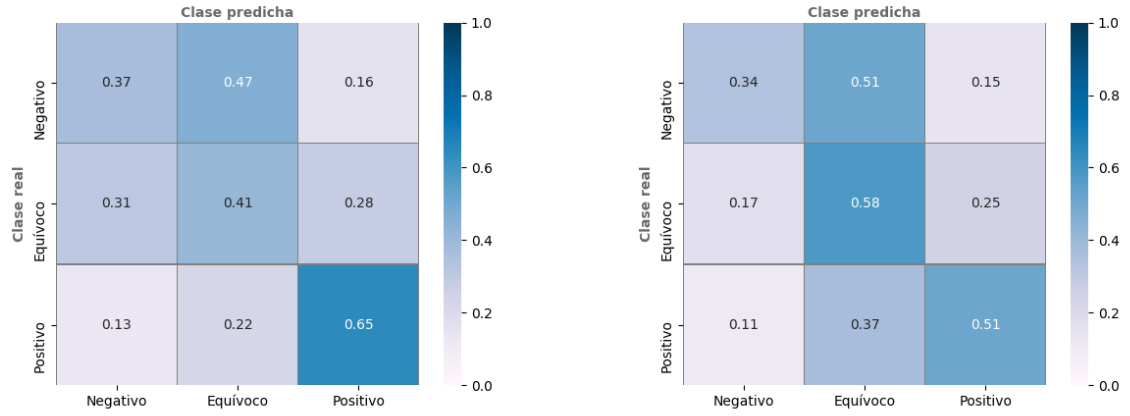
(a) Configuración simple con magnificación 20x.

(b) *Fine tuning* y *data augmentation* con magnificación 20x.

Figura 4.1.8: Matrices de confusión normalizadas de experimentos con mejores resultados para magnificación 20x.

	Negativo	Equívoco	Positivo	**Promedio**
Simple				
Precisión	0.54	0.22	0.53	0.46
Recuperación	0.25	0.48	0.45	0.38
F1-score	0.34	0.3	0.49	0.39
Exactitud	-	-	-	0.38
Data augmentation				
Precisión	0.47	0.23	0.5	0.43
Recuperación	0.11	0.6	0.4	0.33
F1-score	0.17	0.33	0.44	0.31
Exactitud	-	-	-	0.33
Data augmentation + fine tuning				
Precisión	0.62	0.28	0.63	0.54
Recuperación	0.34	0.58	0.51	0.46
F1-score	0.44	0.38	0.56	0.47
Exactitud	-	-	-	0.46
Reentrenamiento total				
Precisión	0.55	0.26	0.65	0.52
Recuperación	0.37	0.41	0.65	0.48
F1-score	0.44	0.32	0.65	0.49
Exactitud	-	-	-	0.48

Tabla 4.1.8: Resumen de resultados conseguidos con magnificación 40x. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.



(a) Reentrenamiento total con magnificación 40x.

(b) *Fine tuning* y *data augmentation* con magnificación 40x.

Figura 4.1.9: Matrices de confusión normalizadas de experimentos con mejores resultados para magnificación 40x.

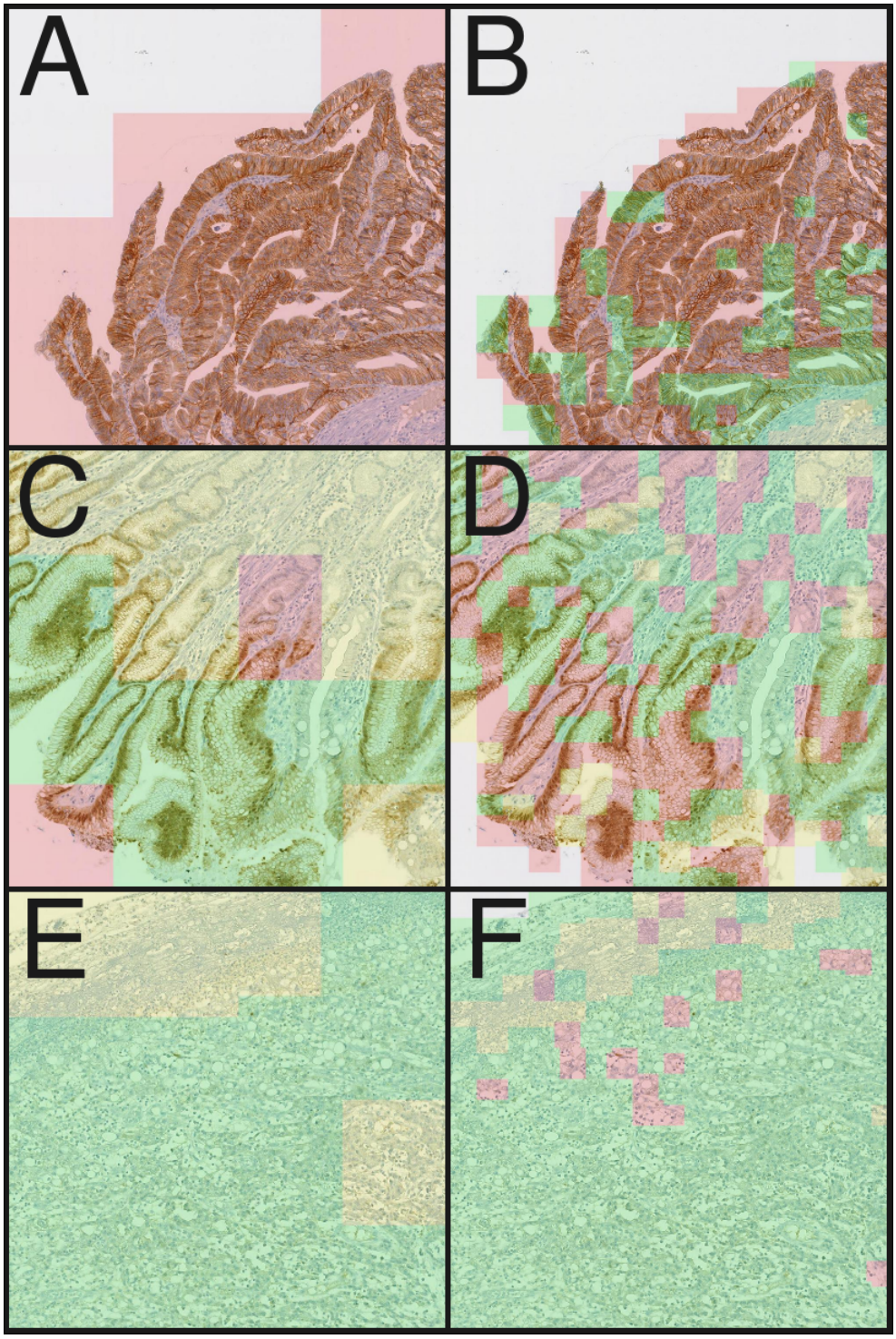


Figura 4.1.10: Ejemplos de clasificación de parches extraídos de ROIs en macroexperimento I. Cada ROI está formado por varios parches, y sobre cada parche, se pintó con transparencia la clasificación predicha (verde: HER2 negativo, amarillo: HER2 equívoco, rojo: HER2 positivo). Visualizaciones de lado izquierdo fueron generadas usando las redes entrenadas en el experimento de reentrenamiento total a magnificación 10x, mientras que las del lado derecho corresponden a reentrenamiento total con magnificación 40x. A) y B) ROI de tipo 3+, C) y D), ROI de tipo 2+, E) y F) ROI de tipo 0. Todos los ROIs fueron anotados por el patólogo 2.

4.2. Macroexperimento II

Para llevar a cabo el segundo macroexperimento, nuevamente fueron empleadas las biopsias provenientes del estudio *PRECISO*. Sin embargo, a diferencia del macroexperimento I, donde se empleó el etiquetado ya disponible generado por los patólogos 1 y 2, para este macroexperimento fue necesario obtener un nuevo etiquetado. Tal como se describió en la sección 3.3, se consiguió que un nuevo especialista (patólogo 3) etiquetara las biopsias, generando anotaciones para zonas con sobreexpresión HER2 y también para zonas que no presentan tejido canceroso.

Como se comentó previamente, el macroexperimento I fue considerado como un análisis exploratorio; en particular, fue posible estudiar qué configuración experimental entregaba mejores resultados y aplicar esa configuración para los experimentos del macroexperimento II. Así, se vio que las configuraciones de reentrenamiento total con magnificación 10x y la de entrenamiento simple con magnificación 20x fueron las que obtuvieron mejores resultados (ambas con F1-score = 0.61). Para dirimir entre ambas configuraciones, se tomó en cuenta que, en términos clínicos, es mucho más relevante recuperar correctamente aquellas biopsias con clasificación 2+ o 3+, ya que ambas abren una puerta para un tratamiento que puede alargar la sobrevida de los pacientes. Por ello, se decidió priorizar aquella configuración que obtuviera mejor resultado en las clases *equívoco* y *negativo*, en especial respecto a la métrica de recuperación. De esta forma, tras estudiar los resultados obtenidos en ambas configuraciones, se decidió utilizar la configuración de reentrenamiento total con magnificación 10x para todos los experimentos del macroexperimento II.

Luego, se siguió el mismo proceso de extracción de ROIs y parches que en el macroexperimento I, pero sólo para magnificación 10x. Así, se generó el dataset 2, cuyas estadísticas se detallan en la tabla 4.2.1

Clasificación HER2	Sin filtrar	Filtrado
No tumor	2269	2206
Sin reactividad	1029	1007
Reactividad no lineal	222	222
Reactividad lineal casi imperceptible	238	238
Reactividad lineal débil	99	99
Reactividad lineal fuerte	147	147
Total	4004	3919

Tabla 4.2.1: Estadísticas de dataset 2, formado por los parches extraídos de los ROIs anotados por el patólogo 3. Todos los parches fueron extraídos a magnificación 10x. Filtro aplicado corresponde a parches con proporción de tejido mayor a 20 %.

4.2.1. Metodología experimental

La principal ventaja de este nuevo dataset, además de tener mayor concordancia con lo evaluado originalmente por el patólogo del estudio *PRECISO*, es que dado que contiene anotaciones de zonas no cancerosas, permite generar un modelo que clasifique biopsias completas. De esta forma, fue necesario definir experimentos que fueran en esta línea.

En primer lugar, es posible notar que, a diferencia del dataset 1, en este nuevo conjunto de datos las clases de las biopsias están claramente desbalanceadas, con una fuerte tendencia hacia la clase 0 (tabla 3.3.2). Esto dificulta la estrategia de k-fold sobre biopsias, ya que varias particiones quedarían sin muestras de las clases 2+ y 3+. Por otro lado, un caso de uso real del presente proyecto sería hacer uso del algoritmo entrenado sobre muchas biopsias para evaluar una biopsia completamente nueva, jamás vista previamente por el modelo. Tomando en cuenta estas dos condiciones, se decidió seguir una estrategia de entrenamiento y evaluación conocida como *leave-one-out cross validation* ($LOOCV$). Esta estrategia es equivalente a k-fold con $k = n =$ número de datos, donde en cada iteración se entrena con todos los datos menos uno, y ese dato restante es usado para evaluación. Así, de forma análoga a lo hecho en el macroexperimento I, LOOCV fue aplicado a biopsias, usando los parches extraídos de $n - 1$ biopsias para entrenar, y los parches de la biopsia restante para evaluar. Dado que la complejidad algorítmica de $k-$fold anidado es $O(k^2)$, aplicar esta estrategia cuando $k = 34$ y se realiza un reentrenamiento completo de una red con tantos parámetros como Inception v3 se vuelve computacionalmente costoso. Por ello, se desechó la fase de selección de hiperparámetros y así, para todos los experimentos del macroexperimento II se utilizó un número fijo de épocas $\eta = 30$. Este valor es cercano al promedio de los números de épocas óptimas utilizadas en la configuración de reentrenamiento total con magnificación 10x en el macroexperimento I.

De esta manera, se definieron dos experimentos:

- **Todo en uno:** consiste en el entrenamiento de una sola red para la clasificación de parches en las clases *No tumor*, *Sin reactividad*, *Reactividad no lineal*, *Reactividad lineal casi imperceptible*, *Reactividad lineal débil* y *Reactividad lineal fuerte*.
- **Cascada:** consistente en el el entrenamiento de dos redes, una para clasificación de *tumor/no tumor*, y otra para clasificar las clases relativas a la reactividad HER2. Respecto a la clasificación *tumor/no tumor* es importante notar que:
  - En este entrenamiento, todos los parches relativos a reactividad HER2 son considerados como zonas cancerosas, ya que la sobreexpresión de esta proteína sólo debe ser evaluada en aquellas regiones de la biopsia que sí son cancerosas.
  - Dado que es clasificación binaria, se utilizó una red con sólo una neurona de salida. Por ello, no se empleó la función de activación *Softmax*, ya que bajo esta configuración, esta función de activación siempre entrega como resultado 1. Así, se usó la función *Sigmoide*, definida como $f(x) = \frac{1}{1+\mathrm{e}^{-x}}$. Del mismo modo, la función de costo fue entropía cruzada binaria (*binary cross entropy*), definida como

$$C = -\frac{1}{N}\sum_{\mathrm{i}}^{N} y_{\mathrm{i}} log(a(x_{\mathrm{i}})) + (1 - y_{\mathrm{i}}) log(1 - a(x_{\mathrm{i}})), \tag{4.2.1}$$

	Todo en uno	Cascada	
		Tumor / No tumor	Reactividad HER2
Magnificación	10x	10x	10x
Número de salidas	6	1	5
Arquitectura red	Inception v3	Inception v3	Inception v3
Capas entrenables	Todas	Todas	Todas
Función de activación	Softmax	Sigmoide	Softmax
Algoritmo de optimización	SGD con $\eta = 0{,}0001$	SGD con $\eta = 0{,}0001$	SGD con $\eta = 0{,}0001$
Función de costo	Entropía cruzada categórica	Entropía cruzada binaria	Entropía cruzada categórica
Tamaño batch	32	32	32
Distorsiones aleatorias	Sí	Sí	Sí
Épocas de entrenamiento	30	30	30

Tabla 4.2.2: Resumen de cada configuración experimental del macroexperimento II.

donde N es el número de elementos en el mini-conjunto de entrenamiento (*batch*), x_i es un dato de entrada para la red, $y_i \in \{0, 1\}$ es su clase correspondiente, y $a(x_i)$ es la salida de la red. Esta definición es equivalente a entropía cruzada categórica cuando el número de clases es $C = 2$.

– Dado que la salida de la red es un número continuo $\in [0, 1]$, para realizar clasificación propiamente tal es necesario definir un valor umbral T que divida el espacio de salida. Así, si la salida $a(x_i) < T$, se dirá que x_i pertenece a la clase *No tumor*; en caso contrario, corresponderá a la clase *Tumor*. Este umbral T será definido post-entrenamiento del modelo.

Cada experimento tiene dos fases: entrenamiento y evaluación, llevadas a cabo utilizando la estrategia de LOOCV aplicada a biopsias. La tabla 4.2.2 resume la configuración utilizada en el entrenamiento de cada uno de estos experimentos, mientras que las figuras 4.2.1 y 4.2.2 esquematizan la fase de entrenamiento de los experimentos *todo en uno* y *en cascada*, respectivamente.

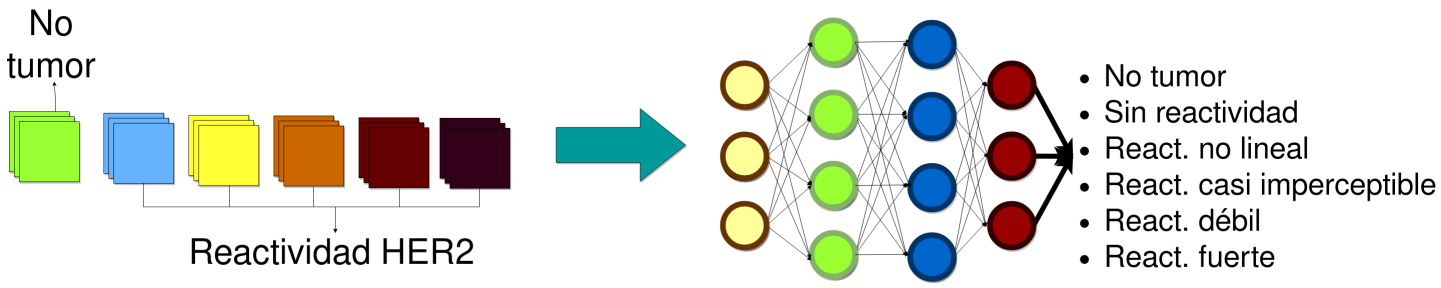


Figura 4.2.1: Esquema del experimento *todo en uno*.

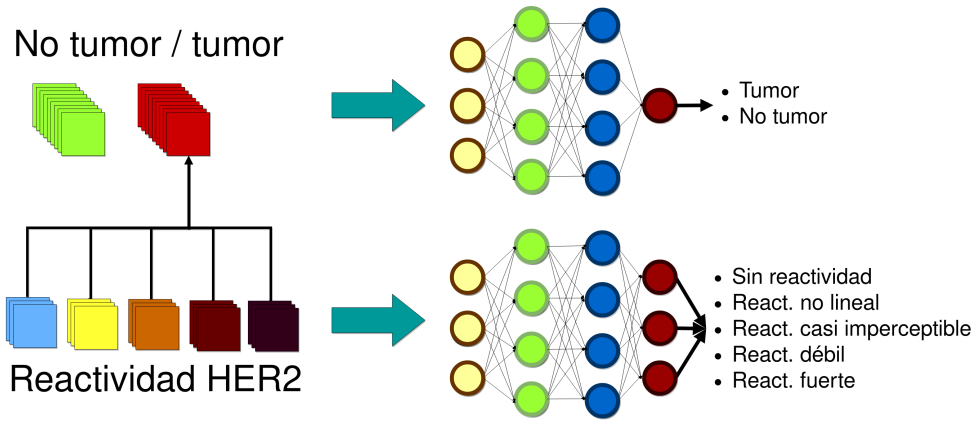


Figura 4.2.2: Esquema del experimento *en cascada*.

Por otra parte, la fase de evaluación de cada uno de estos experimentos está a su vez dividida en dos partes:

- **Evaluación de clasificación de parches de ROIs**: análoga a la evaluación realizada en el macroexperimento I. Se evalúa, mediante distintas métricas, la performance de los clasificadores sobre los parches extraídos de ROIs, cuya etiqueta ya se conoce (bajo el supuesto de que todos los parches extraídos desde un ROI corresponden a la misma etiqueta). Además, en el caso particular de la clasificación binaria de *tumor / no tumor* (parte del esquema *en cascada*), se realiza un análisis utilizando curvas ROC y AUC.
- **Evaluación de clasificación de biopsias**: este proceso de evaluación es distinto, dado que, si bien cada biopsia tiene su clasificación global, no es posible asumir que cada zona de la biopsia es de ese tipo. En particular, la guía clínica muestra que basta con que, por ejemplo, un 10 % de las células de una biopsia por resección tengan reactividad lineal fuerte para que toda la biopsia sea de tipo 3+. Por ello, en este proceso de evaluación se sigue el siguiente algoritmo:
  1. Se extraen todos los parches de una biopsia, sin superposición e independientemente de si están contenidos en un ROI o no.
  2. Se calcula la proporción de tejido contenida en cada parche.
  3. Se utiliza la CNN ya entrenada con todas las biopsias restantes para clasificar cada parche.
  4. Se seleccionan todos los parches que fueron identificados como cancerosos, junto a su reactividad HER2.
  5. Usando la proporción de tejido de cada parche y la correspondiente clasificación predicha por la red, además del mapeo entre reactividad y clasificación HER2 (tabla 3.3.1), se calcula el total de tejido correspondiente a cada clase (0, 1+, 2+, 3+). Es importante notar que, en este punto, se asume una densidad celular constante para cada parche canceroso; es decir, se parte del supuesto que cada parche tiene una cantidad de células proporcional a su cantidad de tejido. Además, si bien existe una diferencia en la densidad celular entre parches cancerosos y sanos, esto no afecta el supuesto realizado, ya que las regiones sanas no influyen en la evaluación de sobreexpresión HER2.

6. Finalmente, siguiendo el algoritmo de las guías clínicas, se genera una clasificación global para la biopsia. Como en este macroexperimento se trabaja sólo con biopsias por resección, el algoritmo usado corresponde sólo a la regla de 10 %, y no se debe evaluar la presencia de clusters como ocurre en el caso de biopsias endoscópicas. Además, si bien no está explicitado en la guía clínica, de acuerdo a lo explicado por patólogos especialistas, la regla del 10 % debe ser aplicada de modo "acumulativo", tal como se detalla en el algoritmo 4.2.1.

En el caso del experimento *todo en uno*, cada parche es clasificado directamente en *No tumor* o su reactividad HER2 correspondiente. Por otra parte, en el experimento *en cascada*, primero cada parche es clasificado en *Tumor* o *No tumor*, y luego todas aquellas imágenes identificadas como cancerosas son ingresadas a una segunda red, que las clasifica de acuerdo a su reactividad HER2. La figura 4.2.3 esquematiza la evaluación de biopsias para el experimento *en cascada*.

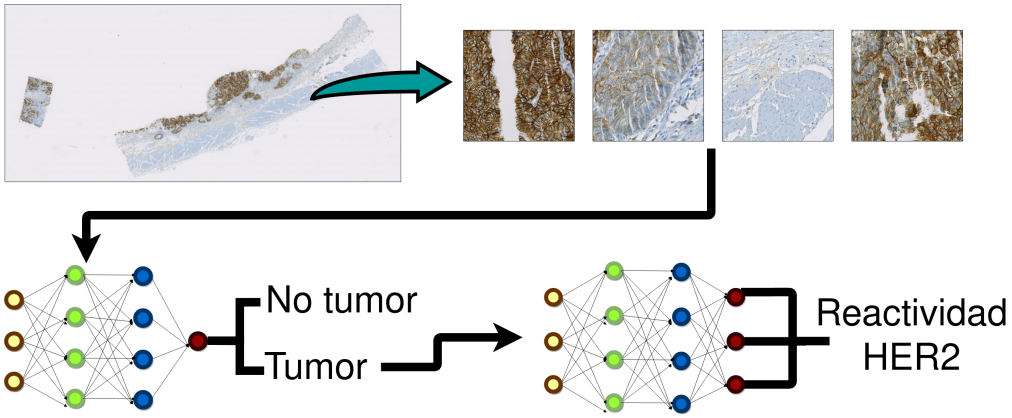


Figura 4.2.3: Esquema de evaluación de parches extraídos directamente de biopsias en experimento *en cascada*.

Algoritmo 4.2.1 Algoritmo "acumulativo" de evaluación de biopsias.

function ClasificarBiopsia(Biopsia i)
  $t^i_0, t^i_{1+}, t^i_{2+}, t^i_{3+} \leftarrow$ cantidad de tejido de clases 0, 1+, 2+ y 3+ en biopsia i.
  $t^i_{total} \leftarrow t^i_0 + t^i_{1+} + t^i_{2+} + t^i_{3+}$
  if $t^i_{3+} \geq 0{,}1 \times t^i_{total}$ **then return** 3+
  else if $t^i_{3+} + t^i_{2+} \geq 0{,}1 \times t^i_{total}$ **then return** 2+
  else if $t^i_{3+} + t^i_{2+} + t^i_{1+} \geq 0{,}1 \times t^i_{total}$ **then return** 1+
  else return 0
  end if
end function

4.2.2. Resultados

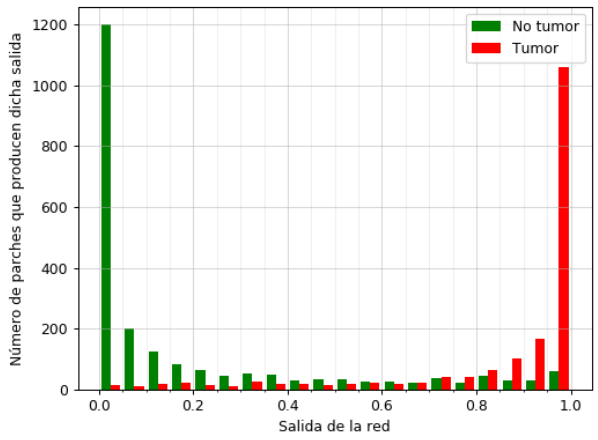
Como se describió previamente, se utilizó la estrategia de LOOCV aplicada a biopsias para el entrenamiento y evaluación de los modelos. De esta forma, fueron generados $k = 34$

modelos para el experimento *todo en uno*, y $2 \times 34 = 68$ modelos para el experimento *en cascada*. Así, en total en el macroexperimento II fueron entrenadas $3 \times 34 = 102$ redes neuronales convolucionales.

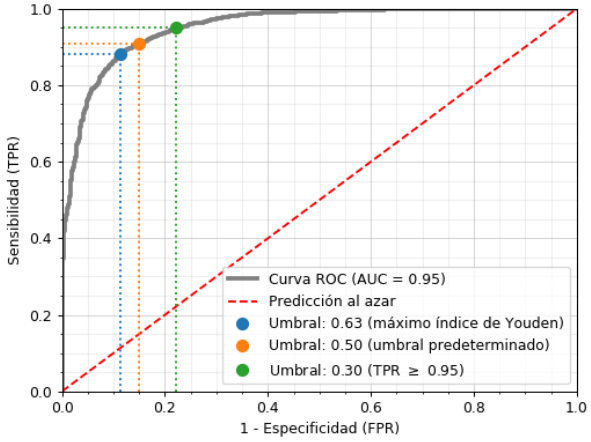
4.2.2.1. Evaluación de clasificación de parches de ROIs

Siguiendo la estrategia de LOOCV, cada modelo fue entrenado con los parches extraídos de los ROIs provenientes de $k-1 = 33$ biopsias, y luego la evaluación se realizó utilizando los parches extraídos de los ROIs de la biopsia restante. Luego, para realizar la evaluación global de cada experimento, se concatenaron los resultados de cada uno de los $k = 34$ modelos entrenados y sobre esos resultados se calcularon las métricas correspondientes.

Los histogramas de la figura 4.2.4a muestran la distribución de los valores de salida producidos por los clasificadores binarios de *tumor / no tumor*. Si bien la separación entre clases es muy marcada, es posible notar que existe superposición en los valores intermedios, entre 0.2 y 0.8. De este modo, se hace necesario seleccionar un umbral que permita, en base a alguna métrica determinada, obtener un mejor desempeño en las clasificaciones. Es importante destacar que esta decisión se debe basar en conocimiento del dominio (*domain knowledge*) sobre el cual se trabaja, evaluando el costo asociado a cada falso positivo y falso negativo que se pueda obtener en la clasificación [42]. Dado el problema médico que se aborda en este trabajo, resulta claro que un falso negativo tiene un costo mayor a un falso positivo; clasificar un parche tumoroso como sano es, en teoría, significativamente más grave que identificar un parche sano como tumoroso. Así, la métrica a optimizar será la de sensibilidad, imponiendo un mínimo de 95 %. La figura 4.2.4b muestra la curva ROC obtenida al evaluar los clasificadores tumor/no tumor ($AUC = 0{,}95$), destacando el umbral que logra la sensibilidad de 95 % ($T = 0{,}3$), además del umbral predeterminado ($T = 0{,}5$) y aquel que maximiza el índice J de Youden ($T = 0{,}63$). Es importante notar también que, en la parte superior de la curva,



(a) Histogramas de valores de salida producidos por el clasificador binario *tumor / no tumor*.



(b) Curva ROC de clasificador binario, destacando umbrales de decisión relevantes.

Figura 4.2.4: Análisis de umbral de decisión para clasificación binaria de *tumor / no tumor*.

pequeñas ganancias en sensibilidad implican detrimentos cada vez mayores en especificidad, por lo cual se decidió no optimizar la sensibilidad más allá del mencionado 95 %.

Tras dicho análisis, se selecciona el umbral $T = 0{,}3$ para la clasificación binaria de tumor/no tumor, correspondiente al esquema *en cascada*. Las tablas 4.2.3 y 4.2.4 detallan las métricas de evaluación obtenidas para los diferentes esquemas, mientras que las figuras 4.2.5 y 4.2.6 muestran las matrices de confusión conseguidas. Algo importante de notar es que, en la clasificación binaria de *tumor / no tumor*, la recuperación de la clase *tumor* corresponde a la sensibilidad del clasificador, mientras que la recuperación de la clase *no tumor* es equivalente a la especificidad.

	Precisión	Recuperación	F1-Score	Exactitud
No tumor	0.91	0.79	0.85	-
Sin reactividad	0.62	0.71	0.67	-
React. no lineal	0.25	0.37	0.3	-
React. lineal casi imperceptible	0.15	0.18	0.16	-
React. lineal débil	0.45	0.37	0.41	-
React. lineal fuerte	0.81	0.8	0.8	-
Promedio	0.74	0.7	0.72	0.7

Tabla 4.2.3: Resumen de resultados conseguidos en experimento *todo en uno*, evaluación de clasificación de parches de ROIs. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.

	Precisión	Recuperación	F1-Score	Exactitud
No tumor / Tumor ($T = 0{,}3$)				
No tumor	0.95	0.78	0.86	-
Tumor	0.77	0.95	0.85	-
Promedio	0.87	0.85	0.85	0.85
Reactividad HER2				
Sin reactividad	0.83	0.73	0.78	-
React. no lineal	0.34	0.39	0.36	-
React. lineal casi imperceptible	0.10	0.13	0.11	-
React. lineal débil	0.54	0.57	0.55	-
React. lineal fuerte	0.89	0.88	0.88	-
Promedio	0.65	0.61	0.63	0.61

Tabla 4.2.4: Resumen de resultados conseguidos en experimento *en cascada*, evaluación de clasificación de parches de ROIs. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.

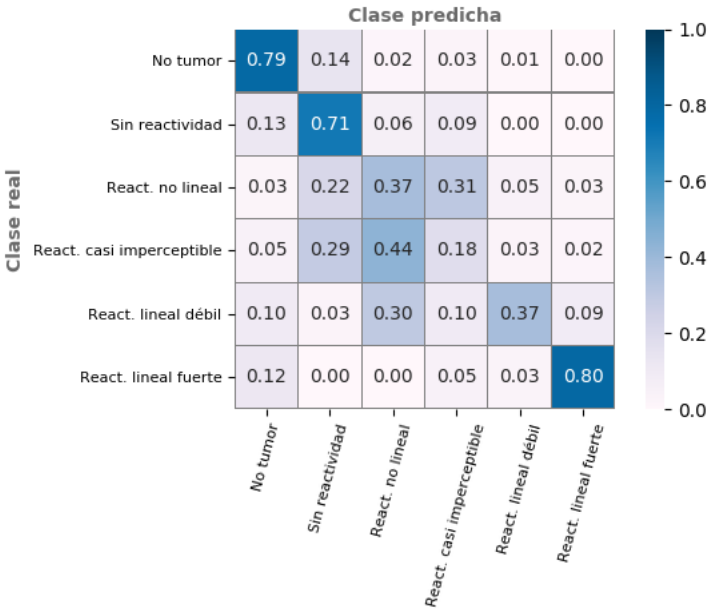
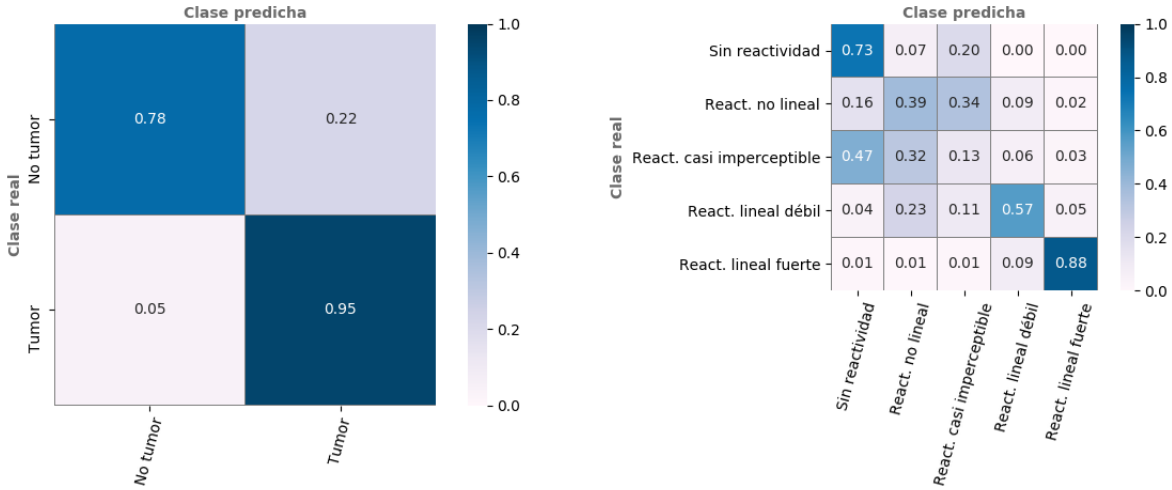


Figura 4.2.5: Matriz de confusión normalizada para experimento *todo en uno*, evaluación de clasificación de parches de ROIs.



(a) No tumor / Tumor (umbral $T = 0{,}3$) (b) Reactividad HER2

Figura 4.2.6: Matrices de confusión normalizadas para experimento *en cascada*, evaluación de clasificación de parches de ROIs.

4.2.2.2. Evaluación de clasificación de biopsias

En la sección previa se evaluó la clasificación de parches extraídos desde ROIs, donde cada parche tiene una etiqueta asociada (la misma etiqueta del ROI del cual fue extraída). Sin embargo, tal como se mencionó previamente, no es posible asumir que todos los parches

de una biopsia tienen la misma clasificación que dicha biopsia; de hecho, en general, tal supuesto no se cumple. Por ello, para realizar la evaluación de la clasificación de biopsias se utilizaron los modelos de ML ya entrenados para etiquetar cada parche, y se "confió" en dicha clasificación. Luego, se seleccionaron todos aquellos parches cancerosos y teniendo ya su clasificación HER2 correspondiente, se procedió a calcular la proporción de tejido correspondiente a cada clase (el tejido control presente en cada biopsia IHC no fue considerado para este análisis). Finalmente, con esos datos, se aplicó el algoritmo detallado en la guía clínica de clasificación de sobreexpresión de HER2 en biopsias IHC en cáncer gástrico, de modo acumulativo tal como fue señalado por los patólogos especialistas (algoritmo 4.2.1).

Algo importante de notar es que dicha guía clínica menciona que se debe cuantificar la cantidad de células, mientras que en el presente trabajo se cuantifica la cantidad de tejido. De esta forma, se asumió que todos los parches cancerosos poseen la misma densidad celular; incógnita, pero constante para todos los parches. Si bien es posible apreciar una diferencia en la cantidad de células presentes en zonas cancerosas versus zonas sanas, esto no influye en los resultados obtenidos, dado que las zonas no cancerosas no son consideradas al aplicar el algoritmo de las guías clínicas. De esta forma, se calcula el porcentaje de tejido de cada clase (0, 1+, 2+, 3+) respecto al tejido canceroso, no respecto al total de tejido.

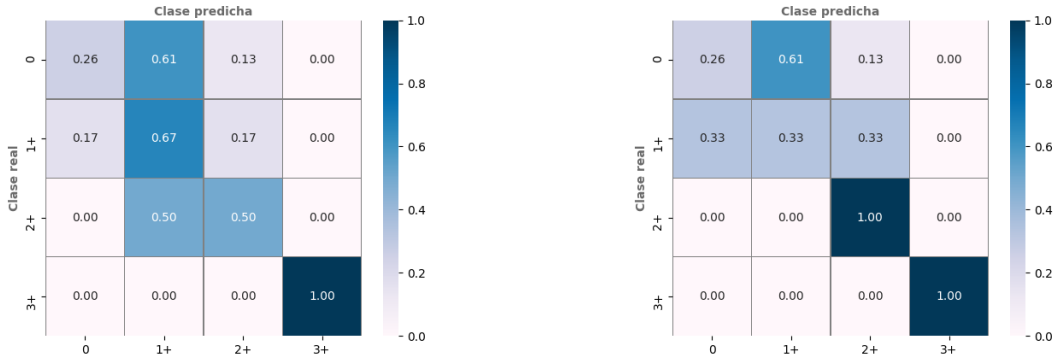
Por otra parte, es importante recordar que el proceso de evaluación de sobreexpresión de proteína HER2 es un proceso inherentemente subjetivo, por lo cual dos especialistas distintos pueden entregar diferentes evaluaciones para una misma biopsia. Si bien los resultados entregados por el patólogo 0 y el patólogo 3 son bastante similares, también tienen diferencias (principalmente discordancias en las clases 0 y 1+, además de dos muestras clasificadas como negativas por el patólogo 0, pero como equívocas por el patólogo 3). Por ello, se evaluó el desempeño de los modelos generados utilizando las evaluaciones de ambos patólogos como clasificación base. Es relevante destacar que todos los modelos fueron entrenados con anotaciones generadas por el patólogo 3, por lo cual es esperable que los modelos tengan mayor similitud con lo expresado por dicho especialista. La tabla 4.2.5 detalla los resultados obtenidos.

Del mismo modo, si bien las clases de sobreexpresión de biopsias son 0, 1+, 2+ y 3+, dado que las clases 0 y 1+ representan un tipo *Negativo*, estas etiquetas se pueden juntar. Así, también se puede analizar el desempeño de los algoritmos en un esquema simplificado, considerando solamente las clases *Negativo* (0 y 1+), *Equívoco* (2+) y *Positivo* (3+). La tabla 4.2.6 muestra los resultados usando estas clases.

Además de las métricas de las tablas 4.2.5 y 4.2.6, es posible evaluar la concordancia entre los modelos *todo en uno* y *en cascada*. Así, en el esquema de cuatro clases (0, 1+, 2+ y 3+) se obtiene un κ de Cohen igual a 0.58, mientras que en el esquema de tres clases (negativo, equívoco y positivo) el κ es de 0.72, lo que indica una concordancia considerable. Asimismo, es posible notar que la mejor concordancia obtenida experimentalmente es entre el patólogo 3 y el modelo *en cascada* ($\kappa = 0{,}61$, usando el esquema de tres clases), la cual también se ubica en el rango de concordancia considerable.

	Precisión	Recuperación	F1-Score	Exactitud	κ de Cohen
Todo en uno, con Patólogo 0 como clasificación base					
0	1.00	0.26	0.41	-	-
1+	0.16	0.75	0.26	-	-
2+	0.00	0.00	0.00	-	-
3+	1.00	1.00	1.00	-	-
Promedio	0.90	0.38	0.45	0.38	0.19
En cascada, con Patólogo 0 como clasificación base					
0	1.00	0.30	0.46	-	-
1+	0.12	0.50	0.20	-	-
2+	0.00	0.00	0.00	-	-
3+	1.00	1.00	1.00	-	-
Promedio	0.90	0.38	0.47	0.38	0.18
Todo en uno, con Patólogo 3 como clasificación base					
0	0.86	0.26	0.40	-	-
1+	0.21	0.67	0.32	-	-
2+	0.20	0.50	0.29	-	-
3+	1.00	1.00	1.00	-	-
Promedio	0.72	0.41	0.43	0.41	0.21
En cascada, con Patólogo 3 como clasificación base					
0	0.75	0.26	0.39	-	-
1+	0.12	0.33	0.18	-	-
2+	0.29	1.00	0.44	-	-
3+	1.00	1.00	1.00	-	-
Promedio	0.63	0.38	0.41	0.38	0.16

Tabla 4.2.5: Resumen de resultados para clasificación de biopsias, considerando clases 0, 1+, 2+ y 3+. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.



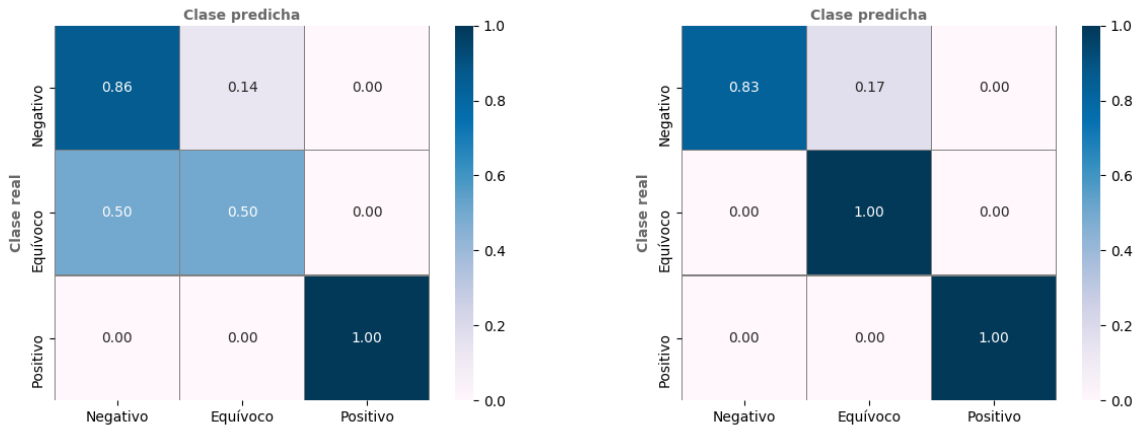
(a) *Todo en uno*.

(b) Cascada.

Figura 4.2.7: Matrices de confusión normalizadas para evaluación de clasificación de biopsias, usando clases 0, 1+, 2+ y 3+. En clasificación binaria de esquema *en cascada* se utilizó un umbral $T = 0{,}3$.

	Precisión	Recuperación	F1-Score	Exactitud	κ de Cohen
Todo en uno, con Patólogo 0 como clasificación base					
Negativo	1.00	0.84	0.91	-	-
Equívoco	0.00	0.00	0.00	-	-
Positivo	1.00	1.00	1.00	-	-
Promedio	1.00	0.85	0.92	0.85	0.50
En cascada, con Patólogo 0 como clasificación base					
Negativo	1.00	0.77	0.87	-	-
Equívoco	0.00	0.00	0.00	-	-
Positivo	1.00	1.00	1.00	-	-
Promedio	1.00	0.79	0.88	0.79	0.41
Todo en uno, con Patólogo 3 como clasificación base					
Negativo	0.96	0.86	0.91	-	-
Equívoco	0.20	0.50	0.29	-	-
Positivo	1.00	1.00	1.00	-	-
Promedio	0.92	0.85	0.88	0.85	0.55
En cascada, con Patólogo 3 como clasificación base					
Negativo	1.00	0.83	0.91	-	-
Equívoco	0.29	1.00	0.44	-	-
Positivo	1.00	1.00	1.00	-	-
Promedio	0.96	0.85	0.89	0.85	**0.61**

Tabla 4.2.6: Resumen de resultados para clasificación de biopsias, considerando clases *Negativo*, *Equívoco* y *Positivo*. En el caso de precisión, recuperación y F1-score, se calculó un promedio ponderado.



(a) *Todo en uno*.

(b) Cascada.

Figura 4.2.8: Matrices de confusión normalizadas para evaluación de clasificación de biopsias, usando clases *Negativo*, *Equívoco* y *Positivo*. En clasificación binaria de esquema *en cascada* se utilizó un umbral $T = 0{,}3$.

4.2.2.3. Visualizaciones

Con las predicciones realizadas por los modelos ya entrenados, fueron generadas visualizaciones de la clasificación HER2 en las clases *Negativo*, *Equívoco* y *Positivo*, además de marcar con un color distinto todas las zonas no cancerosas. Estas visualizaciones fueron construidas usando el lenguaje de programación `Python 3.6`, la librería de manipulación de imágenes `Pillow 6.1.0` y la librería de manipulación y análisis de datos `pandas 0.25.1`.

Si bien por ahora sólo se generan imágenes estáticas, y pese a que aún falta mucho para que pueda ser considerado como un prototipo propiamente tal, esto podría ser un primer paso para la construcción de un sistema de visualización de sobreexpresión HER2 para patólogos, de tal manera que ellos seleccionen una biopsia y visualicen de forma automática y simple qué zonas tienen sobreexpresión de esta proteína. Pese a que el presente trabajo no aborda el concepto de interpretabilidad en el sentido tradicional utilizado en ML (dado que no se busca explicar el motivo de la clasificación de cada parche en particular), este sistema de visualización sí busca aportar a la interpretabilidad desde otra perspectiva, permitiendo entender el porqué de la clasificación de una biopsia completa. Dado el contexto médico del proyecto, se estima que este sería un factor relevante en la adopción de un sistema de ML. De esta forma, las visualizaciones permitirían comprender por qué una biopsia fue clasificada de tal manera, pudiendo contrastar la clasificación de cada región con lo evaluado por el patólogo. Las figuras 4.2.9, 4.2.10 y 4.2.11 ejemplifican las visualizaciones generadas, junto a las biopsias originales respectivas.

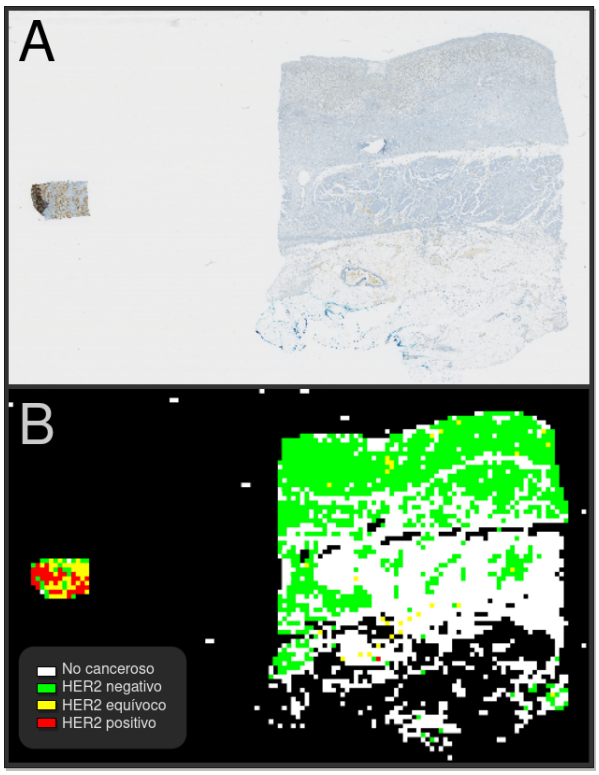


Figura 4.2.9: Biopsia con sobreexpresión HER2 negativa, correctamente clasificada. A) biopsia original, B) visualización generada utilizando la clasificación producida por el modelo *en cascada*.

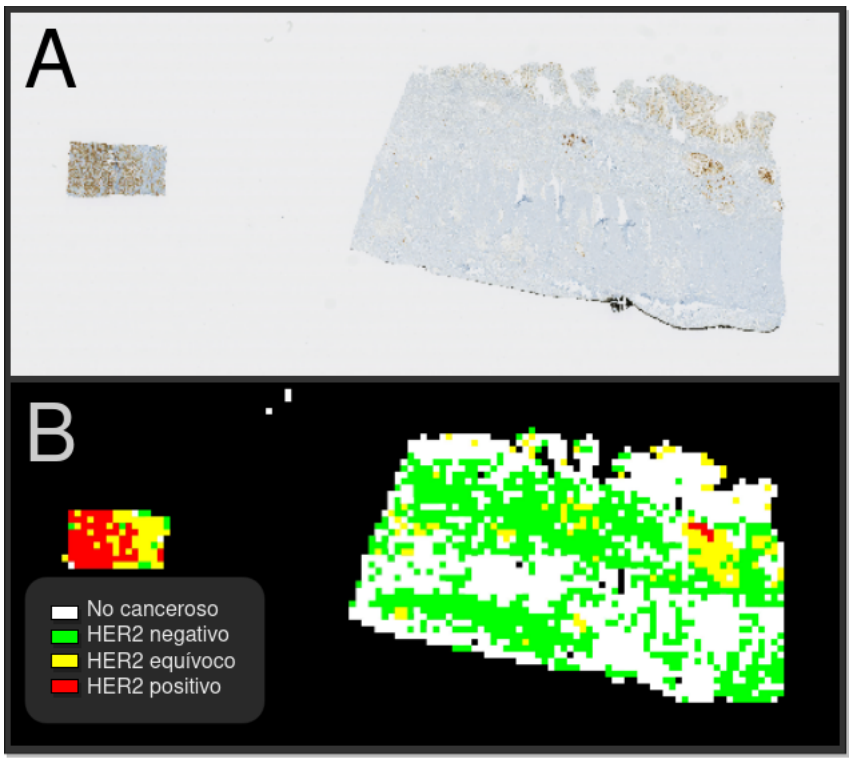


Figura 4.2.10: Biopsia con sobreexpresión HER2 equívoca, correctamente clasificada. A) biopsia original, B) visualización generada utilizando la clasificación producida por el modelo *en cascada*.

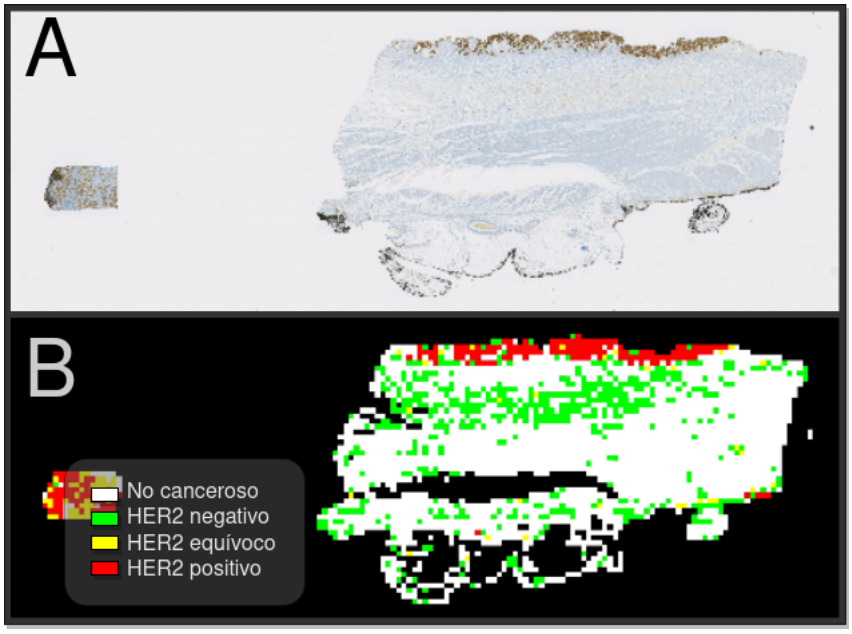


Figura 4.2.11: Biopsia con sobreexpresión HER2 positiva, correctamente clasificada. A) biopsia original, B) visualización generada utilizando la clasificación producida por el modelo *en cascada*.

Capítulo 5

Discusión

En la presenta sección, se discutirán los resultados obtenidos, sus implicancias y cómo estos se relacionan con el estado del arte en clasificación de HER2 en biopsias IHC utilizando ML. De especial relevancia son los resultados obtenidos en el macroexperimento II, dado que allí fue donde se generaron modelos que permiten la clasificación global de una biopsia, mientras el macroexperimento I fue principalmente entendido como una exploración del problema. Además, se discutirá brevemente sobre otros modelos de ML utilizados en la literatura para la clasificación de sobreexpresión HER2 en imágenes de biopsias IHC.

5.1. Comparación con otros modelos de ML

En el presente trabajo sólo se emplearon CNNs profundas. De esta forma, las redes fueron alimentadas directamente con imágenes en bruto, delegándole la tarea de extracción de características a las mismas CNNs. Si bien en un principio se planificó comparar el desempeño de descriptores especialmente diseñados (*hand crafted features*) y de otros modelos de ML, no fue posible realizar estas comparaciones. Esto, debido a la gran cantidad de trabajo que implicó el levantamiento de nuevos datos (todo el proceso, inesperado, de generar el etiquetado 2, debido a los problemas aparecidos en el etiquetado 1), además de todos los experimentos que sí fueron llevados a cabo (402 CNNs entrenadas entre los dos macroexperimentos).

Sin embargo, sí es posible analizar los resultados presentes en la literatura. En Vandenberghe y col. [89], en la tarea de clasificación de células, comparan el desempeño de redes neuronales convolucionales contra modelos clásicos de ML como son los *support vector machines* y los *random forests*. Así, comparan tres experimentos:

- SVM lineal con *hand crafted features*
- Random Forests con *hand crafted features*
- CNNs (extracción de características automática)

Los resultados de la comparación muestran que en la mayoría de las clases, la performance de las CNNs es superior a la de las otras configuraciones, mientras que en las clases restantes

no existen diferencias estadísticamente significativas.

Por otra parte, en Razavian y col. [70] recomiendan utilizar las características aprendidas por una CNN preentrenada como entradas para un SVM lineal, y utilizar esa configuración experimental como línea base. No obstante, los resultados del macroexperimento I sugieren que una CNN preentrenada con imágenes naturales no logra capturar de forma efectiva las propiedades de imágenes de biopsias, generando *features* con baja capacidad discriminatoria. Luego, esto sugiere que, con las imágenes empleadas en el presente trabajo, una configuración de SVM + *features* de una CNN preentrenada en ImageNet probablemente obtendrá peores resultados que utilizar *fine tuning* o reentrenamiento total sobre dicha CNN. Aún así, estas hipótesis deben ser validadas experimentalmente.

5.2. Macroexperimento I

Tras las dos fases del macroexperimento I (selección de hiperparámetros y evaluación de modelos), un hecho que llama la atención es que en las magnificaciones x10 y x40, los resultados obtenidos con la configuración experimental simple (Inception v3 + capa final) son peores que aquellos obtenidos con las configuraciones de *fine tuning* y reentrenamiento total. Además, al realizar la selección de hiperparámetros, fue posible apreciar que en los subexperimentos con configuración simple, el número óptimo de épocas de entrenamiento tiende a ser bajo (promedio 10x: $15{,}6 \pm 12{,}1$, 20x: $12{,}8 \pm 15{,}2$, promedio 40x: $20{,}6 \pm 19{,}2$) y que, en general, en estos subexperimentos los entrenamientos no son estables y no convergen, tal como lo ejemplifica la figura 5.2.1. Por otro lado, en los subexperimentos con *fine tuning* o reentrenamiento total, el número de épocas óptimo suele ser bastante más alto. Además, al revisar los gráficos generados, estos experimentos sí parecen converger o aumentar constantemente la exactitud, como se ve en las figuras 5.2.2 y 5.2.3.

Todo esto parece indicar que usar transferencia de aprendizaje directamente, sin aplicar *fine tuning* en parte alguna, desde un modelo entrenado en imágenes naturales como *ImageNet* a un dominio muy distinto como es el de imágenes celulares no produce buenos resultados. Esto se condice con lo planteado en Yosinski y col. [91], donde se muestra que la transferibilidad de las características decrece a medida que aumenta la distancia entre el dominio original (donde fue entrenado el modelo base) y el dominio nuevo. Del mismo modo, en los apuntes de curso de *Redes Neuronales Convolucionales para Reconocimiento Visual* impartido por la Universidad Stanford, se señala que las primeras capas de una red convolucional parecen aprender características genéricas, y que estas *features* se van especializando a medida que se profundiza la red Karpathy [44]. De esta forma, se explicaría el porqué de que las configuraciones sin *fine tuning* entregaran estos resultados, ya que las capas más profundas de la red están adaptadas para dominios muy distintos de aquel trabajado en este proyecto.

Otro hecho interesante es que los resultados obtenidos con la magnificación 10x fueron mejores que en 20x, y estos a su vez fueron mejores que los resultados en 40x. Esto parece indicar que el supuesto de que todos los parches extraídos desde un ROI corresponden a la misma clase, en general no se cumple, en particular para las clases 2+ y 3+. Dado que los parches tienen un tamaño fijo de 300x300 pixeles, los parches a magnificaciones más bajas

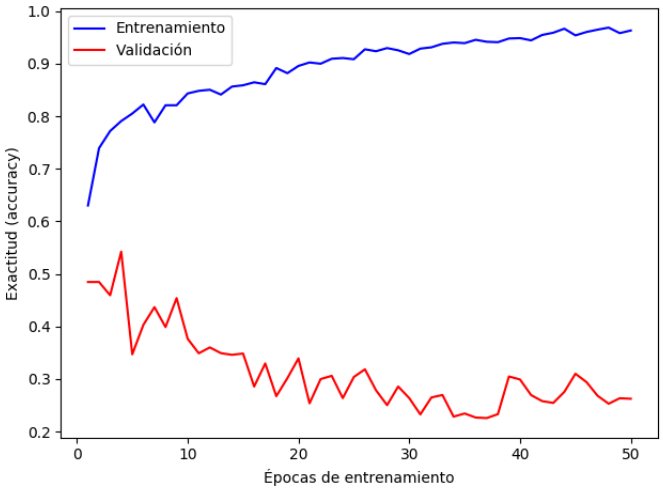


Figura 5.2.1: Subexperimento de selección de parámetros simple utilizando parches a magnificación 40x. Entrenamiento con subconjuntos 2, 3 y 5, y validación con conjunto 1.

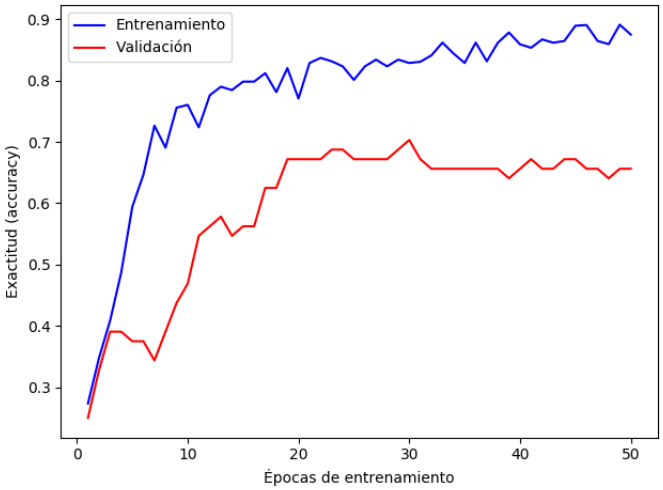


Figura 5.2.2: Subexperimento de selección de parámetros con *fine tuning* y *data augmentation* utilizando parches a magnificación 10x. Entrenamiento con subconjuntos 1, 4 y 5, y validación con conjunto 2.

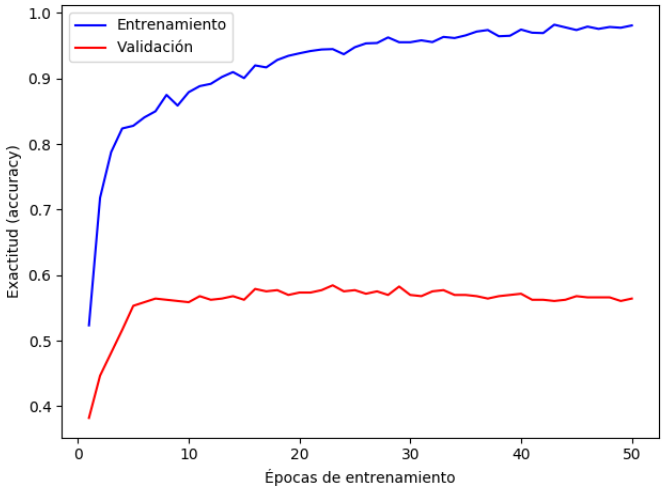


Figura 5.2.3: Subexperimento de selección de parámetros con reentrenamiento total y *data augmentation* utilizando parches a magnificación 20x. Entrenamiento con subconjuntos 1, 3 y 4 y validación con conjunto 5.

abarcan una mayor cantidad de tejido físico, pero a menor resolución. Esto aleja el presente trabajo de lo hecho por Vandenberghe y col. [89], donde la entrada para la red corresponde a la imagen de una célula. Sin embargo, en dicho trabajo, las anotaciones fueron llevadas a cabo de forma detallada, teniendo cuidado de que cada región fuese lo más homogénea posible, lo cual difiere considerablemente de las anotaciones de ROIs utilizadas en este macroexperimento. Por ello, debido a la mayor heterogeneidad que presentan estos ROIs, es que los parches extraídos a magnificaciones más bajas parecen ser más robustos.

De todas formas, estas apreciaciones deben ser tomadas con cuidado, dado que las anotaciones utilizadas en este macroexperimento tienen graves problemas clínicos, lo cual suma un factor de impredictibilidad. Por ello, para confirmar estas apreciaciones, sería necesario reejecutar estos experimentos con una base de datos óptima, como la utilizada en el macroexperimento II.

5.3. Macroexperimento II

Respecto a la evaluación de parches extraídos directamente desde ROIs, es posible apreciar que:

- En ambos experimentos, las clases *No tumor* y *Reactividad lineal fuerte* son las más "fáciles" de identificar (F1-score $\geq 0{,}8$ en modelos *todo en uno* y *en cascada*). Además, la clase *Sin reactividad* también obtiene buenos resultados (F1-score *todo en uno* $= 0.67$, *en cascada* $= 0.78$). Esto tiene sentido, si se considera que una de estas clases representa ausencia de tejido canceroso, y las otras corresponden a sobreexpresión HER2 0 y 3+, las cuales se encuentran en extremos opuestos de la clasificación.
- Por el contrario, las clases *Reactividad no lineal*, *Reactividad lineal casi imperceptible*

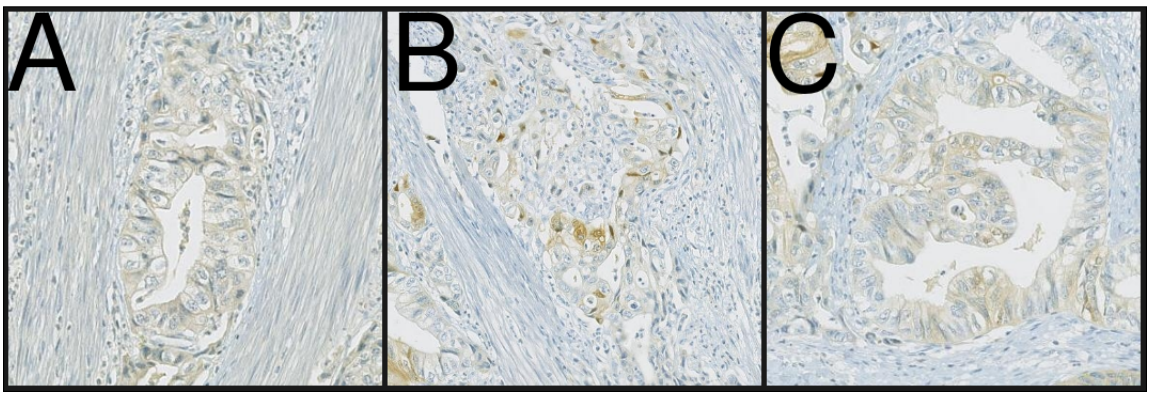


Figura 5.3.1: Ejemplos de ROIs pertenecientes a clases que las redes suelen confundir. A) Reactividad no lineal, B) Reactividad lineal casi imperceptible, C) Reactividad lineal débil. Todos los ROIs provienen de la misma biopsia.

y *Reactividad lineal débil* generan confusión en los modelos entrenados. Estas clases corresponden a sobreexpresión 0 (no lineal), 1+ (lineal casi imperceptible) y 2+ (lineal débil). Además, en el modelo *en cascada*, la clase *Reactividad lineal casi imperceptible* es confundida frecuentemente con la clase *sin reactividad* (correspondiente a sobreexpresión 0). De todas formas, y tal como se aprecia en la figura 5.3.1, distinguir entre estas clases puede ser complicado incluso para el ojo humano.

Luego, respecto a la clasificación binaria de *tumor / no tumor* perteneciente al esquema *en cascada*, existen varios aspectos que deben ser destacados. En primer lugar, es necesario discutir sobre los costos asociados a los errores de clasificación. Como ya se mencionó previamente, dado el contexto médico del presente trabajo, el costo de un falso negativo es mayor al de un falso positivo; clasificar un parche canceroso como sano es en teoría más grave que identificar un parche sano como canceroso. Sin embargo, existe una sutileza: en este proyecto lo que se busca es clasificar la sobreexpresión de HER2 en biopsias, y la clasificación de parches en *tumor / no tumor* es sólo un medio para conseguir tal fin. Es más, es importante considerar que todas las muestras utilizadas en este proyecto sí contienen tejido canceroso; ninguna de las biopsias corresponde a tejido completamente sano.

Así, tomando en cuenta lo anterior, es necesario recordar que el proceso de diagnóstico de cáncer en tejido de cáncer gástrico se realiza utilizando biopsias con tinción H&E. De esta forma, para construir un clasificador que sólo detecte si un tejido es canceroso o no, lo óptimo sería utilizar una base de datos de biopsias H&E. Aún así, el clasificador *tumor / no tumor* construido en el presente trabajo sí tiene un valor por sí mismo: dado que este modelo clasifica parches con tinción IHC, y considerando que la sobreexpresión de HER2 se debe realizar sólo en zonas cancerosas de la biopsia, este sistema podría sugerir a un patólogo regiones de interés para realizar la evaluación HER2.

Ahora bien, siguiendo con la clasificación binaria, resaltan algunos aspectos técnicos. El clasificador entrenado logra separar de manera efectiva las clases en cuestión, tal como se aprecia en los histogramas de la figura 4.2.4a y es refrendado por el $AUC = 0{,}95$ obtenido. Específicamente, un 73 % de los parches de clase *no tumor* obtuvieron una salida $\leq 0{,}2$, mientras que un 81 % de los parches de la clase *tumor* tuvieron una salida $\geq 0{,}8$. Además,

si se utiliza un umbral de decisión $T = 0{,}5$ (el predeterminado o *default*), se obtiene una sensibilidad de 91 % y una especificidad de 85 %. De todas maneras, tomando en cuenta todas las consideraciones mencionadas, se decidió ajustar el umbral de decisión para que el clasificador *tumor / no tumor* tuviera una sensibilidad muy alta, aún cuando esto fuera en desmedro de la especificidad del sistema. Por ello, se seleccionó el umbral de decisión $T = 0{,}3$, que obtiene una sensibilidad de 95 %. No obstante esto, resulta interesante notar que si se realiza la clasificación de biopsias utilizando el umbral $T = 0{,}5$, los resultados obtenidos son prácticamente iguales a los obtenidos utilizando el umbral $T = 0{,}3$ reportado en el presente trabajo. En dicha situación, sólo se ven afectadas dos biopsias negativas: una pasa de ser 0 a 1+, y otra cambia de 1+ a 0.

Por otro lado, respecto a la evaluación global de biopsias también es posible notar varios hechos. Lo primero es que los distintos modelos obtienen, en general, resultados que se acercan más a la clasificación realizada por el patólogo 3, en desmedro del patólogo 0. Esto era esperable, dado que fue este especialista quien realizó las anotaciones utilizadas en el macroexperimento II.

En segundo lugar, todos los modelos entrenados confunden la clase 0 con la clase 1+, siendo esta la principal fuente de errores. Sin embargo, cuando ambas clases se juntan en la etiqueta *Negativo*, se obtienen mejores resultados; usando al patólogo 3 como base, el F1-score del modelo *todo en uno* es de 0.88, mientras el del modelo *en cascada* es 0.89. Además, el modelo *en cascada* obtiene una recuperación de 100 % para las clases 2+ y 3+, tal como se aprecia en la tabla 4.2.6. De esta forma, el modelo *en cascada* es el que obtiene los mejores resultados, y por ello, el resto de la discusión estará centrada en ese modelo.

Algo importante de recordar es que la concordancia entre los patólogos 0 y 3 es de $\kappa = 0{,}73$ en el esquema de tres clases, está en el rango de concordancia considerable, según la tabla 2.7.1 recomendada por Landis y Koch [52]. De esta forma, tanto la mejor concordancia entre patólogos como el mejor modelo generado en el presenta trabajo presentan concordancia considerable. Además, en Kushima y col. [51] se muestra que la concordancia entre patólogos sin entrenamiento específico en HER2 en cáncer gástrico alcanza un $\kappa = 0{,}68$, el cual se eleva a 0.79 tras pasar por un programa de entrenamiento. De todas formas, una de las críticas principales al coeficiente κ de Cohen es que asume que dos observadores pueden acertar por azar, siendo esto no necesariamente cierto y lo cual en la práctica puede disminuir excesivamente el grado de concordancia [58].

Por ello, es necesario hacer un análisis un tanto más cualitativo de los resultados obtenidos. Así, en el modelo *en cascada*, tomando la clasificación del patólogo 3 como base y utilizando el esquema simplificado de tres clases, sólo 5 biopsias fueron mal clasificadas. En particular, todas estas biopsias fueron catalogadas como *equívocas*, cuando su clase real era *negativo*. Desde un punto de vista clínico, si bien esto no es óptimo, tampoco es la peor alternativa. Esto, porque una biopsia mal clasificada como equívoca deberá ser sometida a un examen FISH para determinar si es que existe o no amplificación del gen HER2, y finalmente sí se podrá determinar su clasificación real. Por otro lado, una biopsia mal clasificada como negativa implica que no habrá acceso a tratamiento con Trastuzumab, lo cual es perjudicial para el paciente en cuestión. Además, desde una perspectiva algorítmica, resulta interesante analizar la predicción realizada por las CNNs para los parches de dichas biopsias mal clasificadas. En

particular, es posible revisar la cantidad de tejido etiquetado como 2+ y 3+, y su porcentaje del total de tejido canceroso. Para estas cinco biopsias, los porcentajes de tejido clasificados como 2+ o 3+ son 16,5 %, 13,1 %, 20,1 %, 17,9 % y 10.6 % (promedio 15,6 % ± 3,4 %). Además de la opción obvia de errores en la clasificación de parches, y debido a lo cercano que son los valores previamente mencionados al umbral del 10 %, no es posible descartar un error en la evaluación del patólogo 3. Esto cobra especial relevancia en aquellos casos donde la diferencia entre lo calculado por el modelo y lo estimado por el especialista es menor a un 5 %.

Siguiendo con el análisis cualitativo, aún con el modelo *en cascada* pero retomando el esquema de cuatro clases, existen 14 biopsias clasificadas como 1+ cuando debiesen ser 0, y dos biopsias clasificadas como 0 cuando debiesen ser 1+. Si bien ambas clases son de tipo *negativo*, al analizar en mayor detalle las clasificaciones de los parches de dichas biopsias es posible notar que, en general, los resultados obtenidos exceden por mucho el umbral del 10 %. Así, en los catorce casos mal clasificados como 1+, el promedio de la suma de los tejidos clasificados como 1+, 2+ y 3+ es de 30,1 % ± 12,0 %. Por otro lado, en los casos de las biopsias mal clasificadas como 0, las sumas de los tejidos de tipo 1+, 2+ y 3+ corresponden al 9,2 % y 8,1 % del tejido canceroso, valores que sí son cercanos al umbral del 10 %. De todas maneras, estos errores cobran sentido al recordar lo visto en la evaluación de clasificación de parches en base a su reactividad, donde la clase *reactividad lineal casi imperceptible* (que implica sobreexpresión 1+) es frecuentemente confundida con la clase *reactividad no lineal* (sobreexpresión 0) (figura 4.2.6b).

También se debe destacar que, utilizando regresión u otra técnica similar, tal vez sería posible hallar umbrales de corte distintos al 10 % que permitieran una mejor separación entre las clases 0, 1+, 2+ y 3+ y obtuvieran resultados aún más cercanos a lo evaluado por el patólogo especialista. Sin embargo, esta opción fue descartada y se prefirió mantener la regla del 10 %, en pos de ceñirse a las guías clínicas y permitir una mayor interpretabilidad para los médicos que potencialmente usarán este sistema.

5.3.1. Estado del arte y limitaciones

Como se mencionó en la sección 2.8, la mayor parte de la investigación sobre clasificación de HER2 utilizando ML se ha llevado a cabo en cáncer de mama. Por ejemplo, en Vandenberghe y col. [89], se utiliza ML para clasificar células y luego, aplicando las reglas de la guía clínica correspondiente, calculas la clasificación HER2 global de una biopsia. Así, en dicho trabajo (y utilizando también el esquema de clases *Negativo*, *Equívoco* y *Positivo*), la concordancia con la evaluación de los patólogos es de un 83 %, con un κ de Cohen igual a 0.69. Por su parte, en Saha y Chakraborty [78] también trabajan con cáncer de mama y reportan los porcentajes de concordancia desagregados por clase (0: 100 %, 1+: 98.13 %, 2+: 97.98 %, 3+: 99.97 %).

Respecto a cáncer gástrico, en Nielsen, Nielsen y Vyberg [65] ajustan los parámetros del software *HER2CONNECT* para clasificar biopsias IHC de cáncer gástrico, reportando sensibilidad y especificidad del 100 %. No obstante, estos valores se obtienen cuando utilizan como clasificación base los resultados del examen FISH, el cual sólo produce una salida positiva o negativa. De esta forma, cuando se compara lo obtenido en dicho trabajo con lo evaluado por patólogos mediante IHC, sí existe discrepancia en la clase 2+. Por su parte, en

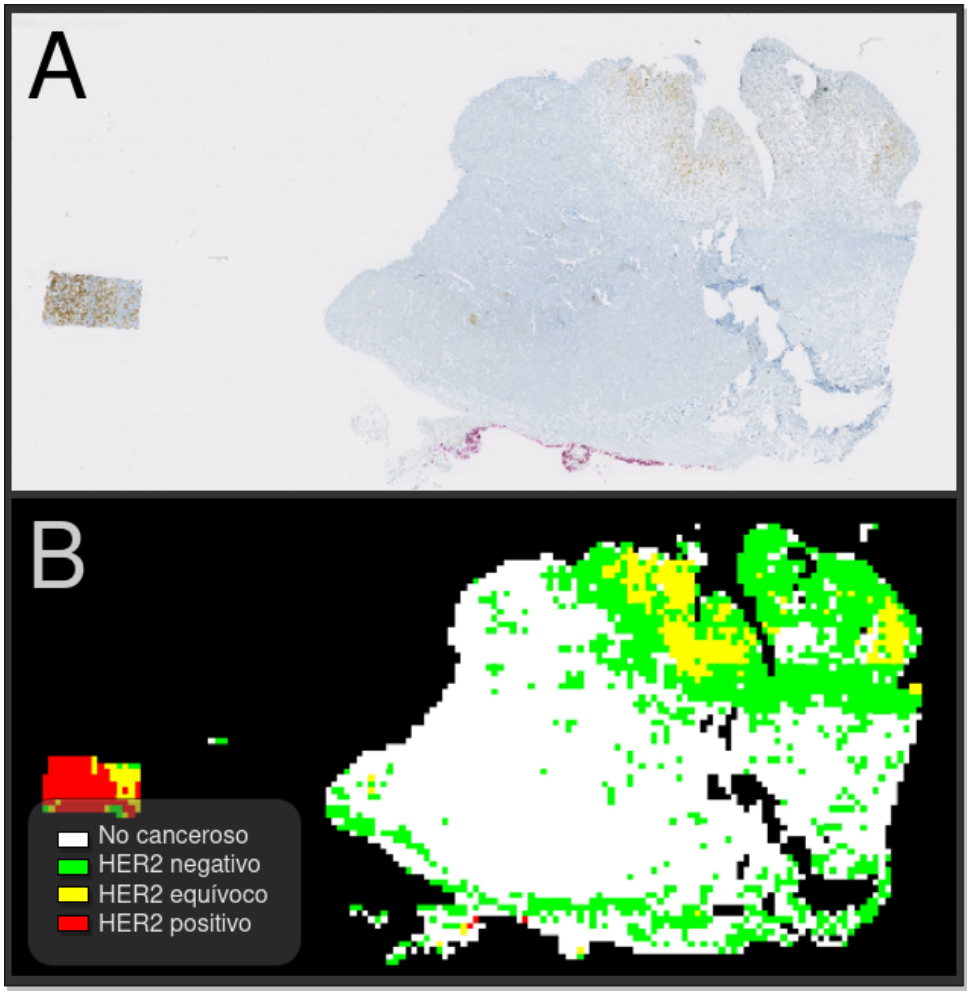


Figura 5.3.2: Biopsia con sobreexpresión HER2 negativa, incorrectamente clasificada como equívoca. Tejido de tipo 2+ o 3+ calculado por algoritmo corresponde a 16,5 % del tejido canceroso. A) biopsia original, B) visualización generada utilizando la clasificación producida por el modelo *en cascada*.

Sharma y col. [81] no clasifican biopsias de forma global.

De todas formas, es necesario tener en cuenta las limitaciones del presente trabajo. Algunas de ellas son:

- Todo el macroexperimento II fue llevado a cabo utilizando las anotaciones de un sólo patólogo. Si bien la concordancia entre este patólogo y el especialista original del estudio *PRECISO* es alta, sí existe un sesgo hacia la clasificación realizada por el patólogo 3, tal como se muestra en las tablas 4.2.5 y 4.2.6.
- El método propuesto no es invariante a cambios en la forma de adquisición de las imágenes. En particular, tanto el kit de tinción utilizado para generar la reacción en biopsias IHC como el escáner empleado para digitalizar los portaobjetos pueden generar diferencias en las imágenes finales, por lo cual es necesario estudiar el comportamiento del sistema generado bajo otras condiciones.
- El modelo construido solamente clasifica la sobreexpresión HER2 en biopsias por resección. Para biopsias endoscópicas, las reglas de evaluación son distintas y se basan en

la presencia de *clusters* de al menos 5 células. Los experimentos del macroexperimento II fueron todos llevados a cabo con magnificación 10x, donde cada parche tiene, en general, mucho más de 5 células. Por ello, adaptar lo construido para que funcione con biopsias endoscópicas resulta difícil y requeriría estrategias distintas.

- La configuración escogida para llevar a cabo el macroexperimento II fue la que obtuvo mejores resultados en el macroexperimento I (reentrenamiento total con magnificación 10x). No obstante, dados los problemas presentes en los datos utilizados en el primer macroexperimento (anotaciones de patólogos 1 y 2), es posible argumentar que dicha configuración no necesariamente es la que mejor resultados obtiene en el macroexperimento II. Así, dicha hipótesis se debe validar, reejecutando el último macroexperimento con varias configuraciones distintas, con lo cual mejore resultados podrían ser conseguidos.

Capítulo 6

Conclusiones

Los resultados más relevantes de este trabajo son aquellos relacionados a la clasificación global de biopsias IHC de cáncer gástrico de acuerdo a su sobreexpresión HER2. En particular, uno de los modelos entrenados consigue recuperar correctamente todas las biopsias de tipo *Equívoco* y *Positivo*, aquellas más relevantes en la práctica clínica, ya que una de ellas (Positivo) permite que al paciente se le administre un tratamiento que permite alargar la sobrevida, mientras que la otra categoría (Equívoco) indica que se debe realizar otro examen para confirmar o descartar el posible tratamiento. De esta forma, y siguiendo el análisis cuantitativo y cualitativo discutido en la sección 5.3, se cumple la hipótesis planteada en este trabajo, de que es posible generar posible generar un sistema automatizado de procesamiento de imágenes para la clasificación de sobreexpresión de la proteína HER2 en muestras de cáncer gástrico con tinción inmunohistoquímica y que obtenga resultados comparables a los entregados por patólogos. Esto, porque la concordancia lograda en este trabajo con respecto al patólogo 3 (aquel que evaluó las biopsias y anotó las imágenes) es considerable, y además se logra recuperar correctamente aquellas clases importantes para la práctica clínica. No obstante, algunas consideraciones que deben ser tomadas en cuenta son el hecho de que todo el entrenamiento fue realizado usando un dataset formado por anotaciones realizadas por sólo un patólogo y que además se está trabajando sólo con biopsias por resección, dejando de lado las biopsias endoscópicas.

Además, los modelos construidos cumplen con los objetivos específicos, a saber:

- **Replicar el proceso diagnóstico llevado a cabo por los médicos especialistas**. Esto, porque sólo se utilizó ML para clasificar parches, mientras que la evaluación global de una biopsia fue realizada siguiendo la regla del 10 % estipulada para biopsias por resección y recomendada en las guías clínicas [10].
- **Ser interpretable por dichos médicos.** Esto, porque con el sistema de visualización desarrollado, un patólogo puede fácilmente identificar la clasificación otorgada por el modelo a cada zona, y decidir en base a su propia experiencia si es que la clasificación fue correcta o no.
- **Obtener resultados comparables a los entregados por los patólogos.** Esto se cumple, ya que la concordancia obtenida por el mejor modelo es considerable respecto a la evaluación realizada por el patólogo, y a su vez, está en el mismo rango de con-

cordancia que el acuerdo entre patólogos (tanto el observado en este trabajo como el reportado en la literatura).

El enfoque utilizado en el presente trabajo, de replicación del proceso diagnóstico utilizando los algoritmos de evaluación de las guías clínicas, es similar a lo realizado en cáncer de mama por otros autores. Y si bien los resultados aquí obtenidos son comparables a lo logrado en uno de dichos trabajos [89], aún así distan mucho de lo logrado en otro de los artículos, donde se logra una concordancia casi perfecta con lo evaluado por patólogos [78]. Ahora bien, respecto a cáncer gástrico, sólo un artículo aborda la clasificación global de biopsias, y la concordancia allí es medida utilizando como base el examen FISH, además de no seguir un enfoque de replicación del proceso diagnóstico [65]. Por ende, los resultados del presente trabajo, donde la interpretabilidad y el apego a las guías clínicas fueron ejes importantes, no son directamente comparables con dicho estudio.

Por otro lado, se debe tener en cuenta que algunas conclusiones sacadas del primer macroexperimento deben ser validadas, como el hecho de que una red preentrenada en un conjunto de imágenes naturales no es óptima *a priori* para clasificar imágenes de biopsias, y que debe ser sometida a *fine tuning* o reentrenamiento total.

6.1. Trabajo futuro

En primera instancia, para construir un modelo más robusto, es necesario contar con más anotaciones, provenientes de distintos patólogos. Con ello, se reducirá el sesgo que tienen los modelos hacia la evaluación realizada por un sólo patólogo, lo que en teoría permitiría una mejor generalización del modelo y mejores resultados.

Otras ideas que resultan interesantes de evaluar son:

- Experimentar con arquitecturas de redes neuronales convolucionales que puedan aprovechar la estructura piramidal de las imágenes. En particular, sería de especial interés poder entrenar con parches extraídos a varias magnificaciones distintas.
- Utilizar técnicas de procesamiento de imágenes para detectar la cantidad de células presentes en cada parche, en lugar de asumir una densidad celular constante en cada parche canceroso, como se hace ahora. Esto podría significar una evaluación global HER2 mucho más exacta.
- En el presente trabajo, se utilizaron biopsias IHC para detectar presencia de cáncer o no. Algo con lo que se podría experimentar es detectar cáncer en biopsias H&E (tinción generalmente utilizada para detectar zonas cancerosas), y luego esa detección trasladarla a biopsias IHC. Una idea similar tiene precedentes en Sharma y col. [81], donde la sobreexpresión HER2 anotada en una biopsia IHC es trasladada a su correspondiente biopsia pareada con tinción H&E.
- Empleando un nuevo etiquetado, se podrían utilizar redes que permitan la segmentación de imágenes, reconociendo los bordes de membranas y núcleos celulares. Alternativas de redes para esta tarea son *Her2Net*, construida por Saha y Chakraborty [78], o *U-Net*, especialmente diseñada para segmentación de imágenes biomédicas [72]. Este enfoque

podría entregar mejores resultados experimentales y además brindar mayor interpretabilidad, ya que un patólogo podría observar en detalle cómo fue clasificada cada célula.

Finalmente, en términos más prácticos, y recordando que el fin último es proveer una herramienta de apoyo a la práctica clínica, este sistema debe ser retroalimentado por la experiencia de patólogos especialistas. Así, luego de robustecer el modelo con datos provenientes de distintos especialistas, se podría facilitar un prototipo del sistema a algunos patólogos, de tal forma que ellos pudieran evaluar el desempeño del modelo y así entregar sus recomendaciones.

Bibliografía

[1] Martín Abadi y col. «TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems». En: *arXiv:1603.04467 [cs.DC]* (2015). Software available from tensorflow.org. URL: http://arxiv.org/abs/1603.04467.

[2] Yaser S. Abu-Mostafa, Malik Magdon-Ismail y Hsuan-Tien Lin. *Learning from data: a short course*. 1ra ed. S.l.: AMLbook, 2012. 201 págs. ISBN: 978-1-60049-006-4.

[3] Omar B. Ahmad y col. *Age standardization of rates: a new WHO standard*. GPE Discussion Paper Series 31. Geneva: World Health Organization, 2001.

[4] Jaffer A. Ajani y col. *NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines)*. versión 5.2017. National Comprehensive Cancer Network, Inc, 2017. URL: https://www.nccn.org.

[5] Luca Albarello, Lorenza Pecciarini y Claudio Doglioni. «HER2 Testing in Gastric Cancer». En: *Advances in Anatomic Pathology* 18.1 (ene. de 2011), págs. 53-59. DOI: 10.1097/PAP.0b013e3182026d72. URL: https://insights.ovid.com/crossref?an=00125480-201101000-00004 (visitado 25-07-2019).

[6] American Cancer Society. *¿Qué es el cáncer de estómago?* 2017. URL: https://www.cancer.org/es/cancer/cancer-de-estomago/acerca/que-es-el-cancer-de-estomago.html (visitado 26-06-2019).

[7] Mahul B. Amin y col., eds. *AJCC Cancer Staging Manual*. 8va ed. Chicago IL: Springer, 2018. 1032 págs. ISBN: 978-3-319-40617-6.

[8] *Applications - Keras Documentation*. URL: https://keras.io/applications/ (visitado 31-10-2019).

[9] Yung-Jue Bang y col. «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». En: *The Lancet* 376.9742 (2010), págs. 687-697. DOI: 10.1016/S0140-6736(10)61121-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S014067361061121X (visitado 28-03-2019).

[10] Angela N. Bartley y col. «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». En: *Journal of Clinical Oncology* 35.4 (2017), págs. 446-464. DOI: 10.1200/JCO.2016.69.4836. URL: http://ascopubs.org/doi/10.1200/JCO.2016.69.4836 (visitado 18-07-2019).

[11] Hans-Michael Behrens y col. «Reproducibility of Her2/neu scoring in gastric cancer and assessment of the 10 % cut-off rule». En: *Cancer Medicine* 4.2 (2015), págs. 235-244.

DOI: 10.1002/cam4.365. URL: http://doi.wiley.com/10.1002/cam4.365 (visitado 28-03-2019).

[12] Fred T. Bosman y col., eds. *WHO classification of tumours of the digestive system*. 4ta ed. Vol. 3. World Health Organization classification of tumours. Lyon, France: International Agency for Research on Cancer, 2010. 417 págs. ISBN: 978-92-832-2432-7.

[13] Freddie Bray y col. «Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: A Cancer Journal for Clinicians* 68.6 (2018), págs. 394-424. DOI: 10.3322/caac.21492. URL: http://doi.wiley.com/10.3322/caac.21492 (visitado 17-07-2019).

[14] Jason Brownlee. *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*. Machine Learning Mastery. 30 de ago. de 2018. URL: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ (visitado 22-02-2020).

[15] Anja Brügmann y col. «Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains». En: *Breast Cancer Research and Treatment* 132.1 (2012), págs. 41-49. DOI: 10.1007/s10549-011-1514-2. URL: http://link.springer.com/10.1007/s10549-011-1514-2 (visitado 28-03-2019).

[16] *Caffe | Model Zoo*. URL: http://caffe.berkeleyvision.org/model_zoo.html (visitado 02-09-2019).

[17] François Chollet. *Keras*. 2015. URL: https://keras.io.

[18] Yosep Chong y col. «DNA methylation status of a distinctively different subset of genes is associated with each histologic Lauren classification subtype in early gastric carcinogenesis». En: *Oncology Reports* 31.6 (2014), págs. 2535-2544. DOI: 10.3892/or.2014.3133. URL: https://www.spandidos-publications.com/10.3892/or.2014.3133 (visitado 23-07-2019).

[19] Corinna Cortes y Vladimir Vapnik. «Support-vector networks». En: *Machine learning* 20.3 (1995), págs. 273-297.

[20] Dako. *HercepTest™ Interpretation Manual - Gastric Cancer*. Glostrup, Dinamarca, 2010. URL: https://www.agilent.com/cs/library/usermanuals/public/29018_05may10_herceptest_interpretation_manual_gastric_cancer.pdf (visitado 23-07-2019).

[21] Shaveta Dargan y col. «A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning». En: *Archives of Computational Methods in Engineering* (2019), págs. 1-22. DOI: 10.1007/s11831-019-09344-w. URL: https://doi.org/10.1007/s11831-019-09344-w (visitado 07-10-2019).

[22] Li Deng y Dong Yu. «Deep learning: methods and applications». En: *Foundations and Trends® in Signal Processing* 7.3 (2014), págs. 197-387.

[23] Rahul C. Deo. «Machine Learning in Medicine». En: *Circulation* 132.20 (2015), págs. 1920-1930. DOI: 10.1161/CIRCULATIONAHA.115.001593. URL: https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.115.001593 (visitado 13-08-2019).

[24] Christophe Deroulers y col. «Analyzing huge pathology images with open source software». En: *Diagnostic Pathology* 8.1 (2013), pág. 92. DOI: 10.1186/1746-1596-8-92. URL: https://diagnosticpathology.biomedcentral.com/articles/10.1186/1746-1596-8-92 (visitado 28-10-2019).

[25] Arden Dertat. *Applied Deep Learning - Part 1: Artificial Neural Networks*. Medium. 9 de oct. de 2017. URL: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6 (visitado 08-10-2019).

[26] Rachel Draelos. *A Short History of Convolutional Neural Networks*. Glass Box Medicine. 13 de abr. de 2019. URL: https://glassboxmedicine.com/2019/04/13/a-short-history-of-convolutional-neural-networks/ (visitado 30-08-2019).

[27] Tom Fawcett. «An introduction to ROC analysis». En: *Pattern Recognition Letters*. ROC Analysis in Pattern Recognition 27.8 (1 de jun. de 2006), págs. 861-874. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2005.10.010. URL: http://www.sciencedirect.com/science/article/pii/S016786550500303X (visitado 22-02-2020).

[28] Jacques Ferlay y col. *Global Cancer Observatory: Cancer Today*. 2018. URL: https://gco.iarc.fr/today (visitado 09-08-2019).

[29] George Forman y Martin Scholz. «Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement». En: *ACM SIGKDD Explorations Newsletter* 12.1 (2010), págs. 49-57. DOI: 10.1145/1882471.1882479. URL: http://doi.acm.org/10.1145/1882471.1882479 (visitado 24-11-2019).

[30] Kunihiko Fukushima. «Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position». En: *Biological Cybernetics* 36.4 (1980), págs. 193-202. DOI: 10.1007/BF00344251. URL: https://doi.org/10.1007/BF00344251 (visitado 30-08-2019).

[31] Rohith Gandhi. *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. 5 de jul. de 2018. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (visitado 04-12-2019).

[32] Rafael C. Gonzalez y Richard E. Woods. *Digital image processing*. 4ta ed. New York, NY: Pearson, 2018. 1168 págs. ISBN: 978-0-13-335672-4.

[33] Raúl S. González. *TNM staging of stomach carcinomas (AJCC 8th edition)*. 2018. URL: https://www.pathologyoutlines.com/topic/stomachstagingcarcinomas.html (visitado 21-11-2019).

[34] Cristina Grávalos y Antonio Jimeno. «HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target». En: *Annals of Oncology* 19.9 (2008), págs. 1523-1529. DOI: 10.1093/annonc/mdn169. URL: https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mdn169 (visitado 23-07-2019).

[35] Sergio B. Guzmán y Enrique M. Norero. «Cáncer gástrico». En: *Revista Médica Clínica Las Condes* 25.1 (2014), págs. 106-113. DOI: 10.1016/S0716-8640(14)70016-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0716864014700161 (visitado 28-03-2019).

[36] David J. Hand y Robert J. Till. «A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems». En: *Machine Learning* 45.2 (1 de nov. de 2001), págs. 171-186. ISSN: 1573-0565. DOI: 10.1023/A:1010920819831. URL: https://doi.org/10.1023/A:1010920819831 (visitado 22-02-2020).

[37] Jonathon S. Hare y col. «Mind the gap: another look at the problem of the semantic gap in image retrieval». En: *Multimedia Content Analysis, Management, and Retrieval 2006*. Vol. 6073. SPIE, 2006, págs. 75-86. DOI: 10.1117/12.647755. URL: https://doi.org/10.1117/12.647755.

[38] Manfred Hofmann y col. «Assessment of a HER2 scoring system for gastric cancer: results from a validation study». En: *Histopathology* 52.7 (2008), págs. 797-805. DOI: 10.1111/j.1365-2559.2008.03028.x. URL: http://doi.wiley.com/10.1111/j.1365-2559.2008.03028.x (visitado 28-03-2019).

[39] Bing Hu y col. «Gastric cancer: Classification, histology and application of molecular pathology». En: *Journal of Gastrointestinal Oncology* 3.3 (2012), págs. 240-250. DOI: 10.3978/j.issn.2078-6891.2012.021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418539/.

[40] Purva Huilgol. *Accuracy vs. F1-Score*. Medium. 24 de ago. de 2019. URL: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2 (visitado 20-11-2019).

[41] International Agency for Research on Cancer. *Glossary of terms*. URL: http://www-dep.iarc.fr/WHOdb/glossary.htm (visitado 17-07-2019).

[42] Gareth James y col. *An Introduction to Statistical Learning: with Applications in R*. 1ra ed. Springer Texts in Statistics. New York: Springer-Verlag, 2013. 426 págs. ISBN: 978-1-4614-7137-0. URL: https://www.springer.com/gp/book/9781461471370 (visitado 16-08-2019).

[43] Weina Jin. *Nested cross validation explained*. Weina Jin, MD. 25 de ago. de 2018. URL: https://weina.me/nested-cross-validation/ (visitado 01-11-2019).

[44] Andrej Karpathy. *CS231n Convolutional Neural Networks for Visual Recognition*. 2016. URL: http://cs231n.github.io/transfer-learning/ (visitado 02-09-2019).

[45] Diederik P. Kingma y Jimmy Ba. «Adam: A Method for Stochastic Optimization». En: *arXiv:1412.6980 [cs.LG]* (2014). URL: http://arxiv.org/abs/1412.6980 (visitado 31-10-2019).

[46] Ross Kleiman y David Page. «AUCμ: A Performance Metric for Multi-Class Machine Learning Models». En: *Proceedings of the 36th International Conference on Machine Learning*. Ed. por Kamalika Chaudhuri y Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 9 de jun. de 2019, págs. 3439-3447. URL: http://proceedings.mlr.press/v97/kleiman19a.html (visitado 22-02-2020).

[47] Igor Kononenko. «Machine learning for medical diagnosis: history, state of the art and perspective». En: *Artificial Intelligence in Medicine* 23.1 (2001), págs. 89-109. DOI: 10.1016/S0933-3657(01)00077-X. URL: http://www.sciencedirect.com/science/article/pii/S093336570100077X (visitado 13-08-2019).

[48] Konstantina Kourou y col. «Machine learning applications in cancer prognosis and prediction». En: *Computational and Structural Biotechnology Journal* 13 (2015), págs. 8-17. DOI: 10.1016/j.csbj.2014.11.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037014000464 (visitado 28-03-2019).

[49] Klaus Krippendorff. *Computing Krippendorff's Alpha-Reliability*. 2011. URL: http://repository.upenn.edu/asc_papers/43.

[50] Alex Krizhevsky, Ilya Sutskever y Geoffrey E. Hinton. «ImageNet classification with deep convolutional neural networks». En: *Advances in neural information processing systems*. Vol. 25. 2012, págs. 1097-1105.

[51] Ryoji Kushima y col. «Interpretation of HER2 tests in gastric cancer: confirmation of interobserver differences and validation of a QA/QC educational program». En: *Virchows Archiv* 464.5 (2014), págs. 539-545. DOI: 10.1007/s00428-014-1567-9.

[52] J. Richard Landis y Gary G. Koch. «The Measurement of Observer Agreement for Categorical Data». En: *Biometrics* 33.1 (1977), págs. 159-174. DOI: 10.2307/2529310. URL: www.jstor.org/stable/2529310 (visitado 25-11-2019).

[53] Pekka Lauren. «The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification». En: *Acta*

Pathologica Et Microbiologica Scandinavica 64.1 (1965), págs. 31-49. DOI: 10.1111/apm.1965.64.1.31.

[54] Yann Lecun y col. «Gradient-based learning applied to document recognition». En: *Proceedings of the IEEE* 86.11 (1998), págs. 2278-2324. DOI: 10.1109/5.726791.

[55] Jimena López Cayo. «Diseño de un proceso de validación de patología digital en cáncer de mama». Tesis para optar al grado de magíster en informática médica. Santiago, Chile: Universidad de Chile, Facultad de Medicina, Escuela de Postgrado, 2017.

[56] Joseph A. Ludwig y John N. Weinstein. «Biomarkers in Cancer Staging, Prognosis and Treatment Selection». En: *Nature Reviews Cancer* 5.11 (2005), págs. 845-856. DOI: 10.1038/nrc1739. URL: http://www.nature.com/articles/nrc1739 (visitado 28-03-2019).

[57] Satya Mallick. *Number of Parameters and Tensor Sizes in a Convolutional Neural Network (CNN) | Learn OpenCV*. 2018. URL: https://www.learnopencv.com/number-of-parameters-and-tensor-sizes-in-convolutional-neural-network/ (visitado 07-10-2019).

[58] Mary L. McHugh. «Interrater reliability: the kappa statistic». En: *Biochemia Medica* 22.3 (2012), págs. 276-282. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/ (visitado 20-11-2019).

[59] Ministerio de Salud. *Guía Clínica AUGE Cáncer Gástrico*. 3ra Edición. Santiago, Chile, 2014. URL: https://web.minsal.cl/sites/default/files/files/GPC%20G%C3%A1strico%20(PL).pdf (visitado 19-11-2019).

[60] Tom M. Mitchell. *The Discipline of Machine Learning*. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

[61] Bettina Müller y Grupo Oncologico Cooperativo Chileno de Investigation. *Prospective Observational Study of Patients With Locally Advanced Gastric Cancer Treated With Perioperative Chemotherapy and Surgery*. ClinicalTrials.gov Identifier: NCT01633203. 2010. URL: https://www.clinicaltrials.gov/ct2/show/NCT01633203?term=GOCCHI&rank=1 (visitado 10-08-2019).

[62] National Cancer Institute. *Diccionario de cáncer*. National Cancer Institute. 2011. URL: https://www.cancer.gov/espanol/publicaciones/diccionario (visitado 26-06-2019).

[63] Andrew Ng. *Machine Learning and AI via Brain simulations*. Stanford University, 2013. URL: http://datascienceassn.org/sites/default/files/Machine%20Learning%20and%20AI%20via%20Brain%20Simulations.pdf (visitado 21-08-2019).

[64] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: http://neuralnetworksanddeeplearning.com (visitado 23-07-2019).

[65] Sofie L. Nielsen, Søren Nielsen y Mogens Vyberg. «Digital Image Analysis of HER2 Immunostained Gastric and Gastroesophageal Junction Adenocarcinomas:» en: *Applied Immunohistochemistry & Molecular Morphology* 25.5 (2017), págs. 320-328. DOI: 10.1097/PAI.0000000000000463. URL: http://Insights.ovid.com/crossref?an=00129039-201705000-00004 (visitado 28-03-2019).

[66] Sebastian Norena. *Python Model Tuning Methods Using Cross Validation and Grid Search*. Medium. 15 de jun. de 2018. URL: https://medium.com/@sebastiannorena/some-model-tuning-methods-bfef3e6544f0 (visitado 08-10-2019).

[67] Raquel Pezoa y col. «Segmentation of HER2 protein overexpression in immunohistochemically stained breast cancer images using Support Vector Machines». En: *Journal of Physics: Conference Series* 762 (2016), pág. 012050. DOI: 10.1088/1742-6596/

762/1/012050. URL: http://stacks.iop.org/1742-6596/762/i=1/a=012050?key=crossref.0b8ab65f1aa18c15eb6c72bbc07daa7c (visitado 28-03-2019).

[68] Tomi Pitkäaho y col. «Classifying HER2 breast cancer cell samples using deep learning». En: *Proceedings of the 18th Irish Machine Vision and Image Processing conference*. Irish Pattern Recognition y Classification Society, 2016, págs. 78-85.

[69] Foster Provost y Pedro Domingos. *Well-Trained PETs: Improving Probability Estimation Trees*. CDER Working Paper #00-04-IS. New York: Stern School of Business, NYU, 3 de nov. de 2000. URL: http://pages.stern.nyu.edu/~fprovost/Papers/pet-wp.pdf (visitado 22-02-2020).

[70] Ali S. Razavian y col. «CNN features off-the-shelf: an astounding baseline for recognition». En: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, págs. 806-813.

[71] Raúl Rojas. *Neural Networks: A Systematic Introduction*. 1ra ed. Berlin: Springer-Verlag Berlin Heidelberg, 1996. 502 págs. ISBN: 978-3-540-60505-8. URL: https://www.springer.com/gp/book/9783540605058 (visitado 20-08-2019).

[72] Olaf Ronneberger, Philipp Fischer y Thomas Brox. «U-net: Convolutional networks for biomedical image segmentation». En: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, págs. 234-241.

[73] Jeffrey S. Ross y Mary Mulcahy. «HER2 Testing in Gastric/Gastroesophageal Junction Adenocarcinomas: Unique Features of a Familiar Test». En: *Gastrointestinal cancer research: GCR* 4.2 (2011), págs. 62-66.

[74] Sebastian Ruder. «An overview of gradient descent optimization algorithms». En: *arXiv:1609.04747 [cs.LG]* (2016). URL: http://arxiv.org/abs/1609.04747 (visitado 20-08-2019).

[75] David E. Rumelhart, Geoffrey E. Hinton y Ronald J. Williams. «Learning representations by back-propagating errors». En: *Nature* 323.6088 (1986), págs. 533-536. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0 (visitado 20-08-2019).

[76] Josef Rüschoff y col. «HER2 testing in gastric cancer: a practical approach». En: *Modern Pathology* 25.5 (2012), págs. 637-650. DOI: 10.1038/modpathol.2011.198. URL: http://www.nature.com/articles/modpathol2011198 (visitado 28-03-2019).

[77] Olga Russakovsky y col. «ImageNet Large Scale Visual Recognition Challenge». En: *International Journal of Computer Vision* 115.3 (2015), págs. 211-252. DOI: 10.1007/s11263-015-0816-y. URL: https://doi.org/10.1007/s11263-015-0816-y (visitado 31-10-2019).

[78] Monjoy Saha y Chandan Chakraborty. «Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation». En: *IEEE Transactions on Image Processing* 27.5 (2018), págs. 2189-2200. DOI: 10.1109/TIP.2018.2795742.

[79] Sumit Saha. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. Medium. 17 de dic. de 2018. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 (visitado 30-08-2019).

[80] Manfred Schwab, ed. *Encyclopedia of cancer*. 2da ed. New York: Springer Science & Business Media, 2008. 3984 págs. ISBN: 978-3-642-16483-5.

[81] Harshita Sharma y col. «Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology». En:

Computerized Medical Imaging and Graphics 61 (2017), págs. 2-13. DOI: 10.1016/j.compmedimag.2017.06.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895611117300502 (visitado 28-03-2019).

[82] Karen Simonyan y Andrew Zisserman. «Very Deep Convolutional Networks for Large-Scale Image Recognition». En: *arXiv:1409.1556 [cs.CV]* (2014). URL: http://arxiv.org/abs/1409.1556 (visitado 07-10-2019).

[83] Dennis J. Slamon y col. «Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2». En: *New England Journal of Medicine* 344.11 (2001), págs. 783-792. DOI: 10.1056/NEJM200103153441101. URL: http://www.nejm.org/doi/abs/10.1056/NEJM200103153441101 (visitado 18-07-2019).

[84] Devin Soni. *Supervised vs. Unsupervised Learning*. Medium. 16 de jul. de 2019. URL: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d (visitado 08-10-2019).

[85] Emilio Soria-Olivas y col. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques*. 1ra ed. Hershey, PA: Information Science Reference, 2009. 834 págs. ISBN: 978-1-60566-766-9.

[86] Christian Szegedy y col. «Rethinking the inception architecture for computer vision». En: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, págs. 2818-2826.

[87] *TensorFlow Hub*. TensorFlow. URL: https://www.tensorflow.org/hub (visitado 02-09-2019).

[88] Sik-Ho Tsang. *Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015*. Medium. 9 de oct. de 2018. URL: https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c (visitado 31-10-2019).

[89] Michel E. Vandenberghe y col. «Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer». En: *Scientific Reports* 7.1 (2017), pág. 45938. DOI: 10.1038/srep45938. URL: http://www.nature.com/articles/srep45938 (visitado 28-03-2019).

[90] Antonio C. Wolff y col. «American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer». En: *Journal of Clinical Oncology* 25.1 (2007), págs. 118-145. DOI: 10.1200/JCO.2006.09.2775. URL: http://ascopubs.org/doi/10.1200/JCO.2006.09.2775 (visitado 28-03-2019).

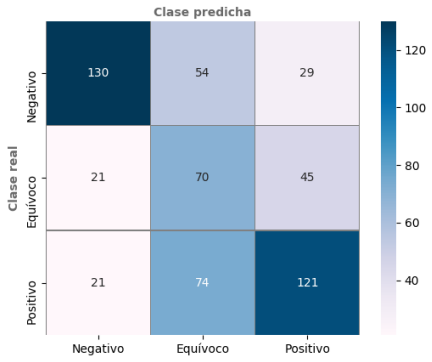
[91] Jason Yosinski y col. «How transferable are features in deep neural networks?» En: *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, págs. 3320-3328. URL: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf.

[92] Falk Zakrzewski y col. «Automated detection of the HER2 gene amplification status in Fluorescence in situ hybridization images for the diagnostics of cancer tissues». En: *Scientific Reports* 9.1 (2019), pág. 8231. DOI: 10.1038/s41598-019-44643-z. URL: http://www.nature.com/articles/s41598-019-44643-z (visitado 08-10-2019).

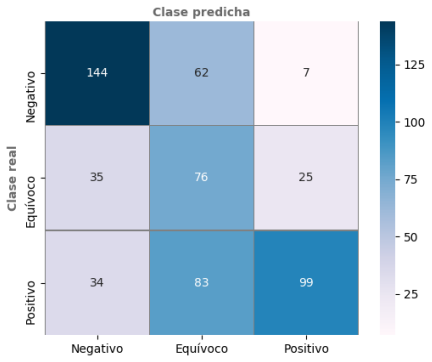
[93] Jiakang Zhang. *The Overview of Gradient Descent Algorithm*. Jiakang's blog. 15 de ene. de 2019. URL: http://www.ikouz.com/2019/01/the-overview-of-gradient-descent-algorithm/ (visitado 08-10-2019).

Apéndice A

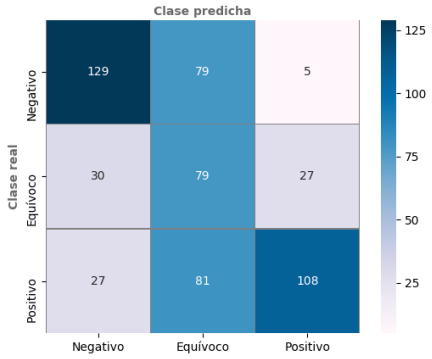
Matrices de confusión sin normalizar, macroexperimento I



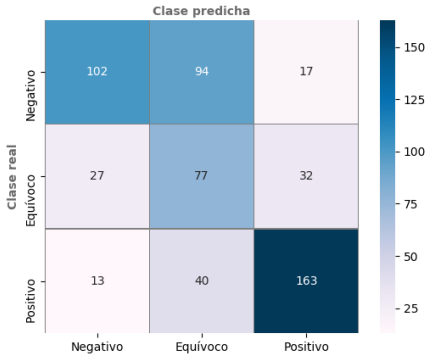
(a) Configuración simple.



(b) *Data augmentation.*

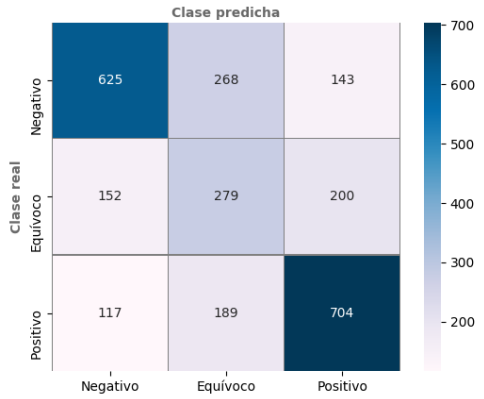


(c) *Fine Tuning* y *data augmentation.*

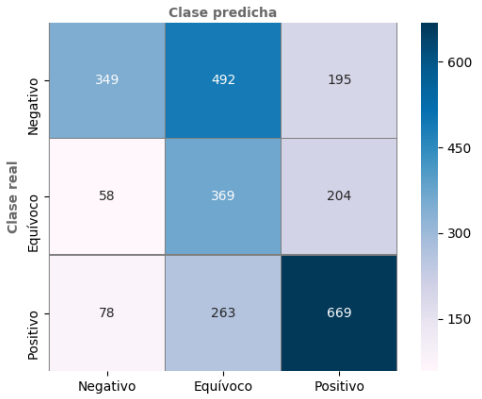


(d) Reentrenamiento total.

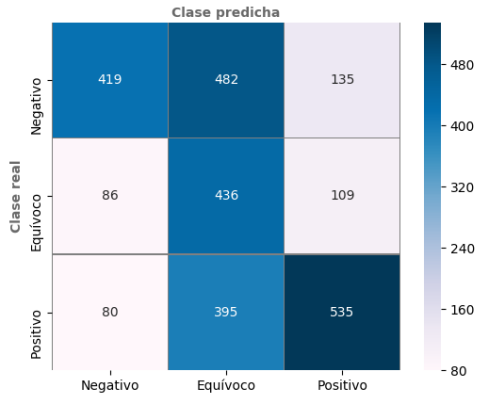
Figura A.1: Matrices de confusión sin normalizar para macroexperimento I, magnificación 10x.



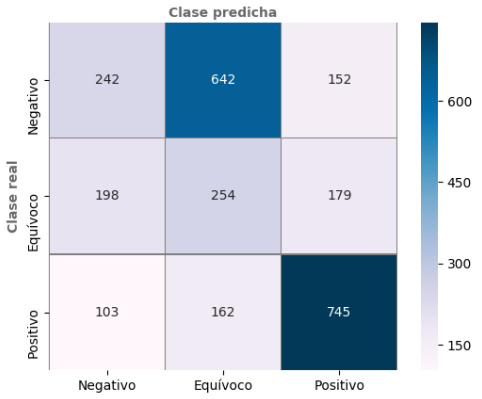
(a) Configuración simple.



(b) *Data augmentation*.

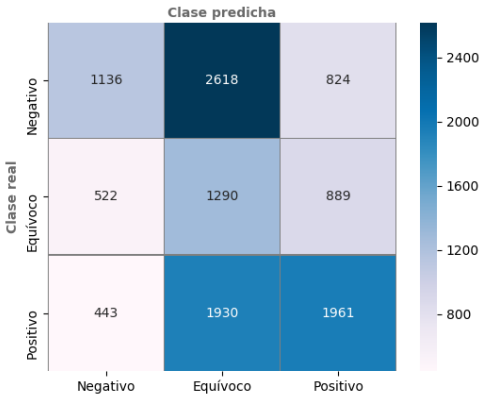


(c) *Fine Tuning* y *data augmentation*.

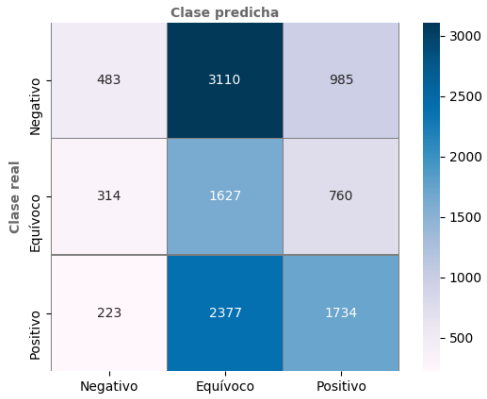


(d) Reentrenamiento total.

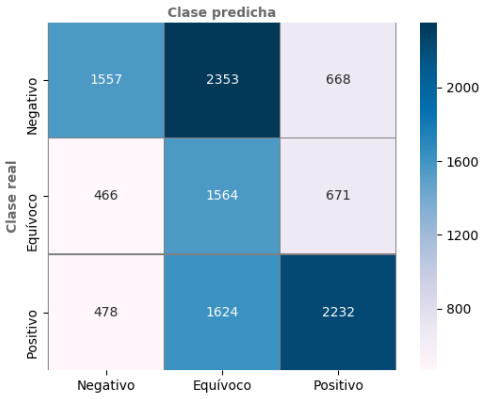
Figura A.2: Matrices de confusión sin normalizar para macroexperimento I, magnificación 20x.



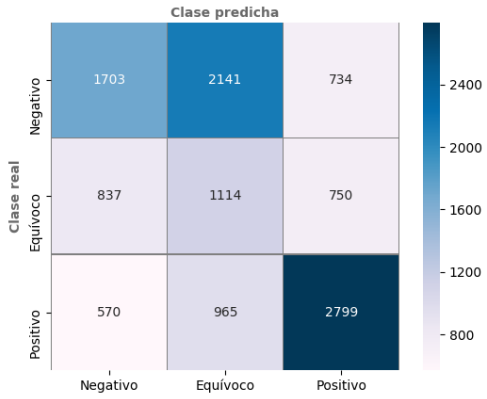
(a) Configuración simple.



(b) *Data augmentation*.



(c) *Fine Tuning* y *data augmentation*.



(d) Reentrenamiento total.

Figura A.3: Matrices de confusión sin normalizar para macroexperimento I, magnificación 40x.

Apéndice B

Matrices de confusión sin normalizar, macroexperimento II

B.1 Evaluación de parches

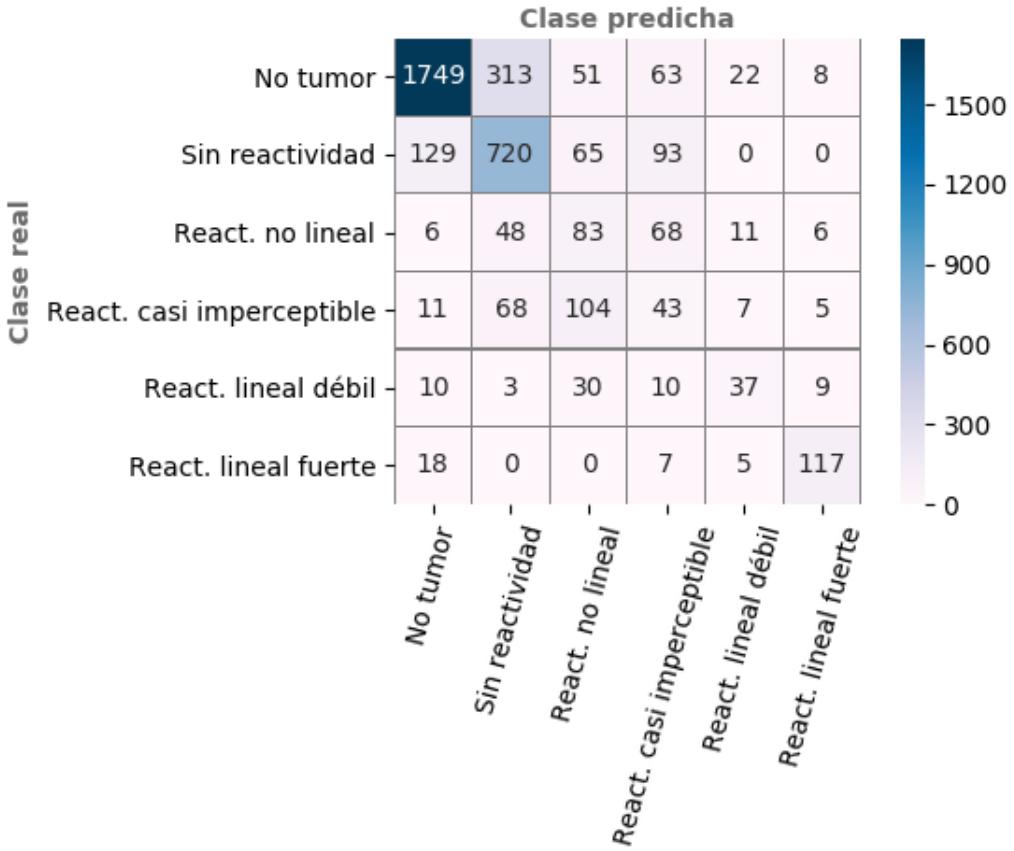
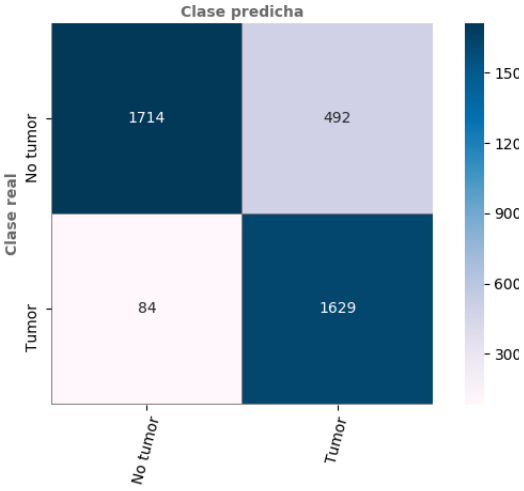
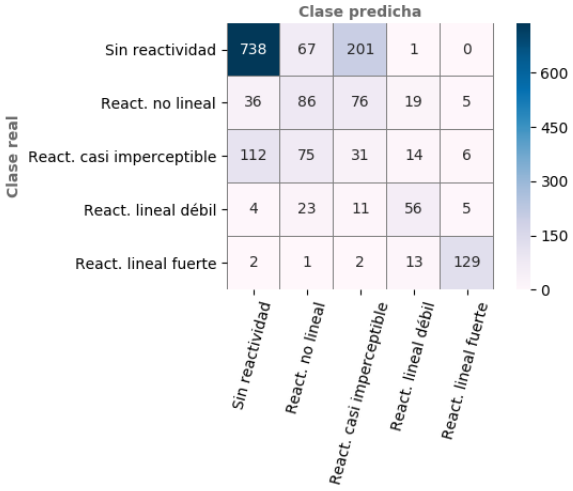


Figura B.1: Matriz de confusión sin normalizar para experimento *todo en uno*, evaluación de clasificación de parches de ROIs.



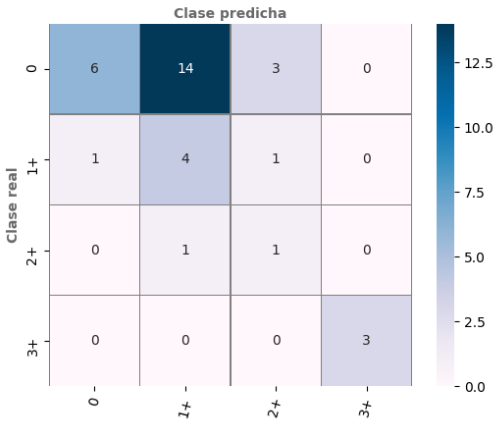
(a) No tumor / Tumor (umbral $T = 0{,}3$).



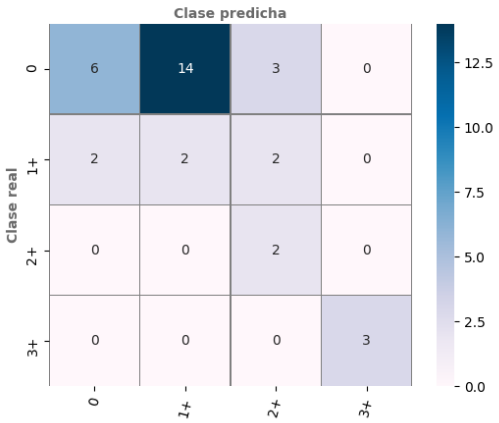
(b) Reactividad HER2.

Figura B.2: Matrices de confusión sin normalizar para experimento *en cascada*, evaluación de clasificación de parches de ROIs.

B.2 Evaluación de biopsias

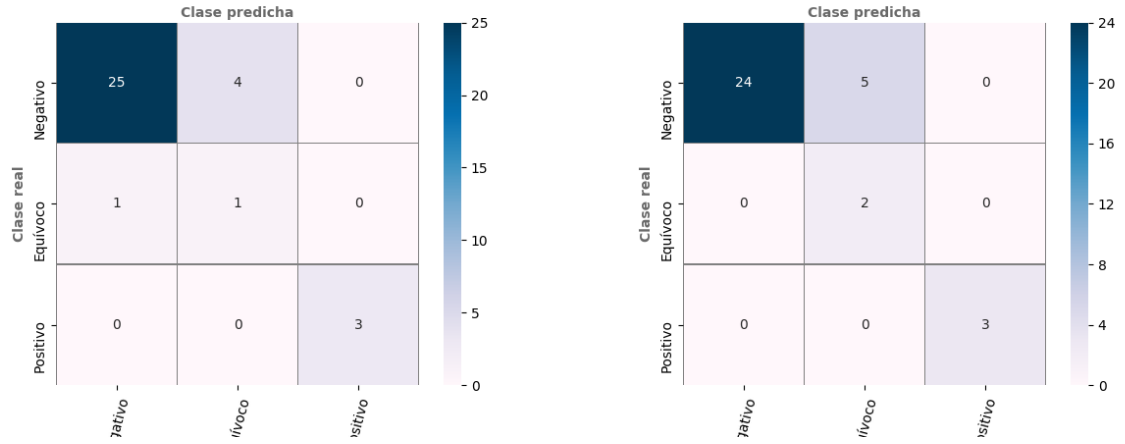


(a) *Todo en uno*.



(b) Cascada.

Figura B.3: Matrices de confusión sin normalizar para evaluación de clasificación de biopsias, usando clases 0, 1+, 2+ y 3+. En clasificación binaria de esquema *en cascada* se utilizó un umbral $T = 0{,}3$.



(a) *Todo en uno.* (b) Cascada.

Figura B.4: Matrices de confusión sin normalizar para evaluación de clasificación de biopsias, usando clases *Negativo*, *Equívoco* y *Positivo*. En clasificación binaria de esquema *en cascada* se utilizó un umbral $T = 0{,}3$.